INTRODUCTION TO EARTH AND SPACE SCIENCES



Robert Wagner
HACC



Introduction to Earth and Space Sciences

Robert Wagner

HACC, Central Pennsylvania's Community College

This text is disseminated via the Open Education Resource (OER) LibreTexts Project (https://LibreTexts.org) and like the thousands of other texts available within this powerful platform, it is freely available for reading, printing, and "consuming."

The LibreTexts mission is to bring together students, faculty, and scholars in a collaborative effort to provide an accessible, and comprehensive platform that empowers our community to develop, curate, adapt, and adopt openly licensed resources and technologies; through these efforts we can reduce the financial burden borne from traditional educational resource costs, ensuring education is more accessible for students and communities worldwide.

Most, but not all, pages in the library have licenses that may allow individuals to make changes, save, and print this book. Carefully consult the applicable license(s) before pursuing such effects. Instructors can adopt existing LibreTexts texts or Remix them to quickly build course-specific resources to meet the needs of their students. Unlike traditional textbooks, LibreTexts' web based origins allow powerful integration of advanced features and new technologies to support learning.



LibreTexts is the adaptable, user-friendly non-profit open education resource platform that educators trust for creating, customizing, and sharing accessible, interactive textbooks, adaptive homework, and ancillary materials. We collaborate with individuals and organizations to champion open education initiatives, support institutional publishing programs, drive curriculum development projects, and more.

The LibreTexts libraries are Powered by NICE CXone Expert and were supported by the Department of Education Open Textbook Pilot Project, the California Education Learning Lab, the UC Davis Office of the Provost, the UC Davis Library, the California State University Affordable Learning Solutions Program, and Merlot. This material is based upon work supported by the National Science Foundation under Grant No. 1246120, 1525057, and 1413739.

Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation nor the US Department of Education.

Have questions or comments? For information about adoptions or adaptions contact info@LibreTexts.org or visit our main website at https://LibreTexts.org.

The LibreTexts name, logo, and book covers are trademarked by LibreTexts, Inc. (a 501(c)(3) not-for-profit organization) and protected against unauthorized use.

This text was compiled on 09/15/2026

TABLE OF CONTENTS

Licensing

1: Studying the Universe

1.1: The Nature of Astronomy

1.2: The Nature of Science

1.3: Consequences of Light Travel Time

1.4: A Tour of the Universe

2: Gravity and Planetary Motion

2.1: The Laws of Planetary Motion

2.2: Newton's Great Synthesis

2.3: Newton's Universal Law of Gravitation

3: Our Solar System

3.1: Overview of Our Planetary System

3.2: Composition and Structure of Planets

3.3: Formation of Planetary Systems

3.3.1: Origin of the Solar System

3.3.2: Formation of the Solar System

3.3.3: Comparison with Other Planetary Systems

4: Light and Matter

4.1: The Behavior of Light

4.2: The Electromagnetic Spectrum

4.3: Spectroscopy in Astronomy

4.4: The Structure of the Atom

4.5: Formation of Spectral Lines

4.6: The Doppler Effect

5: Stars

5.1: The Structure and Composition of the Sun

5.2: Understanding Stars

5.2.1: The Brightness of Stars

5.2.2: Colors of Stars

5.2.3: The Spectra of Stars (and Brown Dwarfs)

5.2.4: Using Spectra to Measure Stellar Radius, Composition, and Motion

5.2.5: The H-R Diagram

5.3: Stellar Evolution

5.3.1: Star Formation

5.3.2: The H-R Diagram and the Study of Stellar Evolution

5.3.3: Evolution from the Main Sequence to Red Giants
5.3.4: The Death of Low-Mass Stars
5.3.5: Evolution of Massive Stars-An Explosive Finish

6: Galaxies and the Universe
6.1: The Architecture of the Galaxy
6.2: Types of Galaxies
6.3: The Expanding Universe
6.4: The Beginning of the Universe
6.5: The Cosmic Microwave Background

7: Earth's Atmosphere
7.1: Overview of Earth's Atmosphere
7.2: Layers of the Atmosphere
7.3: Weather and Climate: What's the Difference?
7.4: Radiation
7.5: Earth-Sun Energy Balance
7.5.1: Albedo
7.5.2: Surface Energy Balance
7.5.3: Radiation Interaction with the Atmosphere
7.6: Air Pollution

8: Energy & Humidity
8.1: Energy Transfer
8.2: Latent Heat
8.3: Saturation
8.4: Humidity
8.5: Why Do We Care So Much About Moisture?
8.6: Atmospheric Stability and Lapse Rates

9: Clouds and Precipitation
9.1: Cloud Condensation Nuclei
9.2: Cloud Naming Conventions
9.3: Cloud Droplets
9.4: Precipitation Types

10: Weather Observation and Maps
10.1: Meteorological Reports and Observations
10.2: Synoptic Weather Maps
10.3: Map Analysis

11: Atmospheric Circulation & Air Masses
11.1: Atmospheric Circulation
11.1.1: Main Forces

11.1.2: Differential Heating
11.1.3: Single-Cell Model
11.1.4: Three-Cell Model
11.1.5: Global Surface Winds
11.1.6: Jet Streams
11.2: Air Masses and Fronts
11.2.1: Air Masses
11.2.2: Surface Fronts
12: Weather Systems and Severe Weather
12.1: Mid-latitude Frontal Cyclones
12.2: Tropical Cyclones
12.2.1: Tropical Cyclone Structure
12.2.2: Intensity and Geographic Distribution
12.2.3: Evolution
12.2.4: Hazards
12.3: Thunderstorms and Severe Weather
12.3.1: Thunderstorms
12.3.2: Atmospheric Instability and Thunderstorms
12.3.3: Precipitation and Hail
12.3.4: Lightning
12.3.5: Tornadoes
12.3.6: Keeping Yourself Safe
13: Earth Materials
13.1: What is Geology?
13.2: Minerals and Rocks
13.3: Mineral Groups
13.4: Silicate Minerals
13.5: Mineral Properties
14: Rocks and Geological Time
14.1: The Three Rock Types
14.1.1: Classification of Igneous Rock
14.1.2: Clastic Sedimentary Rocks
14.1.3: Chemical Sedimentary Rocks
14.1.4: Sedimentary Structures and Fossils
14.1.5: Classification of Metamorphic Rocks
14.2: The Rock Cycle
14.3: Reading Earth's History
14.3.1: Geological Time

14.3.2: The Geological Time Scale
14.3.3: Relative Dating Methods
14.3.4: Isotopic Dating Methods
14.3.5: Understanding Geological Time

15: The Dynamic Earth
15.1: Understanding Earth Through Seismology
15.2: Plate Tectonics
15.2.1: Alfred Wegener- The Father of Plate Tectonics
15.2.2: Plate, Plate Motions, and Plate Boundary Processes
15.2.3: Mechanisms for Plate Motion
15.3: Earthquakes
15.3.1: What is an Earthquake?
15.3.2: Earthquakes and Plate Tectonics
15.3.3: Measuring Earthquakes
15.3.4: The Impacts of Earthquakes
15.4: Volcanoes
15.4.1: Plate Tectonics and Volcanism
15.4.2: Magma Composition and Eruption Style
15.4.3: Types of Volcanoes
15.4.4: Volcanic Hazards

16: Weathering and Soils
16.1: Mechanical Weathering
16.2: Chemical Weathering
16.3: The Products of Weathering and Erosion
16.4: Weathering and the Formation of Soil

17: Water and Landscapes
17.1: Topographic Maps
17.2: Surface Water
17.2.1: The Hydrological Cycle
17.2.2: Drainage Basins
17.2.3: Stream Erosion and Deposition
17.2.4: Stream Types
17.2.5: Flooding
17.3: Groundwater
17.3.1: Groundwater and Aquifers
17.3.2: Groundwater Flow
17.3.3: Groundwater Extraction
17.3.4: Groundwater Quality

17.4: Mass Wasting

17.4.1: Factors That Control Slope Stability

17.4.2: Classification of Mass Wasting

17.4.3: Preventing, Delaying, Monitoring, and Mitigating Mass Wasting

18: Glaciers, Earth's History, and Other Worlds

18.1: How Glaciers Work

18.2: Glacial Erosion

18.3: Glacial Deposition

18.4: Earth’s First 2 Billion Years

18.5: Are There Other Earths?

Index

Glossary

Detailed Licensing

This page titled Introduction to Earth and Space Sciences is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by LibreTexts.

Licensing

A detailed breakdown of this resource's licensing can be found in ***Back Matter/Detailed Licensing***.

This page titled Licensing is shared under a Public Domain license and was authored, remixed, and/or curated by LibreTexts.

1: STUDYING THE UNIVERSE

1.1: The Nature of Astronomy
1.2: The Nature of Science
1.3: Consequences of Light Travel Time
1.4: A Tour of the Universe

Welcome to Physical Science - Chapter 1 - Introduction

Physical science explores the processes and principles that shape the world around us—and the universe beyond it. In this course, we will approach physical science through three connected fields: **astronomy, meteorology, and geology**. Together, these fields allow us to examine nature across an enormous range of scales, from stars and galaxies to Earth's atmosphere, rocks, and landscapes.

We begin with **astronomy**, looking outward before gradually bringing our focus closer to home. During this unit, we will explore how scientists study the universe, how gravity governs the motions of planets, how planetary systems form, how light reveals the properties of distant objects, and how stars and galaxies change over time. Along the way, we will encounter many of the same physical principles that will appear again when we study Earth's atmosphere and geology.

This first chapter provides the foundation for that journey. It introduces astronomy as a science, considers the nature of scientific investigation, and gives us a sense of the scale of the universe. One particularly important idea is that **looking across space also means looking back in time**. Because light travels at a finite speed, the universe we observe is not simply a picture of where things are, but also a record of their past.

As you work through this course, the goal is not simply to memorize facts about space, weather, or rocks. Instead, pay attention to the **evidence scientists use, the patterns they identify, and the physical processes they use to explain what they observe**. Those connections will help tie together the three branches of physical science we will study this semester.

This page titled 1: Studying the Universe is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Robert Wagner.

1.1: The Nature of Astronomy

Astronomy is defined as the study of the objects that lie beyond our planet Earth and the processes by which these objects interact with one another. We will see, though, that it is much more. It is also humanity's attempt to organize what we learn into a clear history of the universe, from the instant of its birth in the Big Bang to the present moment. Throughout this book, we emphasize that science is a *progress report*—one that changes constantly as new techniques and instruments allow us to probe the universe more deeply.

In considering the history of the universe, we will see again and again that the cosmos *evolves*; it changes in profound ways over long periods of time. For example, the universe made the carbon, the calcium, and the oxygen necessary to construct something as interesting and complicated as you. Today, many billions of years later, the universe has evolved into a more hospitable place for life. Tracing the evolutionary processes that continue to shape the universe is one of the most important (and satisfying) parts of modern astronomy.

This page titled 1.1: The Nature of Astronomy is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by OpenStax via source content that was edited to the style and standards of the LibreTexts platform.

1.2: The Nature of Science

The ultimate judge in science is always what nature itself reveals based on observations, experiments, models, and testing. Science is not merely a body of knowledge, but a *method* by which we attempt to understand nature and how it behaves. This method begins with many observations over a period of time. From the trends found through observations, scientists can *model* the particular phenomena we want to understand. Such models are always approximations of nature, subject to further testing.

As a concrete astronomical example, ancient astronomers constructed a model (partly from observations and partly from philosophical beliefs) that Earth was the center of the universe and everything moved around it in circular orbits. At first, our available observations of the Sun, Moon, and planets did fit this model; however, after further observations, the model had to be updated by adding circle after circle to represent the movements of the planets around Earth at the center. As the centuries passed and improved instruments were developed for keeping track of objects in the sky, the old model (even with a huge number of circles) could no longer explain all the observed facts. As we will see in the chapter on Observing the Sky: The Birth of Astronomy, a new model, with the Sun at the center, fit the experimental evidence better. After a period of philosophical struggle, it became accepted as our view of the universe.

When they are first proposed, new models or ideas are sometimes called *hypotheses*. You may think there can be no new hypotheses in a science such as astronomy—that everything important has already been learned. Nothing could be further from the truth. Throughout this textbook you will find discussions of recent, and occasionally still controversial, hypotheses in astronomy. For example, the significance that the huge chunks of rock and ice that hit Earth have for life on Earth itself is still debated. And while the evidence is strong that vast quantities of invisible “dark energy” make up the bulk of the universe, scientists have no convincing explanation for what the dark energy actually is. Resolving these issues will require difficult observations done at the forefront of our technology, and all such hypotheses need further testing before we incorporate them fully into our standard astronomical models.

This last point is crucial: a hypothesis must be a proposed explanation that can be *tested*. The most straightforward approach to such testing in science is to perform an experiment. If the experiment is conducted properly, its results either will agree with the predictions of the hypothesis or they will contradict it. If the experimental result is truly inconsistent with the hypothesis, a scientist must discard the hypothesis and try to develop an alternative. If the experimental result agrees with predictions, this does not necessarily prove that the hypothesis is absolutely correct; perhaps later experiments will contradict crucial parts of the hypothesis. But, the more experiments that agree with the hypothesis, the more likely we are to accept the hypothesis as a useful description of nature.

One way to think about this is to consider a scientist who was born and lives on an island where only black sheep live. Day after day the scientist encounters black sheep only, so he or she hypothesizes that all sheep are black. Although every observed sheep adds confidence to the hypothesis, the scientist only has to visit the mainland and observe one white sheep to prove the hypothesis wrong.

When you read about experiments, you probably have a mental picture of a scientist in a laboratory conducting tests or taking careful measurements. This is certainly the case for a biologist or a chemist, but what can astronomers do when our laboratory is the universe? It’s impossible to put a group of stars into a test tube or to order another comet from a scientific supply company.

As a result, astronomy is sometimes called an *observational* science; we often make our tests by observing many samples of the kind of object we want to study and noting carefully how different samples vary. New instruments and technology can let us look at astronomical objects from new perspectives and in greater detail. Our hypotheses are then judged in the light of this new information, and they pass or fail in the same way we would evaluate the result of a laboratory experiment.

Much of astronomy is also a *historical* science—meaning that what we observe has already happened in the universe and we can do nothing to change it. In the same way, a geologist cannot alter what has happened to our planet, and a paleontologist cannot bring an ancient animal back to life. While this can make astronomy challenging, it also gives us fascinating opportunities to discover the secrets of our cosmic past.

You might compare an astronomer to a detective trying to solve a crime that occurred before the detective arrived at the scene. There is lots of evidence, but both the detective and the scientist must sift through and organize the evidence to test various hypotheses about what actually happened. And there is another way in which the scientist is like a detective: they both must prove their case. The detective must convince the district attorney, the judge, and perhaps ultimately the jury that his hypothesis

is correct. Similarly, the scientist must convince colleagues, editors of journals, and ultimately a broad cross-section of other scientists that her hypothesis is provisionally correct. In both cases, one can only ask for evidence "beyond a reasonable doubt." And sometimes new evidence will force both the detective and the scientist to revise their last hypothesis.

This self-correcting aspect of science sets it off from most human activities. Scientists spend a great deal of time questioning and challenging one another, which is why applications for project funding—as well as reports for publication in academic journals—go through an extensive process of *peer review*, which is a careful examination by other scientists in the same field. In science (after formal education and training), everyone is encouraged to improve upon experiments and to challenge any and all hypotheses. New scientists know that one of the best ways to advance their careers is to find a weakness in our current understanding of something and to correct it with a new or modified hypothesis.

This is one of the reasons science has made such dramatic progress. An undergraduate science major today knows more about science and math than did Sir Isaac Newton, one of the most renowned scientists who ever lived. Even in this introductory astronomy course, you will learn about objects and processes that no one a few generations ago even dreamed existed.

This page titled 1.2: The Nature of Science is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by OpenStax via source content that was edited to the style and standards of the LibreTexts platform.

1.3: Consequences of Light Travel Time

There is another reason the speed of light is such a natural unit of distance for astronomers. Information about the universe comes to us almost exclusively through various forms of light, and all such light travels at the speed of light—that is, 1 light-year every year. This sets a limit on how quickly we can learn about events in the universe. If a star is 100 light-years away, the light we see from it tonight left that star 100 years ago and is just now arriving in our neighborhood. The soonest we can learn about any changes in that star is 100 years after the fact. For a star 500 light-years away, the light we detect tonight left 500 years ago and is carrying 500-year-old news.

Because many of us are accustomed to instant news from the Internet, some might find this frustrating.

"You mean, when I see that star up there," you ask, "I won't know what's actually happening there for another 500 years?"

But this isn't the most helpful way to think about the situation. For astronomers, *now* is when the light reaches us here on Earth. There is no way for us to know anything about that star (or other object) until its light reaches us.

But what at first may seem a great frustration is actually a tremendous benefit in disguise. If astronomers really want to piece together what has happened in the universe since its beginning, they must find evidence about each epoch (or period of time) of the past. Where can we find evidence today about cosmic events that occurred billions of years ago?

The delay in the arrival of light provides an answer to this question. The farther out in space we look, the longer the light has taken to get here, and the longer ago it left its place of origin. By looking billions of light-years out into space, astronomers are actually seeing billions of years into the past. In this way, we can reconstruct the history of the cosmos and get a sense of how it has evolved over time.

This is one reason why astronomers strive to build telescopes that can collect more and more of the faint light in the universe. The more light we collect, the fainter the objects we can observe. On average, fainter objects are farther away and can, therefore, tell us about periods of time even deeper in the past. Instruments such as the Hubble Space Telescope (Figure 1.5) and the Very Large Telescope in Chile (which you will learn about in the chapter on Astronomical Instruments), are giving astronomers views of deep space and deep time better than any we have had before.



Figure 1.5 : Telescope in Orbit. The Hubble Space Telescope, shown here in orbit around Earth, is one of many astronomical instruments in space. (credit: modification of work by European Space Agency)

This page titled 1.3: Consequences of Light Travel Time is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by OpenStax via source content that was edited to the style and standards of the LibreTexts platform.

1.4: A Tour of the Universe

We can now take a brief introductory tour of the universe as astronomers understand it today to get acquainted with the types of objects and distances you will encounter throughout the text. We begin at home with Earth, a nearly spherical planet about 13,000 kilometers in diameter (Figure 1.6). A space traveler entering our planetary system would easily distinguish Earth from the other planets in our solar system by the large amount of liquid water that covers some two thirds of its crust. If the traveler had equipment to receive radio or television signals, or came close enough to see the lights of our cities at night, she would soon find signs that this watery planet has sentient life.

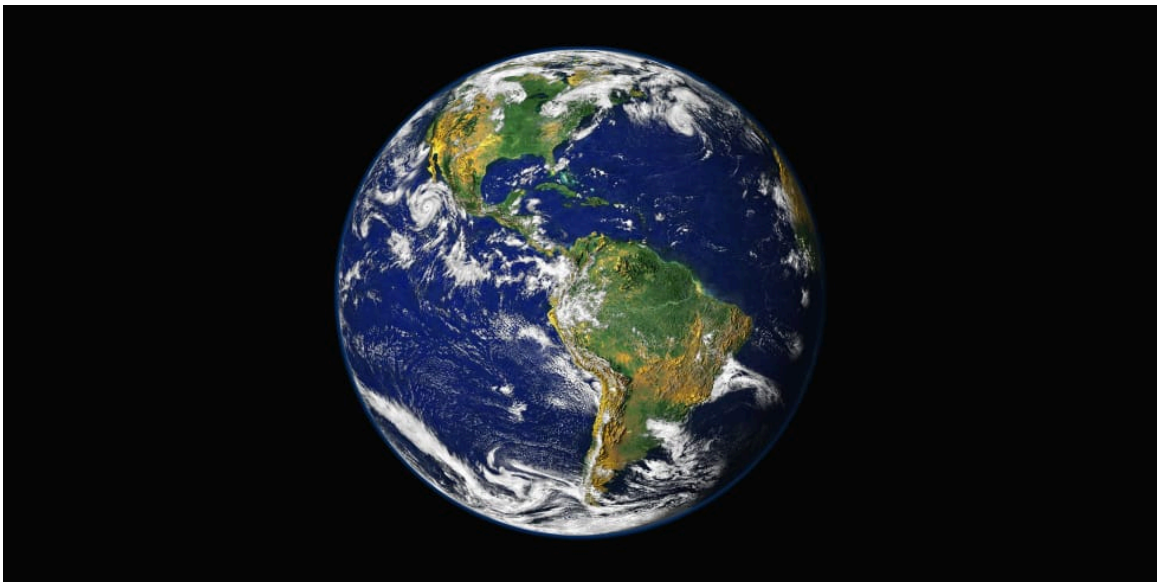


Figure 1.6 : Humanity's Home Base. This image shows the Western hemisphere as viewed from space 35,400 kilometers (about 22,000 miles) above Earth. Data about the land surface from one satellite was combined with another satellite's data about the clouds to create the image. (credit: modification of work by R. Stockli, A. Nelson, F. Hasler, NASA/ GSFC/ NOAA/ USGS)

Our nearest astronomical neighbor is Earth's satellite, commonly called the *Moon*. Figure 1.7 shows Earth and the Moon drawn to scale on the same diagram. Notice how small we have to make these bodies to fit them on the page with the right scale. The Moon's distance from Earth is about 30 times Earth's diameter, or approximately 384,000 kilometers, and it takes about a month for the Moon to revolve around Earth. The Moon's diameter is 3476 kilometers, about one fourth the size of Earth.



Figure 1.7 : Earth and Moon, Drawn to Scale. This image shows Earth and the Moon shown to scale for both size and distance. (credit: modification of work by NASA)

Light (or radio waves) takes 1.3 seconds to travel between Earth and the Moon. If you've seen videos of the Apollo flights to the Moon, you may recall that there was a delay of about 3 seconds between the time Mission Control asked a question and the time the astronauts responded. This was not because the astronauts were thinking slowly, but rather because it took the radio waves almost 3 seconds to make the round trip.

Earth revolves around our star, the Sun, which is about 150 million kilometers away—approximately 400 times as far away from us as the Moon. We call the average Earth–Sun distance an *astronomical unit* (AU) because, in the early days of astronomy, it was the most important measuring standard. Light takes slightly more than 8 minutes to travel 1 astronomical unit, which means the latest news we receive from the Sun is always 8 minutes old. The diameter of the Sun is about 1.5 million kilometers; Earth could fit comfortably inside one of the minor eruptions that occurs on the surface of our star. If the Sun were reduced to the size of a basketball, Earth would be a small apple seed about 30 meters from the ball.

It takes Earth 1 year (3×1073×107 seconds) to go around the Sun at our distance; to make it around, we must travel at approximately 110,000 kilometers per hour. (If you, like many students, still prefer miles to kilometers, you might find the following trick helpful. To convert kilometers to miles, just multiply kilometers by 0.6. Thus, 110,000 kilometers per hour becomes 66,000 miles per hour.) Because gravity holds us firmly to Earth and there is no resistance to Earth's motion in the vacuum of space, we participate in this extremely fast-moving trip without being aware of it day to day.

Earth is only one of eight planets that revolve around the Sun. These planets, along with their moons and swarms of smaller bodies such as dwarf planets, make up the solar system (Figure 1.8). A planet is defined as a body of significant size that orbits a star and does not produce its own light. (If a large body consistently produces its own light, it is then called a *star*.) Later in the book this definition will be modified a bit, but it is perfectly fine for now as you begin your voyage.

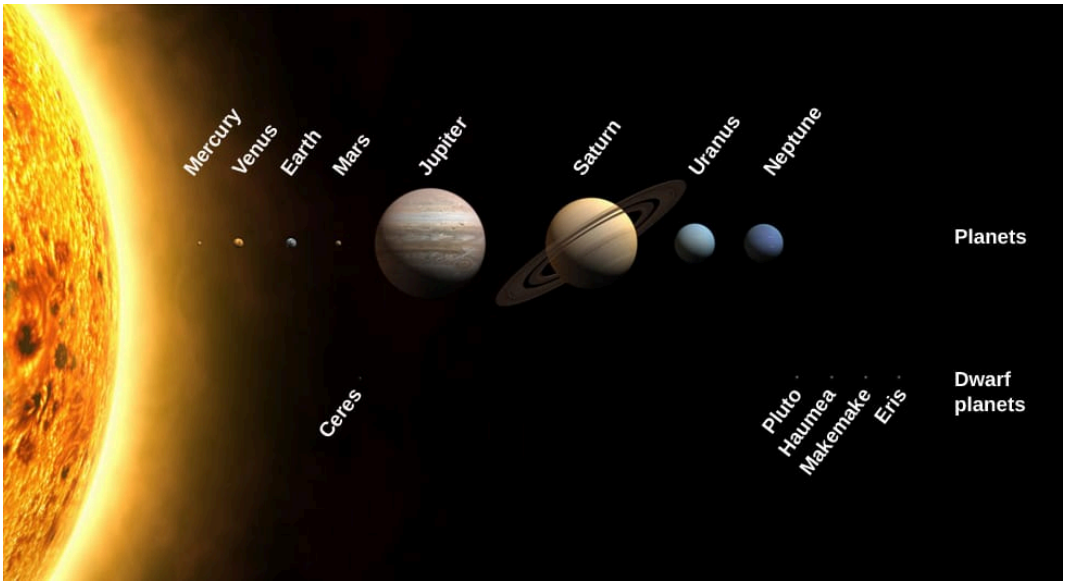


Figure 1.8 : Our Solar Family. The Sun, the planets, and some dwarf planets are shown with their sizes drawn to scale. The orbits of the planets are much more widely separated than shown in this drawing. Notice the size of Earth compared to the giant planets. (credit: modification of work by NASA)

We are able to see the nearby planets in our skies only because they reflect the light of our local star, the Sun. If the planets were much farther away, the tiny amount of light they reflect would usually not be visible to us. The planets we have so far discovered orbiting other stars were found from the pull their gravity exerts on their parent stars, or from the light they block from their stars when they pass in front of them. We can't see most of these planets directly, although a few are now being imaged directly.

The Sun is our local star, and all the other stars are also enormous balls of glowing gas that generate vast amounts of energy by nuclear reactions deep within. We will discuss the processes that cause stars to shine in more detail later in the book. The other stars look faint only because they are so very far away. If we continue our basketball analogy, Proxima Centauri, the nearest star beyond the Sun, which is 4.3 light-years away, would be almost 7000 kilometers from the basketball.

When you look up at a star-filled sky on a clear night, all the stars visible to the unaided eye are part of a single collection of stars we call the *Milky Way Galaxy*, or simply the *Galaxy*. (When referring to the Milky Way, we capitalize *Galaxy*; when talking about other galaxies of stars, we use lowercase *galaxy*.) The Sun is one of hundreds of billions of stars that make up the Galaxy; its extent, as we will see, staggers the human imagination. Within a sphere 10 light-years in radius centered on the Sun, we find roughly ten stars. Within a sphere 100 light-years in radius, there are roughly 10,000 (104)(104) stars—far too many to count or name—but we have still traversed only a tiny part of the Milky Way Galaxy. Within a 1000-light-year sphere, we find some ten million (107)(107) stars; within a sphere of 100,000 light-years, we finally encompass the entire Milky Way Galaxy.

Our Galaxy looks like a giant disk with a small ball in the middle. If we could move outside our Galaxy and look down on the disk of the Milky Way from above, it would probably resemble the galaxy in Figure 1.9, with its spiral structure outlined by the blue light of hot adolescent stars.



Figure 1.9 : Spiral Galaxy. This galaxy of billions of stars, called by its catalog number NGC 1073, is thought to be similar to our own Milky Way Galaxy. Here we see the giant wheel-shaped system with a bar of stars across its middle. (credit: NASA, ESA)

The Sun is somewhat less than 30,000 light-years from the center of the Galaxy, in a location with nothing much to distinguish it. From our position inside the Milky Way Galaxy, we cannot see through to its far rim (at least not with ordinary light) because the space between the stars is not completely empty. It contains a sparse distribution of gas (mostly the simplest element, hydrogen) intermixed with tiny solid particles that we call *interstellar dust*. This gas and dust collect into enormous clouds in many places in the Galaxy, becoming the raw material for future generations of stars. Figure 1.10 shows an image of the disk of the Galaxy as seen from our vantage point.



Figure 1.10 : Milky Way Galaxy. Because we are inside the Milky Way Galaxy, we see its disk in cross-section flung across the sky like a great milky white avenue of stars with dark “rifts” of dust. In this dramatic image, part of it is seen above Trona Pinnacles in the California desert. (credit: Ian Norman)

Typically, the interstellar material is so extremely sparse that the space between stars is a much better vacuum than anything we can produce in terrestrial laboratories. Yet, the dust in space, building up over thousands of light-years, can block the light of more distant stars. Like the distant buildings that disappear from our view on a smoggy day in Los Angeles, the more distant regions of the Milky Way cannot be seen behind the layers of interstellar smog. Luckily, astronomers have found that stars and raw material shine with various forms of light, some of which do penetrate the smog, and so we have been able to develop a pretty good map of the Galaxy.

Recent observations, however, have also revealed a rather surprising and disturbing fact. There appears to be more—much more—to the Galaxy than meets the eye (or the telescope). From various investigations, we have evidence that much of our Galaxy is made of material we cannot currently observe directly with our instruments. We therefore call this component of the Galaxy *dark matter*. We know the dark matter is there by the pull its gravity exerts on the stars and raw material we can observe, but what this dark matter is made of and how much of it exists remain a mystery. Furthermore, this dark matter is not confined to our Galaxy; it appears to be an important part of other star groupings as well.

By the way, not all stars live by themselves, as the Sun does. Many are born in double or triple systems with two, three, or more stars revolving about each other. Because the stars influence each other in such close systems, multiple stars allow us to measure characteristics that we cannot discern from observing single stars. In a number of places, enough stars have formed together that we recognized them as star clusters (Figure 1.11). Some of the largest of the star clusters that astronomers have cataloged contain hundreds of thousands of stars and take up volumes of space hundreds of light-years across.



Figure 1.11 : Star Cluster. This large star cluster is known by its catalog number, M9. It contains some 250,000 stars and is seen more clearly from space using the Hubble Space Telescope. It is located roughly 25,000 light-years away. (credit: NASA, ESA)

You may hear stars referred to as “eternal,” but in fact no star can last forever. Since the “business” of stars is making energy, and energy production requires some sort of fuel to be used up, eventually all stars run out of fuel. This news should not cause you to panic, though, because our Sun still has at least 5 or 6 billion years to go. Ultimately, the Sun and all stars will die, and it is in their death throes that some of the most intriguing and important processes of the universe are revealed. For example, we now know that many of the atoms in our bodies were once inside stars. These stars exploded at the ends of their lives, recycling their material back into the reservoir of the Galaxy. In this sense, all of us are literally made of recycled “star dust.”

This page titled 1.4: A Tour of the Universe is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by OpenStax via source content that was edited to the style and standards of the LibreTexts platform.

2: GRAVITY AND PLANETARY MOTION

2.1: The Laws of Planetary Motion
2.2: Newton’s Great Synthesis
2.3: Newton’s Universal Law of Gravitation

Chapter 2: Gravity and Planetary Motion — Introduction

How would you find a new planet at the outskirts of our solar system that is too dim to be seen with the unaided eye and is so far away that it moves very slowly among the stars? This was the problem confronting astronomers during the nineteenth century as they tried to pin down a full inventory of our solar system.

If we could look down on the solar system from somewhere out in space, interpreting planetary motions would be much simpler. But the fact is, we must observe the positions of all the other planets from our own moving planet. Scientists of the Renaissance did not know the details of Earth's motions any better than the motions of the other planets. **They had to deduce the nature of planetary motion using only their Earth-based observations of the planets' changing positions in the sky.** To solve this complex problem more fully, better observations and better models of the planetary system were needed.

This page titled 2: Gravity and Planetary Motion is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by OpenStax via source content that was edited to the style and standards of the LibreTexts platform.

2.1: The Laws of Planetary Motion

Learning Objectives

By the end of this section, you will be able to:

- Describe how Tycho Brahe and Johannes Kepler contributed to our understanding of how planets move around the Sun
- Explain Kepler's three laws of planetary motion

At about the time that Galileo was beginning his experiments with falling bodies, the efforts of two other scientists dramatically advanced our understanding of the motions of the planets. These two astronomers were the observer Tycho Brahe and the mathematician Johannes Kepler. Together, they placed the speculations of Copernicus on a sound mathematical basis and paved the way for the work of Isaac Newton in the next century.

Tycho Brahe's Observatory

Three years after the publication of Copernicus' *De Revolutionibus*, Tycho Brahe was born to a family of Danish nobility. He developed an early interest in astronomy and, as a young man, made significant astronomical observations. Among these was a careful study of what we now know was an exploding star that flared up to great brilliance in the night sky. His growing reputation gained him the patronage of the Danish King Frederick II, and at the age of 30, Brahe was able to establish a fine astronomical observatory on the North Sea island of Hven (Figure 3.2). Brahe was the last and greatest of the pre-telescopic observers in Europe.



(a) (b)

Figure 3.2 : Tycho Brahe (1546–1601) and Johannes Kepler (1571–1630). (a) A stylized engraving shows Tycho Brahe using his instruments to measure the altitude of celestial objects above the horizon. The large curved instrument in the foreground allowed him to measure precise angles in the sky. Note that the scene includes hints of the grandeur of Brahe's observatory at Hven. (b) Kepler was a German mathematician and astronomer. His discovery of the basic laws that describe planetary motion placed the heliocentric cosmology of Copernicus on a firm mathematical basis.

At Hven, Brahe made a continuous record of the positions of the Sun, Moon, and planets for almost 20 years. His extensive and precise observations enabled him to note that the positions of the planets varied from those given in published tables, which were based on the work of Ptolemy. These data were extremely valuable, but Brahe didn't have the ability to analyze them and develop a better model than what Ptolemy had published. He was further inhibited because he was an extravagant and cantankerous fellow, and he accumulated enemies among government officials. When his patron, Frederick II, died in 1597, Brahe lost his political base and decided to leave Denmark. He took up residence in Prague, where he became court astronomer to Emperor Rudolf of

Bohemia. There, in the year before his death, Brahe found a most able young mathematician, Johannes Kepler, to assist him in analyzing his extensive planetary data.

Johannes Kepler

Johannes Kepler was born into a poor family in the German province of Württemberg and lived much of his life amid the turmoil of the Thirty Years' War (see Figure 3.2). He attended university at Tubingen and studied for a theological career. There, he learned the principles of the Copernican system and became converted to the heliocentric hypothesis. Eventually, Kepler went to Prague to serve as an assistant to Brahe, who set him to work trying to find a satisfactory theory of planetary motion—one that was compatible with the long series of observations made at Hven. Brahe was reluctant to provide Kepler with much material at any one time for fear that Kepler would discover the secrets of the universal motion by himself, thereby robbing Brahe of some of the glory. Only after Brahe's death in 1601 did Kepler get full possession of the priceless records. Their study occupied most of Kepler's time for more than 20 years.

Through his analysis of the motions of the planets, Kepler developed a series of principles, now known as *Kepler's three laws,* which described the behavior of planets based on their paths through space. The first two laws of planetary motion were published in 1609 in *The New Astronomy*. Their discovery was a profound step in the development of modern science.

The First Two Laws of Planetary Motion

The path of an object through space is called its orbit. Kepler initially assumed that the orbits of planets were circles, but doing so did not allow him to find orbits that were consistent with Brahe's observations. Working with the data for Mars, he eventually discovered that the orbit of that planet had the shape of a somewhat flattened circle, or ellipse. Next to the circle, the ellipse is the simplest kind of closed curve, belonging to a family of curves known as *conic sections* (Figure 3.3).

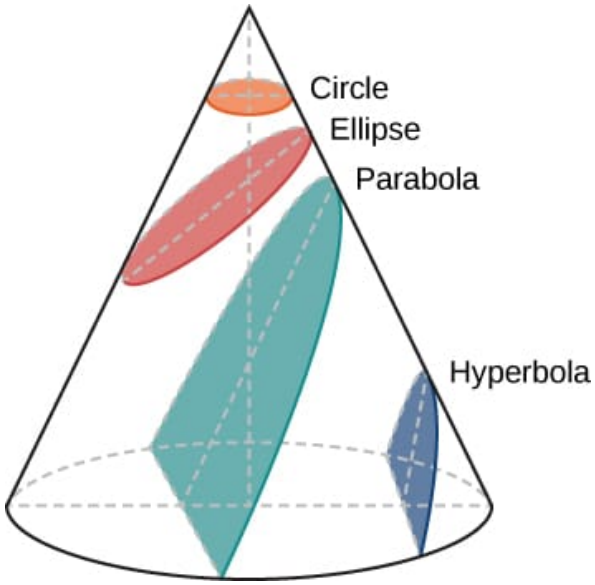


Figure 3.3 : Conic Sections. The circle, ellipse, parabola, and hyperbola are all formed by the intersection of a plane with a cone. This is why such curves are called conic sections.

You might recall from math classes that in a circle, the center is a special point. The distance from the center to anywhere on the circle is exactly the same. In an ellipse, the sum of the distance from two special points inside the ellipse to any point on the ellipse is always the same. These two points inside the ellipse are called its foci (singular: focus), a word invented for this purpose by Kepler.

This property suggests a simple way to draw an ellipse (Figure 3.4). We wrap the ends of a loop of string around two tacks pushed through a sheet of paper into a drawing board, so that the string is slack. If we push a pencil against the string, making the string taut, and then slide the pencil against the string all around the tacks, the curve that results is an ellipse. At any point where the pencil may be, the sum of the distances from the pencil to the two tacks is a constant length—the length of the string. The tacks are at the two foci of the ellipse.

The widest diameter of the ellipse is called its major axis. Half this distance—that is, the distance from the center of the ellipse to one end—is the semimajor axis, which is usually used to specify the size of the ellipse. For example, the semimajor axis of the orbit of Mars, which is also the planet's average distance from the Sun, is 228 million kilometers.

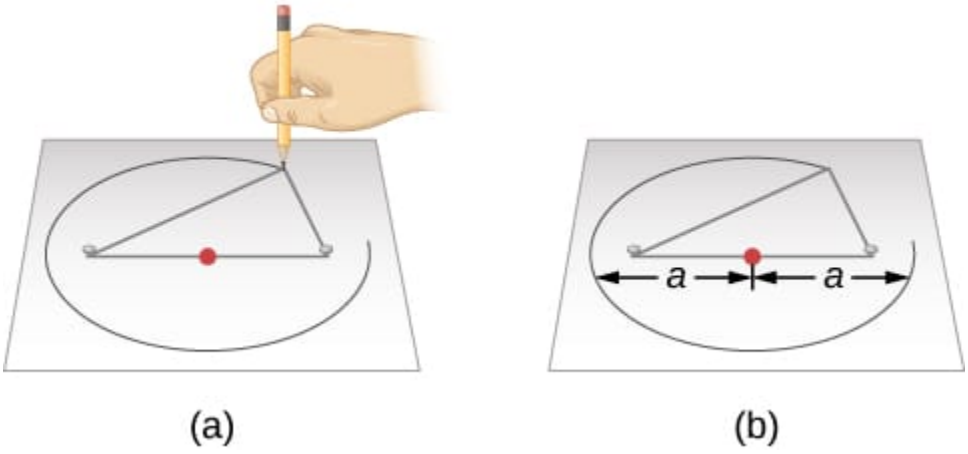


Figure 3.4 : Drawing an Ellipse. (a) We can construct an ellipse by pushing two tacks (the white objects) into a piece of paper on a drawing board, and then looping a string around the tacks. Each tack represents a focus of the ellipse, with one of the tacks being the Sun. Stretch the string tight using a pencil, and then move the pencil around the tacks. The length of the string remains the same, so that the sum of the distances from any point on the ellipse to the foci is always constant. (b) In this illustration, each semimajor axis is denoted by *a*. The distance 2*a* is called the major axis of the ellipse.

The shape (roundness) of an ellipse depends on how close together the two foci are, compared with the major axis. The ratio of the distance between the foci to the length of the major axis is called the eccentricity of the ellipse.

If the foci (or tacks) are moved to the same location, then the distance between the foci would be zero. This means that the eccentricity is zero and the ellipse is just a circle; thus, a circle can be called an ellipse of zero eccentricity. In a circle, the semimajor axis would be the radius.

Next, we can make ellipses of various elongations (or extended lengths) by varying the spacing of the tacks (as long as they are not farther apart than the length of the string). The greater the eccentricity, the more elongated is the ellipse, up to a maximum eccentricity of 1.01.0, when the ellipse becomes "flat," the other extreme from a circle.

The size and shape of an ellipse are completely specified by its semimajor axis and its eccentricity. Using Brahe's data, Kepler found that Mars has an elliptical orbit, with the Sun at one focus (the other focus is empty). The eccentricity of the orbit of Mars is only about 0.10.1; its orbit, drawn to scale, would be practically indistinguishable from a circle, but the difference turned out to be critical for understanding planetary motions.

Kepler generalized this result in his first law and said that *the orbits of all the planets are ellipses*. Here was a decisive moment in the history of human thought: it was not necessary to have only circles in order to have an acceptable cosmos. The universe could be a bit more complex than the Greek philosophers had wanted it to be.

Kepler's second law deals with the speed with which each planet moves along its ellipse, also known as its orbital speed. Working with Brahe's observations of Mars, Kepler discovered that the planet speeds up as it comes closer to the Sun and slows down as it pulls away from the Sun. He expressed the precise form of this relationship by imagining that the Sun and Mars are connected by a straight, elastic line. When Mars is closer to the Sun (positions 1 and 2 in Figure 3.5), the elastic line is not stretched as much, and the planet moves rapidly. Farther from the Sun, as in positions 3 and 4, the line is stretched a lot, and the planet does not move so fast. As Mars travels in its elliptical orbit around the Sun, the elastic line sweeps out areas of the ellipse as it moves (the colored regions in our figure). Kepler found that in equal intervals of time (t), the areas swept out in space by this imaginary line are always equal; that is, the area of the region B from 1 to 2 is the same as that of region A from 3 to 4.

If a planet moves in a circular orbit, the elastic line is always stretched the same amount and the planet moves at a constant speed around its orbit. But, as Kepler discovered, in most orbits that speed of a planet orbiting its star (or moon orbiting its planet) tends to vary because the orbit is elliptical.

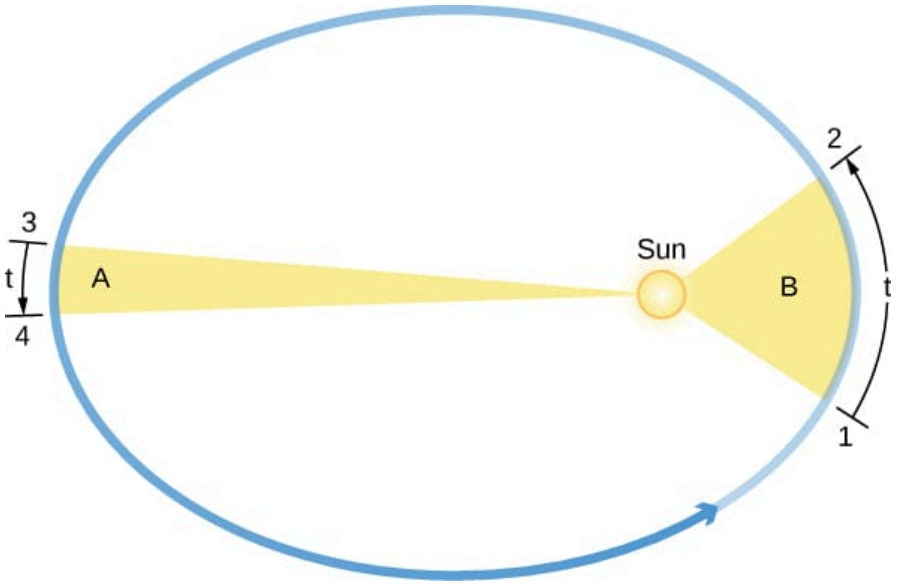


Figure 3.5 : Kepler's Second Law: The Law of Equal Areas. The orbital speed of a planet traveling around the Sun (the circular object inside the ellipse) varies in such a way that in equal intervals of time (t), a line between the Sun and a planet sweeps out equal areas (A and B). Note that the eccentricities of the planets' orbits in our solar system are substantially less than shown here.

Link to Learning

The Kepler's Second Law demonstrator from CCNY's ScienceSims project shows how an orbiting planet sweeps out the same area in the same time.

Kepler's Third Law

Kepler's first two laws of planetary motion describe the shape of a planet's orbit and allow us to calculate the speed of its motion at any point in the orbit. Kepler was pleased to have discovered such fundamental rules, but they did not satisfy his quest to fully understand planetary motions. He wanted to know why the orbits of the planets were spaced as they are and to find a mathematical pattern in their movements—a "harmony of the spheres" as he called it. For many years he worked to discover mathematical relationships governing planetary spacing and the time each planet took to go around the Sun.

In 1619, Kepler discovered a basic relationship to relate the planets' orbits to their relative distances from the Sun. We define a planet's orbital period, (***P***), as the time it takes a planet to travel once around the Sun. Also, recall that a planet's semimajor axis, *a,* is equal to its average distance from the Sun. The relationship, now known as *Kepler's third law*, says that a planet's orbital period squared is proportional to the semimajor axis of its orbit cubed, or

P2∝a3P2∝a3

When *P* (the orbital period) is measured in years, and *a* is expressed in a quantity known as an astronomical unit (AU), the two sides of the formula are not only proportional but equal. One AU is the average distance between Earth and the Sun and is approximately equal to 1.5×1081.5×108 kilometers. In these units,

P2=a3P2=a3

Kepler's third law applies to all objects orbiting the Sun, including Earth, and provides a means for calculating their relative distances from the Sun from the time they take to orbit. Let's look at a specific example to illustrate how useful Kepler's third law is.

For instance, suppose you time how long Mars takes to go around the Sun (in Earth years). Kepler's third law can then be used to calculate Mars' average distance from the Sun. Mars' orbital period (1.88 Earth years) squared, or P2P2, is 1.882=3.531.882=3.53, and according to the equation for Kepler's third law, this equals the cube of its semimajor axis, or a3a3. So what number must be cubed to give 3.53? The answer is 1.521.52(since 1.52×1.52×1.52=3.53)(since 1.52×1.52×1.52=3.53). Thus, Mars' semimajor axis in

astronomical units must be 1.52 AU. In other words, to go around the Sun in a little less than two years, Mars must be about 50% (half again) as far from the Sun as Earth is.

Example 3.1: Calculating Periods

Imagine an object is traveling around the Sun. What would be the orbital period of the object if its orbit has a semimajor axis of 50 AU?

Solution

From Kepler's third law, we know that (when we use units of years and AU)

P2=a3P2=a3

If the object's orbit has a semimajor axis of 50 AU (a = 50), we can cube 50 and then take the square root of the result to get P:

P=a3P=50×50×50=125,000=353.6yearsP=a3P=50×50×50=125,000=353.6years

Therefore, the orbital period of the object is about 350 years. This would place our hypothetical object beyond the orbit of Pluto.

Exercise 2.1.1

What would be the orbital period of an asteroid (a rocky chunk between Mars and Jupiter) with a semimajor axis of 3 AU?

Answer available at the end of this section.

Kepler's three laws of planetary motion can be summarized as follows:

- Kepler's first law: Each planet moves around the Sun in an orbit that is an ellipse, with the Sun at one focus of the ellipse.
- Kepler's second law: The straight line joining a planet and the Sun sweeps out equal areas in space in equal intervals of time.
- Kepler's third law: The square of a planet's orbital period is directly proportional to the cube of the semimajor axis of its orbit.

Kepler's three laws provide a precise geometric description of planetary motion within the framework of the Copernican system. With these tools, it was possible to calculate planetary positions with greatly improved precision. Still, Kepler's laws are purely descriptive: they do not help us understand what forces of nature constrain the planets to follow this particular set of rules. That step was left to Isaac Newton.

Example 3.2: Applying Kepler's Third Law

Using the orbital periods and semimajor axes for Venus and Earth that are provided here, calculate p2p2 and a3a3, and verify that they obey Kepler's third law. Venus' orbital period is 0.62 year, and its semimajor axis is 0.72 AU. Earth's orbital period is 1.00 year, and its semimajor axis is 1.00 AU.

Solution

We can use the equation for Kepler's third law, $P^2 \propto a^3$. For Venus, P2=0.62×0.62=0.38P2=0.62×0.62=0.38 and a3=0.72×0.72×0.72=.037a3=0.72×0.72×0.72=.037 (rounding numbers sometimes causes minor discrepancies like this). The square of the orbital period (0.38) approximates the cube of the semimajor axis (0.37). Therefore, Venus obeys Kepler's third law. For Earth, P2=1.00×1.00=1.00P2=1.00×1.00=1.00 and a3=1.00×1.00×1.00=1.00a3=1.00×1.00×1.00=1.00. The square of the orbital period (1.00) approximates (in this case, equals) the cube of the semimajor axis (1.00). Therefore, Earth obeys Kepler's third law.

Exercise 2.1.1

Using the orbital periods and semimajor axes for Saturn and Jupiter that are provided here, calculate P^2 and a^3, and verify that they obey Kepler's third law. Saturn's orbital period is 29.46 years, and its semimajor axis is 9.54 AU. Jupiter's orbital period is 11.86 years, and its semimajor axis is 5.20 AU.

Answer available at the end of this section.

Link to Learning

In honor of the scientist who first devised the laws that govern the motions of planets, the team that built the first spacecraft to search for planets orbiting other stars decided to name the probe "Kepler." Visit NASA's Kepler website to learn more about Johannes Kepler's life and his laws of planetary motion. NASA's Kepler website and follow the links that interest you.

Answers

Exercise 2.1.1

P=3×3×3=27=5.2yearsP=3×3×3=27=5.2years

Exercise 2.1.1

For Saturn, P2=29.46×29.46=867.9P2=29.46×29.46=867.9 and a3=9.54×9.54×9.54=868.3a3=9.54×9.54×9.54=868.3. The square of the orbital period (867.9) approximates the cube of the semimajor axis (868.3). Therefore, Saturn obeys Kepler's third law.

This page titled 2.1: The Laws of Planetary Motion is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by OpenStax via source content that was edited to the style and standards of the LibreTexts platform.

2.2: Newton's Great Synthesis

Learning Objectives

By the end of this section, you will be able to:

- Describe Newton's three laws of motion
- Explain how Newton's three laws of motion relate to momentum
- Define mass, volume, and density and how they differ
- Define angular momentum

It was the genius of Isaac Newton that found a conceptual framework that completely explained the observations and rules assembled by Galileo, Brahe, Kepler, and others. Newton was born in Lincolnshire, England, in the year after Galileo's death (Figure 3.6). Against the advice of his mother, who wanted him to stay home and help with the family farm, he entered Trinity College at Cambridge in 1661 and eight years later was appointed professor of mathematics. Among Newton's contemporaries in England were architect Christopher Wren, authors Aphra Behn and Daniel Defoe, and composer G. F. Handel.

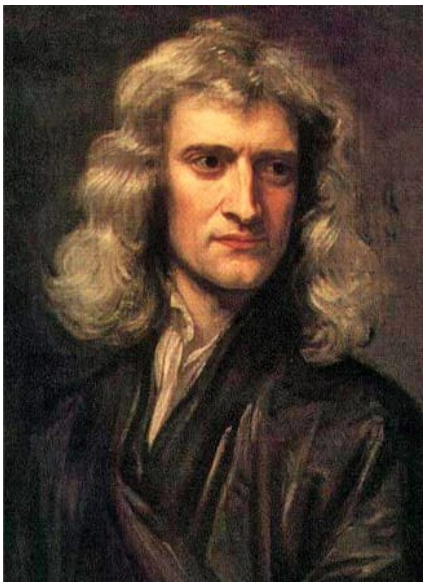


Figure 3.6 : Isaac Newton (1643–1727), 1689 Portrait by Sir Godfrey Kneller. Isaac Newton's work on the laws of motion, gravity, optics, and mathematics laid the foundations for much of physical science.

Newton's Laws of Motion

As a young man in college, Newton became interested in natural philosophy, as science was then called. He worked out some of his first ideas on machines and optics during the plague years of 1665 and 1666, when students were sent home from college. Newton, a moody and often difficult man, continued to work on his ideas in private, even inventing new mathematical tools to help him deal with the complexities involved. Eventually, his friend Edmund Halley (profiled in Comets and Asteroids: Debris of the Solar System) prevailed on him to collect and publish the results of his remarkable investigations on motion and gravity. The result was a volume that set out the underlying system of the physical world, *Philosophiae Naturalis Principia Mathematica*. The *Principia*, as the book is generally known, was published at Halley's expense in 1687.

At the very beginning of the *Principia*, Newton proposes three laws that would govern the motions of all objects:

- Newton's first law: Every object will continue to be in a state of rest or move at a constant speed in a straight line unless it is compelled to change by an outside force.
- Newton's second law: The change of motion of a body is proportional to and in the direction of the force acting on it.
- Newton's third law: For every action there is an equal and opposite reaction (*or:* the mutual actions of two bodies upon each

other are always equal and act in opposite directions).

In the original Latin, the three laws contain only 59 words, but those few words set the stage for modern science. Let us examine them more carefully.

Interpretation of Newton's Laws

Newton's first law is a restatement of one of Galileo's discoveries, called the *conservation of momentum*. The law states that in the absence of any outside influence, there is a measure of a body's motion, called its momentum, that remains unchanged. You may have heard the term momentum used in everyday expressions, such as "This bill in Congress has a lot of momentum; it's going to be hard to stop."

Newton's first law is sometimes called the *law of inertia*, where inertia is the tendency of objects (and legislatures) to keep doing what they are already doing. In other words, a stationary object stays put, and a moving object keeps moving unless some force intervenes.

Let's define the precise meaning of momentum—it depends on three factors: (1) speed—how fast a body moves (zero if it is stationary), (2) the direction of its motion, and (3) its mass—a measure of the amount of matter in a body, which we will discuss later. Scientists use the term velocity to describe the speed and direction of motion. For example, 20 kilometers per hour due south is velocity, whereas 20 kilometers per hour just by itself is speed. Momentum then can be defined as an object's mass times its velocity.

It's not so easy to see this rule in action in the everyday world because of the many forces acting on a body at any one time. One important force is friction, which generally slows things down. If you roll a ball along the sidewalk, it eventually comes to a stop because the sidewalk exerts a rubbing force on the ball. But in the space between the stars, where there is so little matter that friction is insignificant, objects can in fact continue to move (to coast) indefinitely.

The momentum of a body can change only under the action of an outside influence. Newton's second law expresses *force* in terms of its ability to change momentum with time. A force (a push or a pull) has both size and direction. When a force is applied to a body, the momentum changes in the direction of the applied force. This means that a force is required to change either the speed or the direction of a body, or both—that is, to start it moving, to speed it up, to slow it down, to stop it, or to change its direction.

As you learned in Observing the Sky: The Birth of Astronomy, the rate of change in an object's velocity is called *acceleration*. Newton showed that the acceleration of a body was proportional to the force being applied to it. Suppose that after a long period of reading, you push an astronomy book away from you on a long, smooth table. (We use a smooth table so we can ignore friction.) If you push the book steadily, it will continue to speed up as long as you are pushing it. The harder you push the book, the larger its acceleration will be. How much a force will accelerate an object is also determined by the object's mass. If you kept pushing a pen with the same force with which you pushed the textbook, the pen—having less mass—would be accelerated to a greater speed.

Newton's third law is perhaps the most profound of the rules he discovered. Basically, it is a generalization of the first law, but it also gives us a way to define mass. If we consider a system of two or more objects isolated from outside influences, Newton's first law says that the total momentum of the objects should remain constant. Therefore, any change of momentum within the system must be balanced by another change that is equal and opposite so that the momentum of the entire system is not changed.

This means that forces in nature do not occur alone: we find that in each situation there is always a *pair* of forces that are equal to and opposite each other. If a force is exerted on an object, it must be exerted by something else, and the object will exert an equal and opposite force back on that something. We can look at a simple example to demonstrate this.

Suppose that a daredevil astronomy student—and avid skateboarder—wants to jump from his second-story dorm window onto his board below (we don't recommend trying this!). The force pulling him down after jumping (as we will see in the next section) is the force of gravity between him and Earth. Both he and Earth must experience the same total change of momentum because of the influence of these mutual forces. So, both the student and Earth are accelerated by each other's pull. However, the student does much more of the moving. Because Earth has enormously greater mass, it can experience the same change of momentum by accelerating only a very small amount. Things fall toward Earth all the time, but the acceleration of our planet as a result is far too small to be measured.

A more obvious example of the mutual nature of forces between objects is familiar to all who have batted a baseball. The recoil you feel as you swing your bat shows that the ball exerts a force on it during the impact, just as the bat does on the ball. Similarly, when a rifle you are bracing on your shoulder is discharged, the force pushing the bullet out of the muzzle is equal to the force pushing backward upon the gun and your shoulder.

This is the principle behind jet engines and rockets: the force that discharges the exhaust gases from the rear of the rocket is accompanied by the force that pushes the rocket forward. The exhaust gases need not push against air or Earth; a rocket actually operates best in a vacuum (Figure 3.7).



Figure 3.7 : Demonstrating Newton's Third Law. The U.S. Space Shuttle (here launching *Discovery*), powered by three fuel engines burning liquid oxygen and liquid hydrogen, with two solid fuel boosters, demonstrates Newton's third law. (credit: modification of work by NASA)

Link to Learning

For more about Isaac Newton's life and work, check out this timeline page with snapshots from his career, produced by the British Broadcasting Corporation (BBC).

Mass, Volume, and Density

Before we go on to discuss Newton's other work, we want to take a brief look at some terms that will be important to sort out clearly. We begin with mass, which is a measure of the amount of material within an object.

The *volume* of an object is the measure of the physical space it occupies. Volume is measured in cubic units, such as cubic centimeters or liters. The volume is the "size" of an object. A penny and an inflated balloon may both have the same mass, but they have very different volumes. The reason is that they also have very different *densities*, which is a measure of how much mass there is per unit volume. Specifically, density is the mass divided by the volume. Note that in everyday language we often use "heavy" and "light" as indications of density (rather than weight) as, for instance, when we say that iron is heavy or that whipped cream is light.

The units of density that will be used in this book are grams per cubic centimeter (g/cm^3).[1] If a block of some material has a mass of 300 grams and a volume of 100 cm^3, its density is 3 g/cm^3. Familiar materials span a considerable range in density, from artificial materials such as plastic insulating foam (less than 0.1 g/cm^3) to gold (19.3 g/cm^3). Table 3.1 gives the densities of some familiar materials. In the astronomical universe, much more remarkable densities can be found, all the way from a comet's tail (10^{-16} g/cm^3) to a collapsed "star corpse" called a neutron star (10^{15} g/cm^3).

Table 3.1: Densities of Common Materials

Material	Density (g/cm^3)
Gold	19.3
Lead	11.3
Iron	7.9
Earth (bulk)	5.5
Rock (typical)	2.5
Water	1
Wood (typical)	0.8
Insulating foam	0.1
Silica gel	0.02

To sum up, mass is *how much*, volume is *how big*, and density is *how tightly packed*.

Link to Learning

You can play with a simple animation demonstrating the relationship between the concepts of density, mass, and volume, and find out why objects like wood float in water.

Angular Momentum

A concept that is a bit more complex, but important for understanding many astronomical objects, is angular momentum, which is a measure of the rotation of a body as it revolves around some fixed point (an example is a planet orbiting the Sun). The angular momentum of an object is defined as the product of its mass, its velocity, and its distance from the fixed point around which it revolves.

If these three quantities remain constant—that is, if the motion of a particular object takes place at a constant velocity at a fixed distance from the spin center—then the angular momentum is also a constant. Kepler's second law is a consequence of the *conservation of angular momentum*. As a planet approaches the Sun on its elliptical orbit and the distance to the spin center decreases, the planet speeds up to conserve the angular momentum. Similarly, when the planet is farther from the Sun, it moves more slowly.

The conservation of angular momentum is illustrated by figure skaters, who bring their arms and legs in to spin more rapidly, and extend their arms and legs to slow down (Figure 3.8). You can duplicate this yourself on a well-oiled swivel stool by starting yourself spinning slowly with your arms extended and then pulling your arms in. Another example of the conservation of angular momentum is a shrinking cloud of dust or a star collapsing on itself (both are situations that you will learn about as you read on). As material moves to a lesser distance from the spin center, the speed of the material increases to conserve angular momentum.

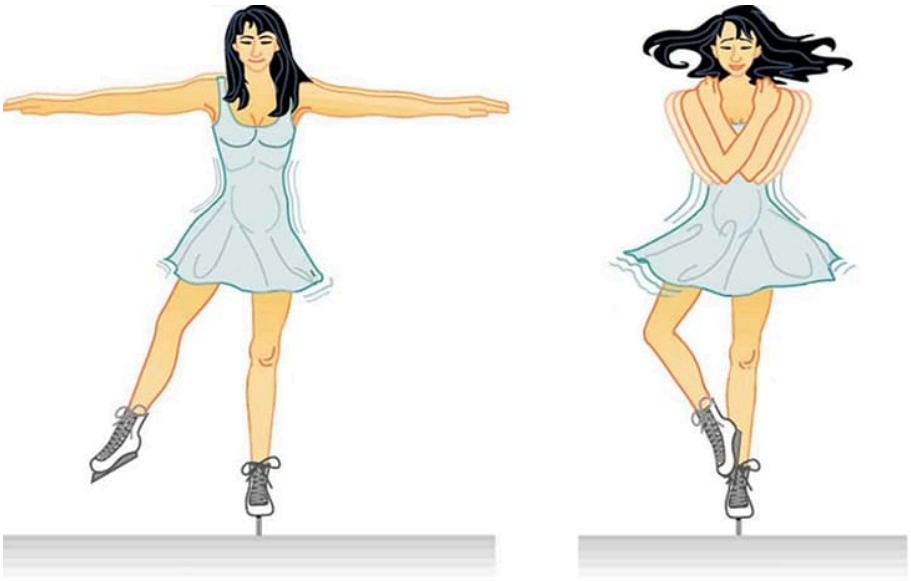


Figure 3.8 : Conservation of Angular Momentum. When a spinning figure skater brings in her arms, their distance from her spin center is smaller, so her speed increases. When her arms are out, their distance from the spin center is greater, so she slows down.

Footnotes

- 1Generally we use standard metric (or SI) units in this book. The proper metric unit of density in that system is kg/m^3. But to most people, g/cm^3 provides a more meaningful unit because the density of water is exactly 1 g/cm^3, and this is useful information for comparison. Density expressed in g/cm^3 is sometimes called specific density or specific weight.

This page titled 2.2: Newton's Great Synthesis is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by OpenStax via source content that was edited to the style and standards of the LibreTexts platform.

2.3: Newton's Universal Law of Gravitation

Learning Objectives

By the end of this section, you will be able to:

- Explain what determines the strength of gravity
- Describe how Newton's universal law of gravitation extends our understanding of Kepler's laws

Newton's laws of motion show that objects at rest will stay at rest and those in motion will continue moving uniformly in a straight line unless acted upon by a force. Thus, it is the *straight line* that defines the most natural state of motion. But the planets move in ellipses, not straight lines; therefore, some force must be bending their paths. That force, Newton proposed, was gravity.

In Newton's time, gravity was something associated with Earth alone. Everyday experience shows us that Earth exerts a gravitational force upon objects at its surface. If you drop something, it accelerates toward Earth as it falls. Newton's insight was that Earth's gravity might extend as far as the Moon and produce the force required to curve the Moon's path from a straight line and keep it in its orbit. He further hypothesized that gravity is not limited to Earth, but that there is a general force of attraction between all material bodies. If so, the attractive force between the Sun and each of the planets could keep them in their orbits. (This may seem part of our everyday thinking today, but it was a remarkable insight in Newton's time.)

Once Newton boldly hypothesized that there was a universal attraction among all bodies everywhere in space, he had to determine the exact nature of the attraction. The precise mathematical description of that gravitational force had to dictate that the planets move exactly as Kepler had described them to (as expressed in Kepler's three laws). Also, that gravitational force had to predict the correct behavior of falling bodies on Earth, as observed by Galileo. How must the force of gravity depend on distance in order for these conditions to be met?

The answer to this question required mathematical tools that had not yet been developed, but this did not deter Isaac Newton, who invented what we today call calculus to deal with this problem. Eventually he was able to conclude that the magnitude of the force of gravity must decrease with increasing distance between the Sun and a planet (or between any two objects) in proportion to the inverse square of their separation. In other words, if a planet were twice as far from the Sun, the force would be (12)2(12)2, or 1414 as large. Put the planet three times farther away, and the force is (13)2(13)2, or 1919 as large.

Newton also concluded that the gravitational attraction between two bodies must be proportional to their masses. The more mass an object has, the stronger the pull of its gravitational force. The gravitational attraction between any two objects is therefore given by one of the most famous equations in all of science:

Fgravity=GM1M2R2Fgravity=GM1M2R2

where F_{gravity} is the gravitational force between two objects, M_1 and M_2 are the masses of the two objects, and R is their separation. G is a constant number known as the *universal gravitational constant*, and the equation itself symbolically summarizes Newton's *universal law of gravitation*. With such a force and the laws of motion, Newton was able to show mathematically that the only orbits permitted were exactly those described by Kepler's laws.

Newton's universal law of gravitation works for the planets, but is it really universal? The gravitational theory should also predict the observed acceleration of the Moon toward Earth as it orbits Earth, as well as of any object (say, an apple) dropped near Earth's surface. The falling of an apple is something we can measure quite easily, but can we use it to predict the motions of the Moon?

Recall that according to Newton's second law, forces cause acceleration. Newton's universal law of gravitation says that the force acting upon (and therefore the acceleration of) an object toward Earth should be inversely proportional to the square of its distance from the center of Earth. Objects like apples at the surface of Earth, at a distance of one Earth-radius from the center of Earth, are observed to accelerate downward at 9.8 meters per second per second (9.8 m/s2)(9.8 m/s2).

It is this force of gravity on the surface of Earth that gives us our sense of *weight*. Unlike your mass, which would remain the same on any planet or moon, your weight depends on the local force of gravity. So you would weigh less on Mars and the Moon than on Earth, even though there is no change in your mass. (Which means you would still have to go easy on the desserts in the college

cafeteria when you got back!)

The Moon is 60 Earth radii away from the center of Earth. If gravity (and the acceleration it causes) gets weaker with distance squared, the acceleration the Moon experiences should be a lot less than for the apple. The acceleration should be (1/60)2=1/3600(1/60)2=1/3600 (or 3600 times less—about 0.00272 m/s20.00272 m/s2). This is precisely the observed acceleration of the Moon in its orbit. (As we shall see, the Moon does not fall *to* Earth with this acceleration, but falls *around* Earth.) Imagine the thrill Newton must have felt to realize he had discovered, and verified, a law that holds for Earth, apples, the Moon, and, as far as he knew, everything in the universe.

Example 3.3: Calculating Weight

By what factor would a person's weight at the surface of Earth change if Earth had its present mass but eight times its present volume?

Solution

With eight times the volume, Earth's radius would double. This means the gravitational force at the surface would reduce by a factor of (12)2=14 (12)2=14 , so a person would weigh only one-fourth as much.

Exercise 2.3.1

By what factor would a person's weight at the surface of Earth change if Earth had its present size but only one-third its present mass?

Answer available at the end of this section.

Gravity is a "built-in" property of mass. Whenever there are masses in the universe, they will interact via the force of gravitational attraction. The more mass there is, the greater the force of attraction. Here on Earth, the largest concentration of mass is, of course, the planet we stand on, and its pull dominates the gravitational interactions we experience. But everything with mass attracts everything else with mass anywhere in the universe.

Newton's law also implies that gravity never becomes zero. It quickly gets weaker with distance, but it continues to act to some degree no matter how far away you get. The pull of the Sun is stronger at Mercury than at Pluto, but it can be felt far beyond Pluto, where astronomers have good evidence that it continuously makes enormous numbers of smaller icy bodies move around huge orbits. And the Sun's gravitational pull joins with the pull of billions of others stars to create the gravitational pull of our Milky Way Galaxy. That force, in turn, can make other smaller galaxies orbit around the Milky Way, and so on.

Why is it then, you may ask, that the astronauts aboard the Space Shuttle appear to have no gravitational forces acting on them when we see images on television of the astronauts and objects floating in the spacecraft? After all, the astronauts in the shuttle are only a few hundred kilometers above the surface of Earth, which is not a significant distance compared to the size of Earth, so gravity is certainly not a great deal weaker that much farther away. The astronauts feel "weightless" (meaning that they don't feel the gravitational force acting on them) for the same reason that passengers in an elevator whose cable has broken or in an airplane whose engines no longer work feel weightless: they are falling (Figure 3.9).[2]



Figure 3.9 : Astronauts in Free Fall. While in space, astronauts are falling freely, so they experience “weightlessness.” Clockwise from top left: Tracy Caldwell Dyson (NASA), Naoko Yamazaki (JAXA), Dorothy Metcalf-Lindenburger (NASA), and Stephanie Wilson (NASA). (credit: NASA)

When *falling*, they are in free fall and accelerate at the same rate as everything around them, including their spacecraft or a camera with which they are taking photographs of Earth. When doing so, astronauts experience no additional forces and therefore feel “weightless.” Unlike the falling elevator passengers, however, the astronauts are falling *around* Earth, not *to* Earth; as a result they will continue to fall and are said to be “in orbit” around Earth (see the next section for more about orbits).

Orbital Motion and Mass

Kepler’s laws describe the orbits of the objects whose motions are described by Newton’s laws of motion and the law of gravity. Knowing that gravity is the force that attracts planets toward the Sun, however, allowed Newton to rethink Kepler’s third law. Recall that Kepler had found a relationship between the orbital period of a planet’s revolution and its distance from the Sun. But Newton’s formulation introduces the additional factor of the masses of the Sun (M_1) and the planet (M_2), both expressed in units of the Sun’s mass. Newton’s universal law of gravitation can be used to show mathematically that this relationship is actually

a3=(M1+M2)×P2a3=(M1+M2)×P2

where a is the semimajor axis and P is the orbital period.

How did Kepler miss this factor? In units of the Sun’s mass, the mass of the Sun is 1, and in units of the Sun’s mass, the mass of a typical planet is a negligibly small factor. This means that the sum of the Sun’s mass and a planet’s mass, (M1+M2)(M1+M2), is very, very close to 1. This makes Newton’s formula appear almost the same as Kepler’s; the tiny mass of the planets compared to the Sun is the reason that Kepler did not realize that both masses had to be included in the calculation. There are many situations in astronomy, however, in which we *do* need to include the two mass terms—for example, when two stars or two galaxies orbit each other.

Including the mass term allows us to use this formula in a new way. If we can measure the motions (distances and orbital periods) of objects acting under their mutual gravity, then the formula will permit us to deduce their masses. For example, we can calculate the mass of the Sun by using the distances and orbital periods of the planets, or the mass of Jupiter by noting the motions of its moons.

Indeed, Newton’s reformulation of Kepler’s third law is one of the most powerful concepts in astronomy. Our ability to deduce the masses of objects from their motions is key to understanding the nature and evolution of many astronomical bodies. We will use this law repeatedly throughout this text in calculations that range from the orbits of comets to the interactions of galaxies.

Example 3.4: Calculating the Effects of Gravity

A planet like Earth is found orbiting its star at a distance of 1 AU in 0.710.71 Earth-year. Can you use Newton's version of Kepler's third law to find the mass of the star? (Remember that compared to the mass of a star, the mass of an earthlike planet can be considered negligible.)

Solution

In the formula a3=(M1+M2)×P2 a3=(M1+M2)×P2 , the factor M1+M2 M1+M2 would now be approximately equal to M1M1 (the mass of the star), since the planet's mass is so small by comparison. Then the formula becomes a3=M1×P2a3=M1×P2, and we can solve for M1M1:

M1=a3P2M1=a3P2

Since a=1,a3=1a=1,a3=1, so

M1=1P2=10.712=10.5=2M1=1P2=10.712=10.5=2

So the mass of the star is twice the mass of our Sun. (Remember that this way of expressing the law has units in terms of Earth and the Sun, so masses are expressed in units of the mass of our Sun.)

Exercise 2.3.1

Suppose a star with twice the mass of our Sun had an earthlike planet that took 4 years to orbit the star. At what distance (semimajor axis) would this planet orbit its star?

Answer available at the end of this section.

Link to Learning

You might like to try a simulation that lets you move the Sun, Earth, Moon, and space station to see the effects of changing their distances on their gravitational forces and orbital paths. You can even turn off gravity and see what happens.

Footnotes

- 2In the film *Apollo 13*, the scenes in which the astronauts were "weightless" were actually filmed in a falling airplane. As you might imagine, the plane fell for only short periods before the engines engaged again.

Answers

Exercise 2.3.1

With one-third its present mass, the gravitational force at the surface would reduce by a factor of 1313, so a person would weight only one-third as much.

Exercise 2.3.1

Again, we can neglect the mass of the planet. So M1=2 M1=2 and P=4P=4 years. The formula is a3=M1×P2 a3=M1×P2 , so a3=2×42=2×16=32 a3=2×42=2×16=32. So *a* is the cube root of 32. To find this, you can just ask Google, "What is the cube root of 32?" and get the answer 3.2 AU.

This page titled 2.3: Newton's Universal Law of Gravitation is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by OpenStax via source content that was edited to the style and standards of the LibreTexts platform.

3: OUR SOLAR SYSTEM

3.1: Overview of Our Planetary System

3.2: Composition and Structure of Planets

3.3: Formation of Planetary Systems

3.3.1: Origin of the Solar System

3.3.2: Formation of the Solar System

3.3.3: Comparison with Other Planetary Systems

Chapter 3: Our Solar System — Introduction

Our solar system contains a remarkable variety of worlds, from small rocky planets to giant planets surrounded by extensive systems of moons and rings. Despite these differences, the planets share patterns in their composition, structure, and motion that provide important clues about how the solar system formed.

In this chapter, we will compare the major types of planets and examine the characteristics that distinguish terrestrial worlds from the giant planets. We will then use these observations as evidence for the origin and formation of the solar system. The leading model describes the solar system as forming about 4.5 billion years ago from a rotating cloud of gas and dust. As this material collapsed, the Sun formed near the center while the remaining material in a surrounding disk gradually assembled into planets and other smaller bodies.

Our solar system is also only one example of a planetary system. Discoveries of planets around other stars allow astronomers to test ideas about planetary formation and show us that planetary systems can take many different forms. By comparing these systems with our own, we can better understand both how our solar system formed and what makes it distinctive.

This page titled 3: Our Solar System is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Robert Wagner.

3.1: Overview of Our Planetary System

Learning Objectives

By the end of this section, you will be able to:

- Describe how the objects in our solar system are identified, explored, and characterized
- Describe the types of small bodies in our solar system, their locations, and how they formed
- Model the solar system with distances from everyday life to better comprehend distances in space

The solar system[1] consists of the Sun and many smaller objects: the planets, their moons and rings, and such "debris" as asteroids, comets, and dust. Decades of observation and spacecraft exploration have revealed that most of these objects formed together with the Sun about 4.5 billion years ago. They represent clumps of material that condensed from an enormous cloud of gas and dust. The central part of this cloud became the Sun, and a small fraction of the material in the outer parts eventually formed the other objects.

During the past 50 years, we have learned more about the solar system than anyone imagined before the space age. In addition to gathering information with powerful new telescopes, we have sent spacecraft directly to many members of the planetary system. (Planetary astronomy is the only branch of our science in which we can, at least vicariously, travel to the objects we want to study.) With evocative names such as *Voyager*, *Pioneer*, *Curiosity*, and *Pathfinder*, our robot explorers have flown past, orbited, or landed on every planet, returning images and data that have dazzled both astronomers and the public. In the process, we have also investigated two dwarf planets, hundreds of fascinating moons, four ring systems, a dozen asteroids, and several comets (smaller members of our solar system that we will discuss later).

Our probes have penetrated the atmosphere of Jupiter and landed on the surfaces of Venus, Mars, our Moon, Saturn's moon Titan, the asteroids Eros, Itokawa, Ryugu, and Bennu, and the Comet Churyumov-Gerasimenko (usually referred to as 67P). Humans have set foot on the Moon and returned samples of its surface soil for laboratory analysis (Figure 7.2). We have flown a helicopter drone on Mars. We have even discovered other places in our solar system that might be able to support some kind of life.



Figure 7.2 : Astronauts on the Moon. The lunar lander and surface rover from the Apollo 15 mission are seen in this view of the one place beyond Earth that has been explored directly by humans. (credit: modification of work by David R. Scott, NASA)

Link to Learning

View this gallery of NASA images that trace the history of the Apollo mission.

An Inventory

The Sun, a star that is brighter than about 80% of the stars in the Galaxy, is by far the most massive member of the solar system, as shown in Table 7.1. It is an enormous ball about 1.4 million kilometers in diameter, with surface layers of incandescent gas and an interior temperature of millions of degrees. The Sun will be discussed in later chapters as our first, and best-studied, example of a star.

Table 7.1: Mass of Members of the Solar System

Object	Percentage of Total Mass of Solar System
Sun	99.80
Jupiter	0.10
Comets	0.0005–0.03 (estimate)
All other planets and dwarf planets	0.04
Moons and rings	0.00005
Asteroids	0.000002 (estimate)
Cosmic dust	0.0000001 (estimate)

Table 7.1 also shows that most of the material of the planets is actually concentrated in the largest one, Jupiter, which is more massive than all the rest of the planets combined. Astronomers were able to determine the masses of the planets centuries ago using Kepler's laws of planetary motion and Newton's law of gravity to measure the planets' gravitational effects on one another or on moons that orbit them (see Orbits and Gravity). Today, we make even more precise measurements of their masses by tracking their gravitational effects on the motion of spacecraft that pass near them.

Beside Earth, five other planets were known to the ancients—Mercury, Venus, Mars, Jupiter, and Saturn—and two were discovered after the invention of the telescope: Uranus and Neptune. The eight planets all revolve in the same direction around the Sun. They orbit in approximately the same plane, like cars traveling on concentric tracks on a giant, flat racecourse. Each planet stays in its own "traffic lane," following a nearly circular orbit about the Sun and obeying the "traffic" laws discovered by Galileo, Kepler, and Newton. Besides these planets, we have also been discovering smaller worlds beyond Neptune that are called trans-Neptunian objects or TNOs (see Figure 7.3). The first to be found, in 1930, was Pluto, but others have been discovered during the twenty-first century. One of them, Eris, is about the same size as Pluto and has at least one moon (Pluto has five known moons.) The largest TNOs are also classed as *dwarf planets,* as is the largest asteroid, Ceres. (Dwarf planets will be discussed further in the chapter on Rings, Moons, and Pluto). To date, more than 2600 of these TNOs have been discovered, and one, called Arrokoth, was explored by the New Horizons spacecraft.

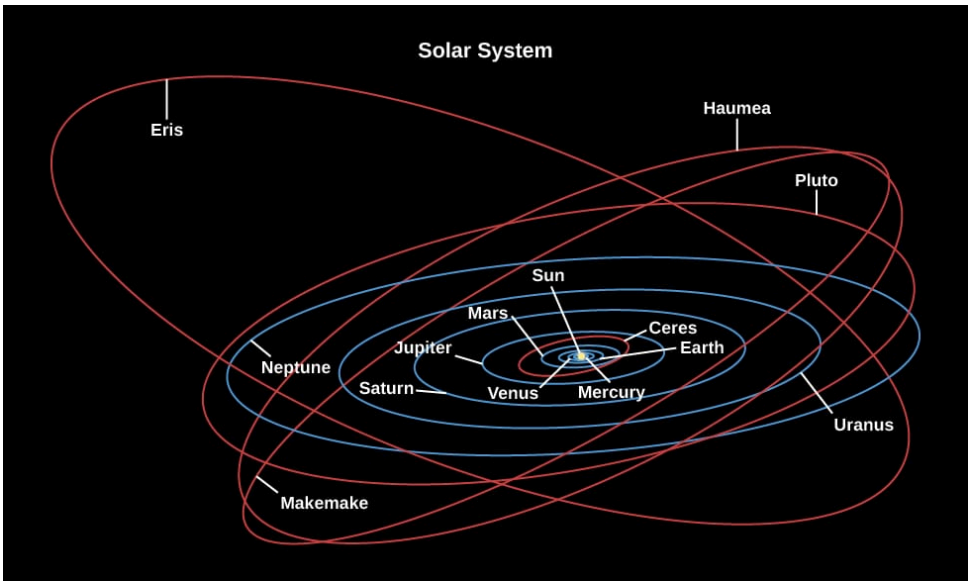


Figure 7.3 : Orbits of the Planets. All eight major planets orbit the Sun in roughly the same plane. The five currently known dwarf planets are also shown: Eris, Haumea, Pluto, Ceres, and Makemake. Note that Pluto's orbit is not in the plane of the planets.

Each of the planets and dwarf planets also rotates (spins) about an axis running through it, and in most cases the direction of rotation is the same as the direction of revolution about the Sun. The exceptions are Venus, which rotates backward very slowly (that is, in a retrograde direction), and Uranus and Pluto, which also have strange rotations, each spinning about an axis tipped nearly on its side. We do not yet know the spin orientations of Eris, Haumea, and Makemake.

The four planets closest to the Sun (Mercury through Mars) are called the inner or terrestrial planets. Often, the Moon is also discussed as a part of this group, bringing the total of terrestrial objects to five. (We generally call Earth's satellite "the Moon," with a capital M, and the other satellites "moons," with lowercase m's.) The terrestrial planets are relatively small worlds, composed primarily of rock and metal. All of them have solid surfaces that bear the records of their geological history in the forms of craters, mountains, and volcanoes (Figure 7.4).

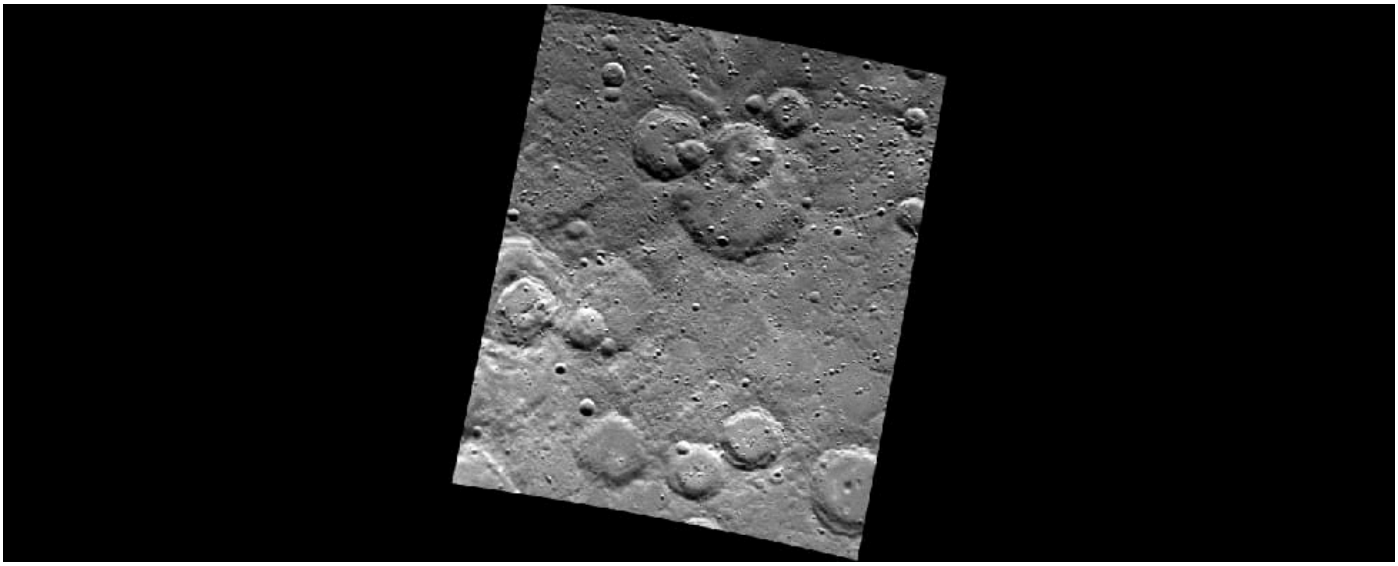


Figure 7.4 : Surface of Mercury. The pockmarked face of the terrestrial world of Mercury is more typical of the inner planets than the watery surface of Earth. This black-and-white image, taken with the Mariner 10 spacecraft, shows a region more than 400 kilometers wide. (credit: modification of work by NASA/John Hopkins University Applied Physics Laboratory/Carnegie Institution of Washington)

The next four planets (Jupiter through Neptune) are much larger and are composed primarily of lighter ices, liquids, and gases. We call these four the jovian planets (after "Jove," another name for Jupiter in mythology) or giant planets—a name they richly

deserve (Figure 7.5). About 1,300 Earths could fit inside Jupiter, for example. These planets do not have solid surfaces on which future explorers might land. They are more like vast, spherical oceans with much smaller, dense cores.

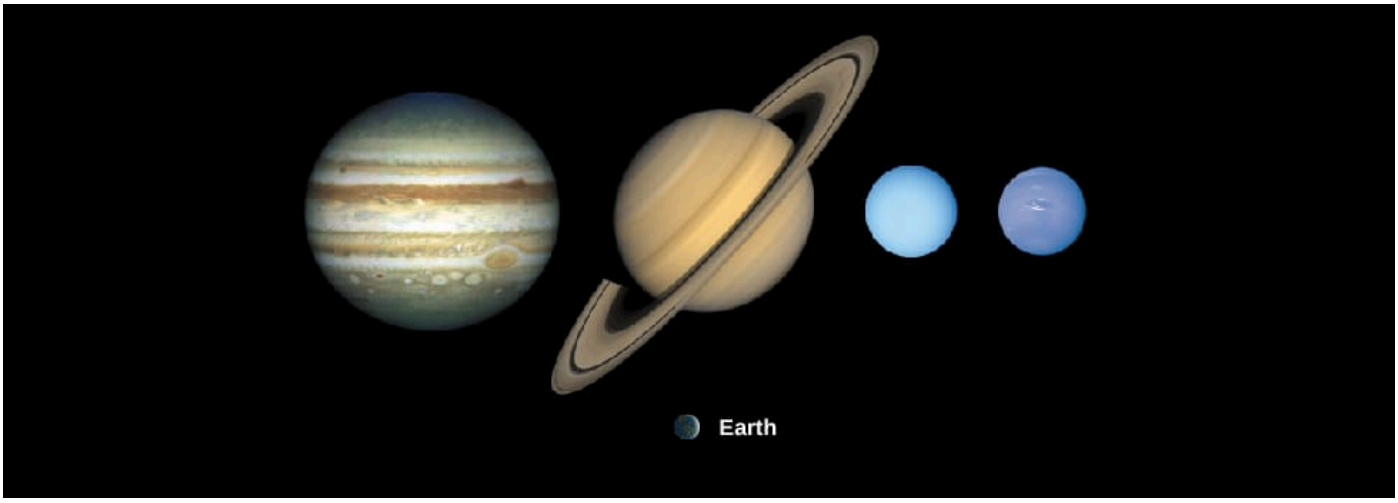


Figure 7.5 : The Four Giant Planets. This montage shows the four giant planets: Jupiter, Saturn, Uranus, and Neptune. Below them, Earth is shown to scale. (credit: modification of work by NASA, Solar System Exploration)

Near the outer edge of the system lies Pluto, which was the first of the distant icy worlds to be discovered beyond Neptune (Pluto was visited by a spacecraft, the NASA New Horizons mission, in 2015 [see Figure 7.6]). Table 7.2 summarizes some of the main facts about the planets.

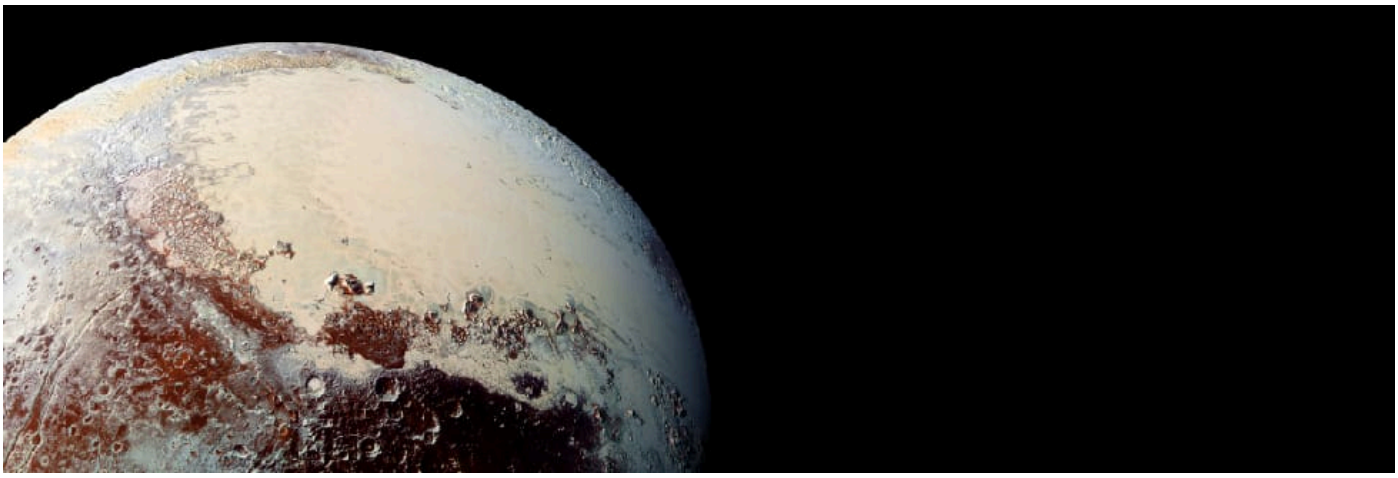


Figure 7.6 : Pluto Close-up. This intriguing image from the New Horizons spacecraft, taken when it flew by the dwarf planet in July 2015, shows some of its complex surface features. The rounded white area is called the Sputnik Plain, after humanity's first spacecraft. (credit: modification of work by NASA/Johns Hopkins University Applied Physics Laboratory/Southwest Research Institute)

Table 7.2: The Planets

Name	Distance from Sun (AU)[2]	Revolution Period (y)	Diameter (km)	Mass (10^{23} kg)	Density (g/cm^3)[3]
Mercury	0.39	0.24	4,878	3.3	5.4
Venus	0.72	0.62	12,120	48.7	5.2
Earth	1.00	1.00	12,756	59.8	5.5
Mars	1.52	1.88	6,787	6.4	3.9
Jupiter	5.20	11.86	142,984	18,991	1.3

Name	Distance from Sun (AU)[2]	Revolution Period (y)	Diameter (km)	Mass (10^{23} kg)	Density (g/cm^3)[3]
Saturn	9.54	29.46	120,536	5686	0.7
Uranus	19.18	84.07	51,118	866	1.3
Neptune	30.06	164.82	49,660	1030	1.6

Example 7.1: Comparing Densities

Let's compare the densities of several members of the solar system. The density of an object equals its mass divided by its volume. The volume (V) of a sphere (like a planet) is calculated using the equation

V=43πR3V=43πR3

where π (the Greek letter pi) has a value of approximately 3.14. Although planets are not perfect spheres, this equation works well enough. The masses and diameters of the planets are given in Table 7.2. For data on selected moons, see Appendix G. Let's use Saturn's moon Mimas as our example, with a mass of 4×10194×1019 kg and a diameter of approximately 400 km (radius, 200 km=2×105m 200 km=2×105m).

Solution

The volume of Mimas is

43×3.14×(2×105m)3=3.3×1016m3.43×3.14×(2×105m)3=3.3×1016m3.

Density is mass divided by volume:

4×1019kg3.3×1016m3=1.2×103kg/m3.4×1019kg3.3×1016m3=1.2×103kg/m3.

Note that the density of water in these units is 1000 kg/m^3, so Mimas must be made mainly of ice, not rock. (Note that the density of Mimas given in Appendix G is 1.2, but the units used there are different. In that table, we give density in units of g/cm^3, for which the density of water equals 1. Can you show, by converting units, that 1 g/cm^3 is the same as 1000 kg/m^3?)

Exercise 3.1.1

Calculate the average density of our own planet, Earth. Show your work. How does it compare to the density of an ice moon like Mimas? See Table 7.2 for data.

Answer available at the end of this section.

Link to Learning

Learn more about NASA's mission to Pluto and see high-resolution images of Pluto and its moon Charon.

Smaller Members of the Solar System

Most of the planets are accompanied by one or more moons; only Mercury and Venus move through space alone. There are more than 210 known moons orbiting planets and dwarf planets (see Appendix G for a listing of the larger ones), and undoubtedly many other small ones remain undiscovered. The largest of the moons are as big as small planets and just as interesting. In addition to our Moon, they include the four largest moons of Jupiter (called the Galilean moons, after their discoverer) and the largest moons of Saturn and Neptune (confusingly named Titan and Triton).

Each of the giant planets also has rings made up of countless small bodies ranging in size from mountains to mere grains of dust, all in orbit about the equator of the planet. The bright rings of Saturn are, by far, the easiest to see. They are among the most beautiful sights in the solar system (Figure 7.7). But, all four ring systems are interesting to scientists because of their complicated forms, influenced by the pull of the moons that also orbit these giant planets.

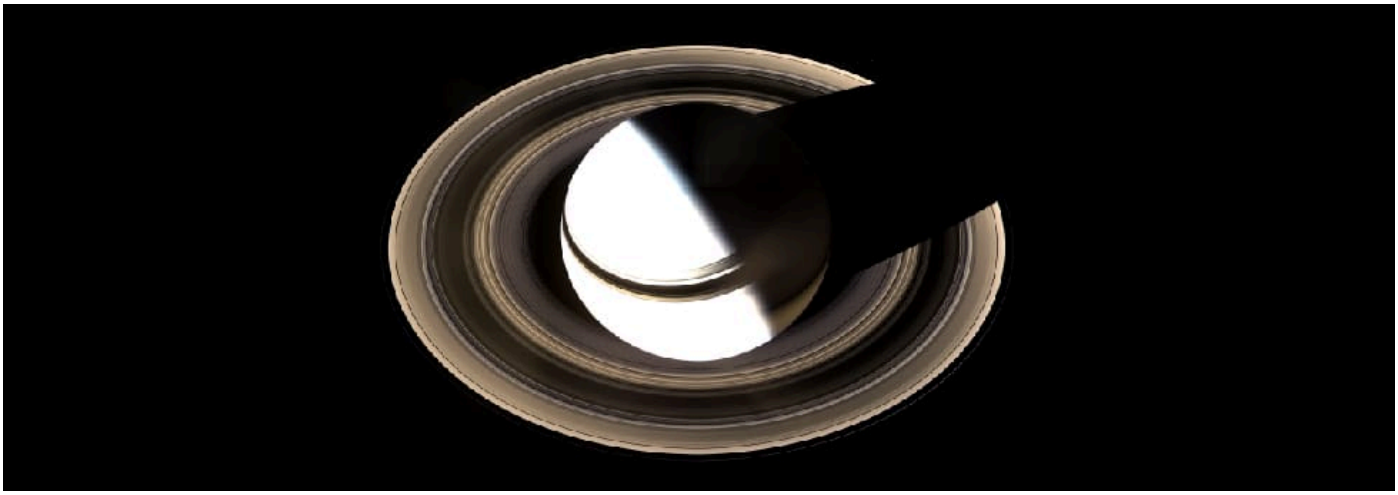


Figure 7.7 : Saturn and Its Rings. This 2007 Cassini image shows Saturn and its complex system of rings, taken from a distance of about 1.2 million kilometers. This natural-color image is a composite of 36 images taken over the course of 2.5 hours. (credit: modification of work by NASA/JPL/Space Science Institute)

The solar system has many other less-conspicuous members. Another group is the asteroids, rocky bodies that orbit the Sun like miniature planets, mostly in the space between Mars and Jupiter (although some do cross the orbits of planets like Earth—see Figure 7.8). Most asteroids are remnants of the initial population of the solar system that existed before the planets themselves formed. Some of the smallest moons of the planets, such as the moons of Mars, are very likely captured asteroids.

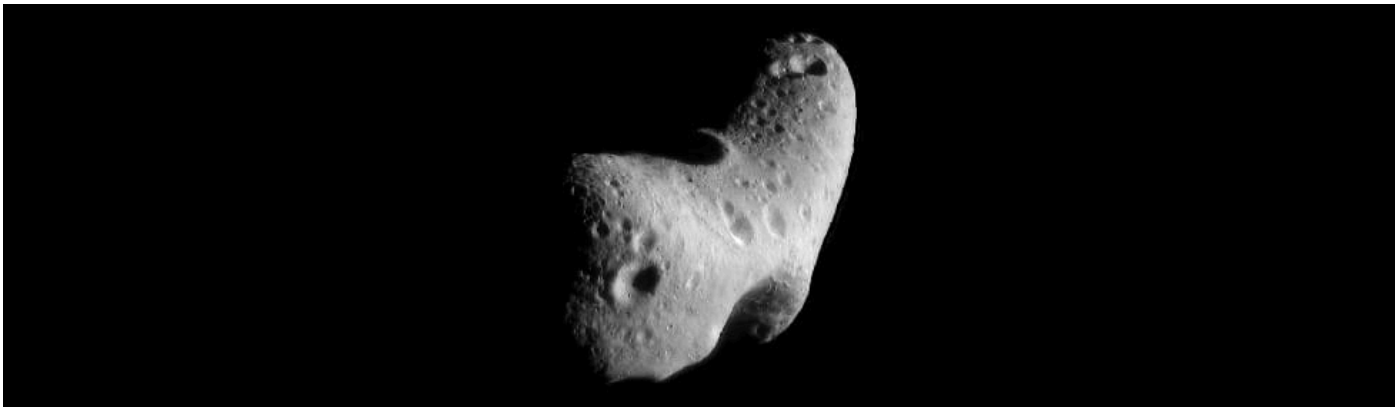


Figure 7.8 : Asteroid Eros. This small Earth-crossing asteroid image was taken by the NEAR-Shoemaker spacecraft from an altitude of about 100 kilometers. This view of the heavily cratered surface is about 10 kilometers wide. The spacecraft orbited Eros for a year before landing gently on its surface. (credit: modification of work by NASA/JHUAPL)

Another class of small bodies is composed mostly of ice, made of frozen gases such as water, carbon dioxide, and carbon monoxide; these objects are called comets (see Figure 7.9). Comets also are remnants from the formation of the solar system, but they were formed and continue (with rare exceptions) to orbit the Sun in distant, cooler regions—stored in a sort of cosmic deep freeze. This is also the realm of the larger icy worlds, called dwarf planets.

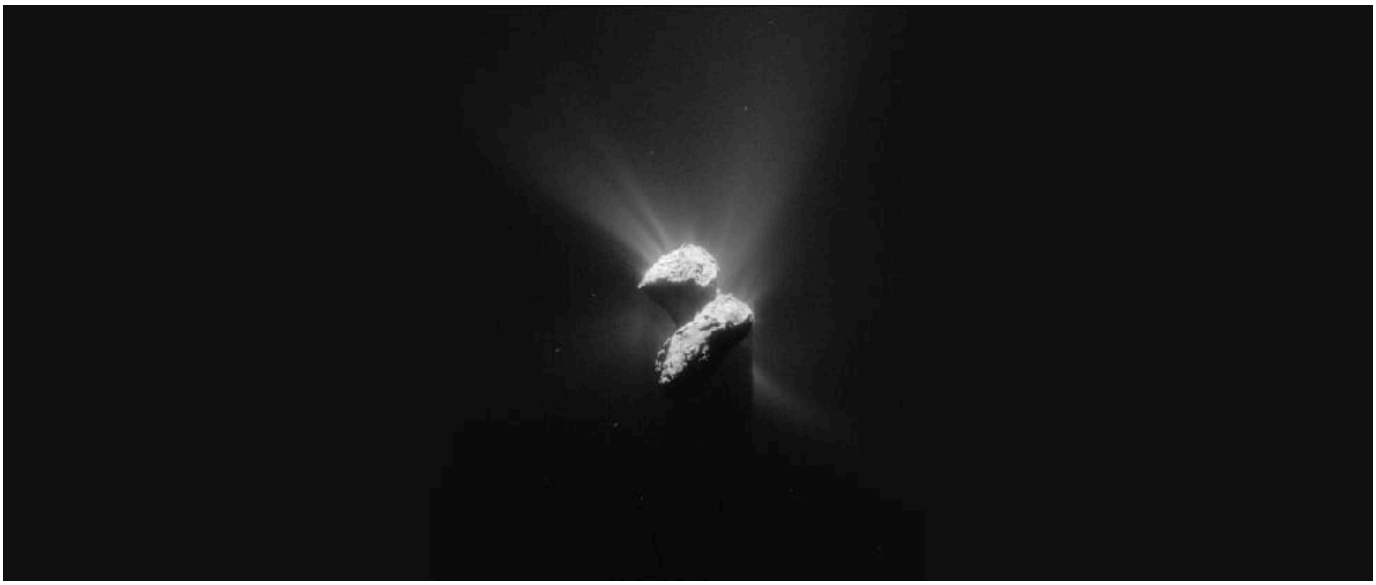


Figure 7.9 : Comet Churyumov-Gerasimenko (67P). This image shows Comet Churyumov-Gerasimenko, also known as 67P, near its closest approach to the Sun in 2015, as seen from the *Rosetta* spacecraft. Note the jets of gas escaping from the solid surface. (credit: modification of work by ESA/Rosetta/NAVACAM, CC BY-SA IGO 3.0)

Finally, there are countless grains of broken rock, which we call cosmic dust, scattered throughout the solar system. When these particles enter Earth's atmosphere (as millions do each day) they burn up, producing a brief flash of light in the night sky known as a meteor (meteors are often referred to as shooting stars). Occasionally, some larger chunk of rocky or metallic material survives its passage through the atmosphere and lands on Earth. Any piece that strikes the ground is known as a meteorite. (You can see meteorites on display in many natural history museums and can sometimes even purchase pieces of them from gem and mineral dealers.)

Voyagers in Astronomy: Carl Sagan: Solar System Advocate

The best-known astronomer in the world during the 1970s and 1980s, Carl Sagan devoted most of his professional career to studying the planets and considerable energy to raising public awareness of what we can learn from exploring the solar system (see Figure 7.10). Born in Brooklyn, New York, in 1934, Sagan became interested in astronomy as a youngster; he also credits science fiction stories for sustaining his fascination with what's out in the universe.

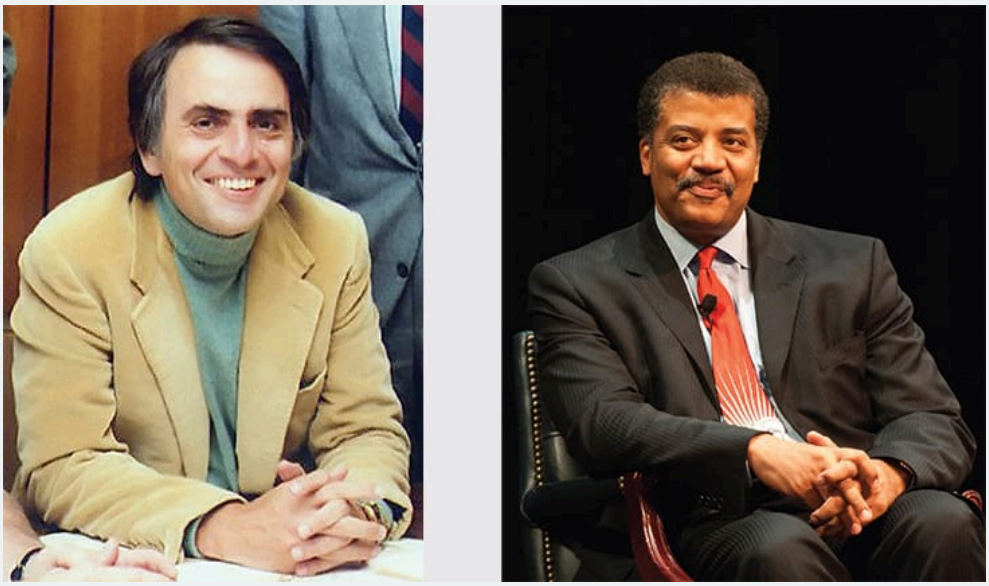


Figure 7.10 : Carl Sagan (1934–1996) and Neil deGrasse Tyson. Sagan was Tyson's inspiration to become a scientist. (credit "Sagan": modification of work by NASA, JPL; credit "Tyson": modification of work by Bruce F. Press)

In the early 1960s, when many scientists still thought Venus might turn out to be a hospitable place, Sagan calculated that the thick atmosphere of Venus could act like a giant greenhouse, keeping the heat in and raising the temperature enormously. He showed that the seasonal changes astronomers had seen on Mars were caused, not by vegetation, but by wind-blown dust. He was a member of the scientific teams for many of the robotic missions that explored the solar system and was instrumental in getting NASA to put a message-bearing plaque aboard the Pioneer spacecraft, as well as audio-video records on the Voyager spacecraft—all of them destined to leave our solar system entirely and send these little bits of Earth technology out among the stars.

To encourage public interest and public support of planetary exploration, Sagan helped found The Planetary Society, now the largest space-interest organization in the world. He was a tireless and eloquent advocate of the need to study the solar system close-up and the value of learning about other worlds in order to take better care of our own.

Sagan simulated conditions on early Earth to demonstrate how some of life's fundamental building blocks might have formed from the "primordial soup" of natural compounds on our planet. In addition, he and his colleagues developed computer models showing the consequences of nuclear war for Earth would be even more devastating than anyone had thought (this is now called the nuclear winter hypothesis) and demonstrating some of the serious consequences of continued pollution of our atmosphere.

Sagan was perhaps best known, however, as a brilliant popularizer of astronomy and the author of many books on science, including the best-selling *Cosmos*, and several evocative tributes to solar system exploration such as *The Cosmic Connection* and *Pale Blue Dot*. His book *The Demon Haunted World*, completed just before his death in 1996, is perhaps the best antidote to fuzzy thinking about pseudo-science and irrationality in print today. An intriguing science fiction novel he wrote, titled *Contact*, which became a successful film as well, is still recommended by many science instructors as a scenario for making contact with life elsewhere that is much more reasonable than most science fiction.

Sagan was a master, too, of the television medium. His 13-part public television series, *Cosmos*, was seen by an estimated 500 million people in 60 countries and has become one of the most-watched series in the history of public broadcasting. A few astronomers scoffed at a scientist who spent so much time in the public eye, but it is probably fair to say that Sagan's enthusiasm and skill as an explainer won more friends for the science of astronomy than anyone or anything else in the second half of the twentieth century.

In the two decades since Sagan's death, no other scientist has achieved the same level of public recognition. Perhaps closest is the director of the Hayden Planetarium, Neil deGrasse Tyson, who followed in Sagan's footsteps by making an updated version of the *Cosmos* program in 2014. Tyson is quick to point out that Sagan was his inspiration to become a scientist, telling how Sagan invited him to visit for a day at Cornell when he was a high school student looking for a career. However, the media environment has fragmented a great deal since Sagan's time. It is interesting to speculate whether Sagan could have adapted his communication style to the world of cable television, Twitter, Facebook, and podcasts.

Link to Learning

Two imaginative videos provide a tour of the solar system objects we have been discussing. Shane Gellert's I Need Some Space uses NASA photography and models to show the various worlds with which we share our system. In the more science fiction-oriented Wanderers video, we see some of the planets and moons as tourist destinations for future explorers, with commentary taken from recordings by Carl Sagan.

A Scale Model of the Solar System

Astronomy often deals with dimensions and distances that far exceed our ordinary experience. What does 1.4 billion kilometers—the distance from the Sun to Saturn—really mean to anyone? It can be helpful to visualize such large systems in terms of a scale model.

In our imaginations, let us build a scale model of the solar system, adopting a scale factor of 1 billion (10^9)—that is, reducing the actual solar system by dividing every dimension by a factor of 10^9. Earth, then, has a diameter of 1.3 centimeters, about the size of a grape. The Moon is a pea orbiting this at a distance of 40 centimeters, or a little more than a foot away. The Earth-Moon system fits into a standard backpack.

In this model, the Sun is nearly 1.5 meters in diameter, about the average height of an adult, and our Earth is at a distance of 150 meters—about one city block—from the Sun. Jupiter is five blocks away from the Sun, and its diameter is 15 centimeters, about the size of a very large grapefruit. Saturn is 10 blocks from the Sun; Uranus, 20 blocks; and Neptune, 30 blocks. Pluto, with a distance that varies quite a bit during its 249-year orbit, is currently just beyond 30 blocks and getting farther with time. Most of the moons of the outer solar system are the sizes of various kinds of seeds orbiting the grapefruit, oranges, and lemons that represent the outer planets.

In our scale model, a human is reduced to the dimensions of a single atom, and cars and spacecraft to the size of molecules. Sending the Voyager spacecraft to Neptune involves navigating a single molecule from the Earth–grape toward a lemon 5 kilometers away with an accuracy equivalent to the width of a thread in a spider's web.

If that model represents the solar system, where would the nearest stars be? If we keep the same scale, the closest stars would be tens of thousands of kilometers away. If you built this scale model in the city where you live, you would have to place the representations of these stars on the other side of Earth or beyond.

By the way, model solar systems like the one we just presented have been built in cities throughout the world. In Sweden, for example, Stockholm's huge Globe Arena has become a model for the Sun, and Pluto is represented by a 12-centimeter sculpture in the small town of Delsbo, 300 kilometers away. Another model solar system is in Washington on the Mall between the White House and Congress (perhaps proving they are worlds apart?).

Link to Learning

This model of the solar system shows all orbits and sizes to scale, and it lets you fly between the planets at an enhanced speed.

Making Connections: Names in the Solar System

We humans just don't feel comfortable until something has a name. Types of butterflies, new elements, and the mountains of Venus all need names for us to feel we are acquainted with them. How do we give names to objects and features in the solar system?

Planets and moons are named after gods and heroes in Greek and Roman mythology (with a few exceptions among the moons of Uranus, which have names drawn from English literature). When William Herschel, a German immigrant to England, first discovered the planet we now call Uranus, he wanted to name it Georgium Sidus (George's star) after King George III of his adopted country. This caused such an outcry among astronomers in other nations, however, that the classic tradition was upheld—and has been maintained ever since. Luckily, there were a lot of minor gods in the ancient pantheon, so plenty of names are left for the many small moons we are discovering around the giant planets. (Appendix G lists the larger moons). More recently, the names of dwarf planets and their moons have been drawn from the mythology of other cultures besides Greek and Roman.

Comets are often named after their discoverers (offering an extra incentive to comet hunters). Asteroids are named by their discoverers after just about anyone or anything they want. Recently, asteroid names have been used to recognize people who have made significant contributions to astronomy, including the three senior authors of this book.

That was pretty much all the naming that was needed while our study of the solar system was confined to Earth. But now, our spacecraft have surveyed and photographed many worlds in great detail, and each world has a host of features that also need names. To make sure that naming things in space remains multinational, rational, and somewhat dignified, astronomers have given the responsibility of approving names to a special committee of the International Astronomical Union (IAU), the body that includes scientists from every country that does astronomy.

This IAU committee has developed a set of rules for naming features on other worlds. For example, craters on Venus are named for women who have made significant contributions to human knowledge and welfare. Volcanic features on Jupiter's moon Io, which is in a constant state of volcanic activity, are named after gods of fire and thunder from the mythologies of many cultures. Craters on Mercury commemorate famous novelists, playwrights, artists, and composers. On Saturn's moon Tethys, all the features are named after characters and places in Homer's great epic poem, *The Odyssey*. As we explore

further, it may well turn out that more places in the solar system need names than Earth history can provide. Perhaps by then, explorers and settlers on these worlds will be ready to develop their own names for the places they may (if but for a while) call home.

You may be surprised to know that the meaning of the word *planet* has recently become controversial because we have discovered many other planetary systems that don't look very much like our own. Even within our solar system, the planets differ greatly in size and chemical properties. The biggest dispute concerns Pluto, which is much smaller than the other eight major planets. The category of dwarf planet was invented to include Pluto and similar icy objects beyond Neptune. But is a dwarf planet also a planet? Logically, it should be, but even this simple issue of grammar has been the subject of heated debate among both astronomers and the general public.

Footnotes

- 1The generic term for a group of planets and other bodies circling a star is *planetary system*. Ours is called the *solar system* because our Sun is sometimes called *Sol*. Strictly speaking, then, there is only one solar system; planets orbiting other stars are in planetary systems.
- 2An AU (or astronomical unit) is the distance from Earth to the Sun.
- 3We give densities in units where the density of water is 1 g/cm^3. To get densities in units of kg/m^3, multiply the given value by 1000.

Answers

Exercise 3.1.1

For a sphere,
density=mass(43πR3)kg/m3.density=mass(43πR3)kg/m3.
For Earth, then,
density=6×1024kg4.2×2.6×1020m3=5.5×103kg/m3.density=6×1024kg4.2×2.6×1020m3=5.5×103kg/m3.
This density is four to five times greater than Mimas'. In fact, Earth is the densest of the planets.

This page titled 3.1: Overview of Our Planetary System is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by OpenStax via source content that was edited to the style and standards of the LibreTexts platform.

3.2: Composition and Structure of Planets

Learning Objectives

By the end of this section, you will be able to:

- Describe the characteristics of the giant planets, terrestrial planets, and small bodies in the solar system
- Explain what influences the temperature of a planet's surface
- Explain why there is geological activity on some planets and not on others

The fact that there are two distinct kinds of planets—the rocky terrestrial planets and the gas-rich jovian planets—leads us to believe that they formed under different conditions. Certainly their compositions are dominated by different elements. Let us look at each type in more detail.

The Giant Planets

The two largest planets, Jupiter and Saturn, have nearly the same chemical makeup as the Sun; they are composed primarily of the two elements hydrogen and helium, with 75% of their mass being hydrogen and 25% helium. On Earth, both hydrogen and helium are gases, so Jupiter and Saturn are sometimes called gas planets. But, this name is misleading. Jupiter and Saturn are so large that the gas is compressed in their interior until the hydrogen becomes a liquid. Because the bulk of both planets consists of compressed, liquefied hydrogen, we should really call them liquid planets.

Under the force of gravity, the heavier elements sink toward the inner parts of a liquid or gaseous planet. Both Jupiter and Saturn, therefore, have cores composed of heavier rock, metal, and ice, but we cannot see these regions directly. In fact, when we look down from above, all we see is the atmosphere with its swirling clouds (Figure 7.11). We must infer the existence of the denser core inside these planets from studies of each planet's gravity.

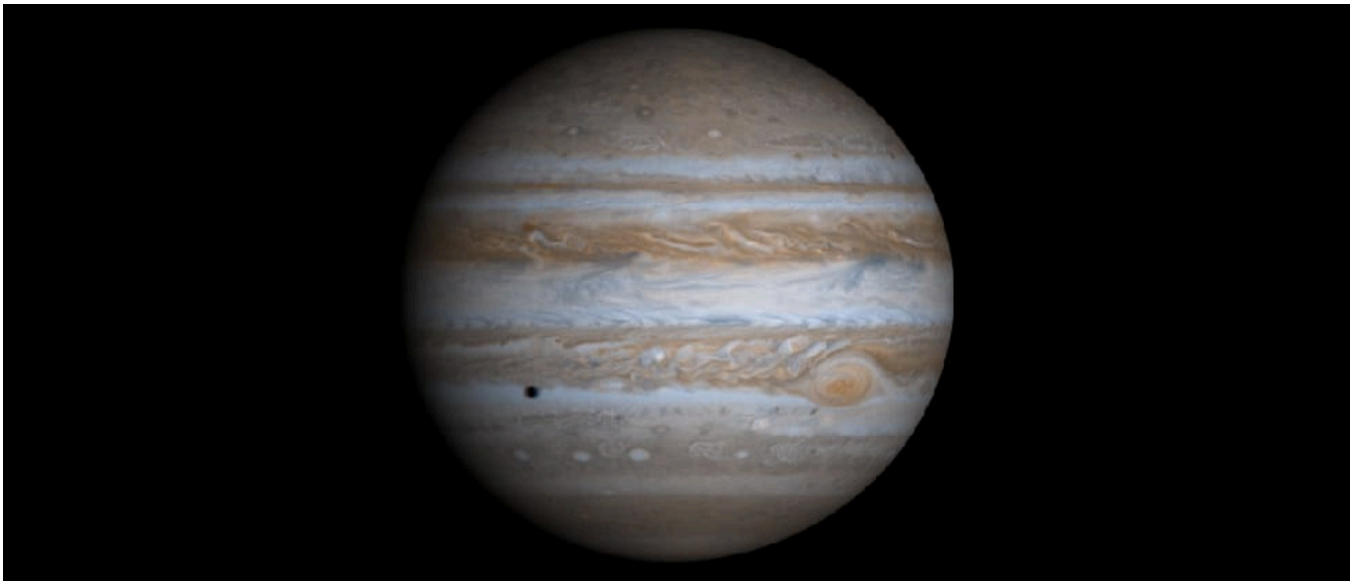


Figure 7.11 : Jupiter. This true-color image of Jupiter was taken from the Cassini spacecraft in 2000. The dark spot is the shadow of one of the giant planet's moons. (credit: modification of work by NASA/JPL/University of Arizona)

Uranus and Neptune are much smaller than Jupiter and Saturn, but each also has a core of rock, metal, and ice. Uranus and Neptune were less efficient at attracting hydrogen and helium gas, so they have much smaller atmospheres in proportion to their cores.

Chemically, each giant planet is dominated by hydrogen and its many compounds. Nearly all the oxygen present is combined

chemically with hydrogen to form water (H_2O). Chemists call such a hydrogen-dominated composition *reduced*. Throughout the outer solar system, we find abundant water (mostly in the form of ice) and reducing chemistry.

The Terrestrial Planets

The terrestrial planets are quite different from the giants. In addition to being much smaller, they are composed primarily of rocks and metals. These, in turn, are made of elements that are less common in the universe as a whole. The most abundant rocks, called silicates, are made of silicon and oxygen, and the most common metal is iron. We can tell from their densities (see Table 7.2) that Mercury has the greatest proportion of metals (which are denser) and the Moon has the lowest. Earth, Venus, and Mars all have roughly similar bulk compositions: about one third of their mass consists of iron-nickel or iron-sulfur combinations; two thirds is made of silicates. Because these planets are largely composed of oxygen compounds (such as the silicate minerals of their crusts), their chemistry is said to be *oxidized*.

When we look at the internal structure of each of the terrestrial planets, we find that the densest metals are in a central core, with the lighter silicates near the surface. If these planets were liquid, like the giant planets, we could understand this effect as the result the sinking of heavier elements due to the pull of gravity. This leads us to conclude that, although the terrestrial planets are solid today, at one time they must have been hot enough to melt.

Differentiation is the process by which gravity helps separate a planet's interior into layers of different compositions and densities. The heavier metals sink to form a core, while the lightest minerals float to the surface to form a crust. Later, when the planet cools, this layered structure is preserved. In order for a rocky planet to differentiate, it must be heated to the melting point of rocks, which is typically more than 1300 K.

Moons, Asteroids, and Comets

Chemically and structurally, Earth's Moon is like the terrestrial planets, but most moons are in the outer solar system, and they have compositions similar to the cores of the giant planets around which they orbit. The three largest moons—Ganymede and Callisto in the jovian system, and Titan in the saturnian system—are composed half of frozen water, and half of rocks and metals. Most of these moons differentiated during formation, and today they have cores of rock and metal, with upper layers and crusts of very cold and—thus very hard—ice (Figure 7.12).



Figure 7.12 : Ganymede. This view of Jupiter's moon Ganymede was taken in June 1996 by the Galileo spacecraft. The brownish gray color of the surface indicates a dusty mixture of rocky material and ice. The bright spots are places where recent impacts have uncovered fresh ice from underneath. (credit: modification of work by NASA/JPL)

Most of the asteroids and comets, as well as the smallest moons, were probably never heated to the melting point. However, some of the largest asteroids, such as Vesta, appear to be differentiated; others are fragments from differentiated bodies. Many of the smaller objects seem to be fragments or rubble piles that are the result of collisions. Because most asteroids and comets retain their original composition, they represent relatively unmodified material dating back to the time of the formation of the solar

system. In a sense, they act as chemical fossils, helping us to learn about a time long ago whose traces have been erased on larger worlds.

Temperatures: Going to Extremes

Generally speaking, the farther a planet or moon is from the Sun, the cooler its surface. The planets are heated by the radiant energy of the Sun, which gets weaker with the square of the distance. You know how rapidly the heating effect of a fireplace or an outdoor radiant heater diminishes as you walk away from it; the same effect applies to the Sun. Mercury, the closest planet to the Sun, has a blistering surface temperature that ranges from 280–430 °C on its sunlit side, whereas the surface temperature on Pluto is only about –220 °C, colder than liquid air.

Mathematically, the temperatures decrease approximately in proportion to the square root of the distance from the Sun. Pluto is about 30 AU at its closest to the Sun (or 100 times the distance of Mercury) and about 49 AU at its farthest from the Sun. Thus, Pluto's temperature is less than that of Mercury by the square root of 100, or a factor of 10: from 500 K to 50 K.

In addition to its distance from the Sun, the surface temperature of a planet can be influenced strongly by its atmosphere. Without our atmospheric insulation (the greenhouse effect, which keeps the heat in), the oceans of Earth would be permanently frozen. Conversely, if Mars once had a larger atmosphere in the past, it could have supported a more temperate climate than it has today. Venus is an even more extreme example, where its thick atmosphere of carbon dioxide acts as insulation, reducing the escape of heat built up at the surface, resulting in temperatures greater than those on Mercury. Today, Earth is the only planet where surface temperatures generally lie between the freezing and boiling points of water. As far as we know, Earth is the only planet to support life.

Astronomy Basics: There's No Place Like Home

In the classic film *The Wizard of Oz*, Dorothy, the heroine, concludes after her many adventures in "alien" environments that "there's no place like home." The same can be said of the other worlds in our solar system. There are many fascinating places, large and small, that we might like to visit, but humans could not survive on any without a great deal of artificial assistance.

A thick carbon dioxide atmosphere keeps the surface temperature on our neighbor Venus at a sizzling 700 K (near 900 °F). Mars, on the other hand, has temperatures generally below freezing, with air (also mostly carbon dioxide) so thin that it resembles that found at an altitude of 30 kilometers (100,000 feet) in Earth's atmosphere. And the red planet is so dry that it has not had any rain for billions of years.

The outer layers of the jovian planets are neither warm enough nor solid enough for human habitation. Any bases we build in the systems of the giant planets may well have to be in space or one of their moons—none of which is particularly hospitable to a luxury hotel with a swimming pool and palm trees. Perhaps we will find warmer havens deep inside the clouds of Jupiter or in the ocean under the frozen ice of its moon Europa.

All of this suggests that we had better take good care of Earth because it is the only site where life as we know it could survive. Recent human activity may be reducing the habitability of our planet by adding pollutants to the atmosphere, especially the potent greenhouse gas carbon dioxide. Human civilization is changing our planet dramatically, and these changes are not necessarily for the better. In a solar system that seems unready to receive us, making Earth less hospitable to life may be a grave mistake.

Geological Activity

The crusts of all of the terrestrial planets, as well as of the larger moons, and Pluto, have been modified over their histories by both internal and external forces. Externally, each has been battered by a slow rain of projectiles from space, leaving their surfaces pockmarked by impact craters of all sizes (see Figure 7.4). We have good evidence that this bombardment was far greater in the early history of the solar system, but it certainly continues to this day, even if at a lower rate. The collision of more than 20 large pieces of Comet Shoemaker–Levy 9 with Jupiter in the summer of 1994 (see Figure 7.13) is one dramatic example of this process.

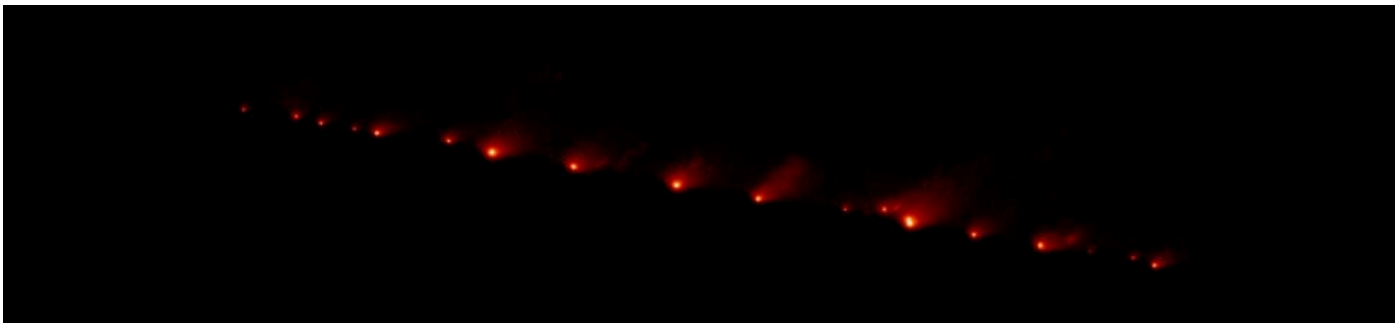


Figure 7.13 : Comet Shoemaker–Levy 9. In this image of Comet Shoemaker–Levy 9 taken on May 17, 1994, by NASA's Hubble Space Telescope, you can see about 20 icy fragments into which the comet broke. The comet was approximately 660 million kilometers from Earth, heading on a collision course with Jupiter. (credit: modification of work by NASA, ESA, H. Weaver (STScI), E. Smith (STScI))

Figure 7.14 shows the aftermath of these collisions, when debris clouds larger than Earth could be seen in Jupiter's atmosphere.

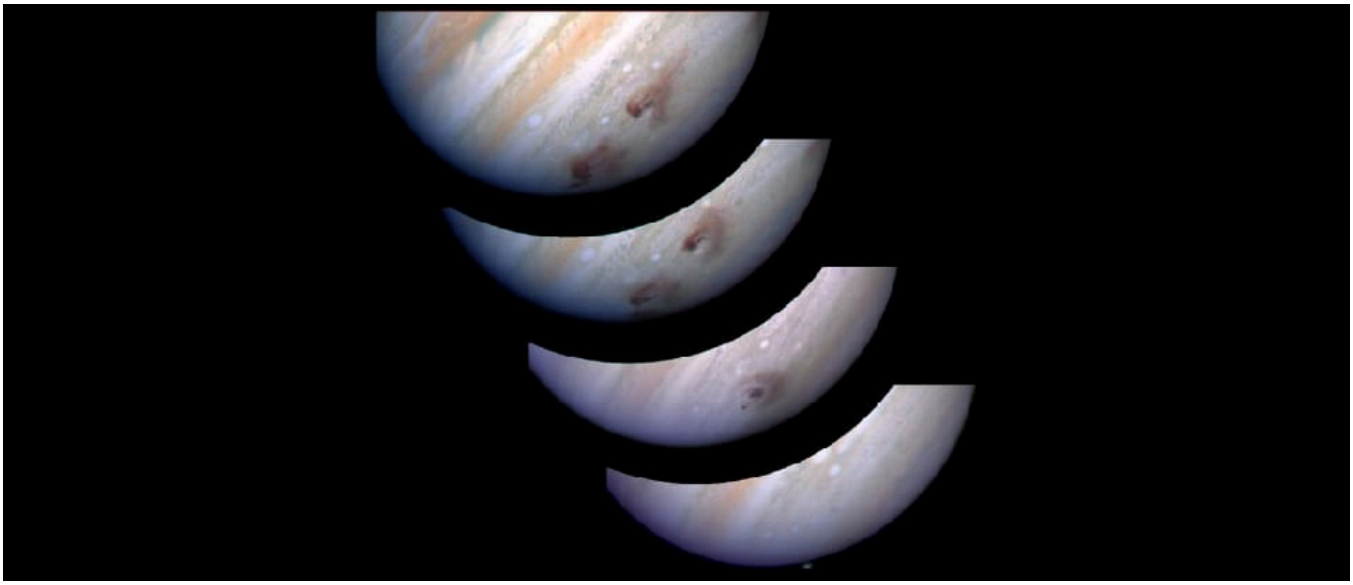


Figure 7.14 : Jupiter with Huge Dust Clouds. The Hubble Space Telescope took this sequence of images of Jupiter in summer 1994, when fragments of Comet Shoemaker–Levy 9 collided with the giant planet. Here we see the site hit by fragment G, from five minutes to five days after impact. Several of the dust clouds generated by the collisions became larger than Earth. (credit: modification of work by H. Hammel, NASA)

During the time all the planets have been subject to such impacts, internal forces on the terrestrial planets have buckled and twisted their crusts, built up mountain ranges, erupted as volcanoes, and generally reshaped the surfaces in what we call geological activity. (The prefix *geo* means "Earth," so this is a bit of an "Earth-chauvinist" term, but it is so widely used that we bow to tradition.) Among the terrestrial planets, Earth and Venus have experienced the most geological activity over their histories, although some of the moons in the outer solar system are also surprisingly active. In contrast, our own Moon is a dead world where geological activity ceased billions of years ago.

Geological activity on a planet is the result of a hot interior. The forces of volcanism and mountain building are driven by heat escaping from the interiors of planets. As we will see, each of the planets was heated at the time of its birth, and this primordial heat initially powered extensive volcanic activity, even on our Moon. But, small objects such as the Moon soon cooled off. The larger the planet or moon, the longer it retains its internal heat, and therefore the more we expect to see surface evidence of continuing geological activity. The effect is similar to our own experience with a hot baked potato: the larger the potato, the more slowly it cools. If we want a potato to cool quickly, we cut it into small pieces.

For the most part, the history of volcanic activity on the terrestrial planets conforms to the predictions of this simple theory. The Moon, the smallest of these objects, is a geologically dead world. Although we know less about Mercury, it seems likely that this planet, too, ceased most volcanic activity about the same time the Moon did. Mars represents an intermediate case. It has been

much more active than the Moon, but less so than Earth. Earth and Venus, the largest terrestrial planets, still have molten interiors even today, some 4.5 billion years after their birth.

This page titled 3.2: Composition and Structure of Planets is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by OpenStax via source content that was edited to the style and standards of the LibreTexts platform.

3.3: Formation of Planetary Systems

This section examines how stars and planetary systems form from clouds of gas and dust. It describes the collapse of a solar nebula, the formation of a rotating disk around the young Sun, and the processes by which small particles gradually combined to form planetesimals and planets. The resulting structure of our solar system provides important evidence for understanding how planetary systems form and evolve.

This page titled 3.3: Formation of Planetary Systems is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Robert Wagner.

3.3.1: Origin of the Solar System

Learning Objectives

By the end of this section, you will be able to:

- Describe the characteristics of planets that are used to create formation models of the solar system
- Describe how the characteristics of extrasolar systems help us to model our own solar system
- Explain the importance of collisions in the formation of the solar system

Much of astronomy is motivated by a desire to understand the origin of things: to find at least partial answers to age-old questions of where the universe, the Sun, Earth, and we ourselves came from. Each planet and moon is a fascinating place that may stimulate our imagination as we try to picture what it would be like to visit. Taken together, the members of the solar system preserve patterns that can tell us about the formation of the entire system. As we begin our exploration of the planets, we want to introduce our modern picture of how the solar system formed.

The recent discovery of thousands of planets in orbit around other stars has shown astronomers that many exoplanetary systems can be quite different from our own solar system. For example, it is common for these systems to include planets intermediate in size between our terrestrial and giant planets. These are often called *superearths*. Some exoplanet systems even have giant planets close to the star, reversing the order we see in our system. In The Birth of Stars and the Discovery of Planets outside the Solar System, we will look at these exoplanet systems. But for now, let us focus on theories of how our own particular system has formed and evolved.

Looking for Patterns

One way to approach our question of origin is to look for regularities among the planets. We found, for example, that all the planets lie in nearly the same plane and revolve in the same direction around the Sun. The Sun also spins in the same direction about its own axis. Astronomers interpret this pattern as evidence that the Sun and planets formed together from a spinning cloud of gas and dust that we call the solar nebula (Figure 7.17).



Figure 7.17 : Solar Nebula. This artist's conception of the solar nebula shows the flattened cloud of gas and dust from which our planetary system formed. Icy and rocky planetesimals (precursors of the planets) can be seen in the foreground. The bright center is where the Sun is forming. (credit: William K. Hartmann, Planetary Science Institute)

The composition of the planets gives another clue about origins. Spectroscopic analysis allows us to determine which elements

are present in the Sun and the planets. The Sun has the same hydrogen-dominated composition as Jupiter and Saturn, and therefore appears to have been formed from the same reservoir of material. In comparison, the terrestrial planets and our Moon are relatively deficient in the light gases and the various ices that form from the common elements oxygen, carbon, and nitrogen. Instead, on Earth and its neighbors, we see mostly the rarer heavy elements such as iron and silicon. This pattern suggests that the processes that led to planet formation in the inner solar system must somehow have excluded much of the lighter materials that are common elsewhere. These lighter materials must have escaped, leaving a residue of heavy stuff.

The reason for this is not hard to guess, bearing in mind the heat of the Sun. The inner planets and most of the asteroids are made of rock and metal, which can survive heat, but they contain very little ice or gas, which evaporate when temperatures are high. (To see what we mean, just compare how long a rock and an ice cube survive when they are placed in the sunlight.) In the outer solar system, where it has always been cooler, the planets and their moons, as well as icy dwarf planets and comets, are composed mostly of ice and gas.

The Evidence from Far Away

A second approach to understanding the origins of the solar system is to look outward for evidence that other systems of planets are forming elsewhere. We cannot look back in time to the formation of our own system, but many stars in space are much younger than the Sun. In these systems, the processes of planet formation might still be accessible to direct observation. We observe that there are many other “solar nebulas” or *circumstellar disks*—flattened, spinning clouds of gas and dust surrounding young stars. These disks resemble our own solar system’s initial stages of formation billions of years ago (Figure 7.18).



Figure 7.18 : Atlas of Planetary Nurseries. These Hubble Space Telescope photos show sections of the Orion Nebula, a relatively close-by region where stars are currently forming. Each image shows an embedded circumstellar disk orbiting a very young star. Seen from different angles, some are energized to glow by the light of a nearby star while others are dark and seen in silhouette against the bright glowing gas of the Orion Nebula. Each is a contemporary analog of our own solar nebula—a location where planets are probably being formed today. (credit: modification of work by NASA/ESA, L. Ricci (ESO))

Building Planets

Circumstellar disks are a common occurrence around very young stars, suggesting that disks and stars form together. Astronomers can use theoretical calculations to see how solid bodies might form from the gas and dust in these disks as they cool. These models show that material begins to coalesce first by forming smaller objects, precursors of the planets, which we call planetesimals.

Today’s fast computers can simulate the way millions of planetesimals, probably no larger than 100 kilometers in diameter, might gather together under their mutual gravity to form the planets we see today. We are beginning to understand that this process was a violent one, with planetesimals crashing into each other and sometimes even disrupting the growing planets themselves. As a consequence of those violent impacts (and the heat from radioactive elements in them), all the planets were heated until they were liquid and gas, and therefore differentiated, which helps explain their present internal structures.

The process of impacts and collisions in the early solar system was complex and, apparently, often random. The solar nebula model can explain many of the regularities we find in the solar system, but the random collisions of massive collections of planetesimals could be the reason for some exceptions to the “rules” of solar system behavior. For example, why do the planets

Uranus and Pluto spin on their sides? Why does Venus spin slowly and in the opposite direction from the other planets? Why does the composition of the Moon resemble Earth in many ways and yet exhibit substantial differences? The answers to such questions probably lie in enormous collisions that took place in the solar system long before life on Earth began.

Today, some 4.5 billion years after its origin, the solar system is—thank goodness—a much less violent place. As we will see, however, some planetesimals have continued to interact and collide, and their fragments move about the solar system as roving "transients" that can make trouble for the established members of the Sun's family, such as our own Earth. (We discuss this "troublemaking" in Comets and Asteroids: Debris of the Solar System.)

Link to Learning

A great variety of infographics at space.com let you explore what it would be like to live on various worlds in the solar system.

This page titled 3.3.1: Origin of the Solar System is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by OpenStax via source content that was edited to the style and standards of the LibreTexts platform.

3.3.2: Formation of the Solar System

Learning Objectives

By the end of this section, you will be able to:

- Describe the motion, chemical, and age constraints that must be met by any theory of solar system formation
- Summarize the physical and chemical changes during the solar nebula stage of solar system formation
- Explain the formation process of the terrestrial and giant planets
- Describe the main events of the further evolution of the solar system

As we have seen, the comets, asteroids, and meteorites are surviving remnants from the processes that formed the solar system. The planets, moons, and the Sun, of course, also are the products of the formation process, although the material in them has undergone a wide range of changes. We are now ready to put together the information from all these objects to discuss what is known about the origin of the solar system.

Observational Constraints

There are certain basic properties of the planetary system that any theory of its formation must explain. These may be summarized under three categories: motion constraints, chemical constraints, and age constraints. We call them *constraints* because they place restrictions on our theories; unless a theory can explain the observed facts, it will not survive in the competitive marketplace of ideas that characterizes the endeavor of science. Let's take a look at these constraints one by one.

There are many regularities to the motions in the solar system. We saw that the planets all revolve around the Sun in the same direction and approximately in the plane of the Sun's own rotation. In addition, most of the planets rotate in the same direction as they revolve, and most of the moons also move in counterclockwise orbits (when seen from the north). With the exception of the comets and other trans-neptunian objects, the motions of the system members define a disk or Frisbee shape. Nevertheless, a full theory must also be prepared to deal with the exceptions to these trends, such as the *retrograde rotation* (not revolution) of Venus.

In the realm of chemistry, we saw that Jupiter and Saturn have approximately the same composition—dominated by hydrogen and helium. These are the two largest planets, with sufficient gravity to hold on to any gas present when and where they formed; thus, we might expect them to be representative of the original material out of which the solar system formed. Each of the other members of the planetary system is, to some degree, lacking in the light elements. A careful examination of the composition of solid solar-system objects shows a striking progression from the metal-rich inner planets, through those made predominantly of rocky materials, out to objects with ice-dominated compositions in the outer solar system. The comets in the Oort cloud and the trans-neptunian objects in the Kuiper belt are also icy objects, whereas the asteroids represent a transitional rocky composition with abundant dark, carbon-rich material.

As we saw in Other Worlds: An Introduction to the Solar System, this general chemical pattern can be interpreted as a temperature sequence: hot near the Sun and cooler as we move outward. The inner parts of the system are generally missing those materials that could not condense (form a solid) at the high temperatures found near the Sun. However, there are (again) important exceptions to the general pattern. For example, it is difficult to explain the presence of water on Earth and Mars if these planets formed in a region where the temperature was too hot for ice to condense, unless the ice or water was brought in later from cooler regions. The extreme example is the observation that there are polar deposits of ice on both Mercury and the Moon; these are almost certainly formed and maintained by occasional comet impacts.

As far as age is concerned, we discussed that radioactive dating demonstrates that some rocks on the surface of Earth have been present for at least 3.8 billion years, and that certain lunar samples are 4.4 billion years old. The primitive meteorites all have radioactive ages near 4.5 billion years. The age of these unaltered building blocks is considered the age of the planetary system. The similarity of the measured ages tells us that planets formed and their crusts cooled within a few tens of millions of years (at most) of the beginning of the solar system. Further, detailed examination of primitive meteorites indicates that they are made primarily from material that condensed or coagulated out of a hot gas; few identifiable fragments appear to have survived from before this hot-vapor stage 4.5 billion years ago.

The Solar Nebula

All the foregoing constraints are consistent with the general idea, introduced in Other Worlds: An Introduction to the Solar System, that the solar system formed 4.5 billion years ago out of a rotating cloud of vapor and dust—which we call the solar nebula—with an initial composition similar to that of the Sun today. As the solar nebula collapsed under its own gravity, material fell toward the center, where things became more and more concentrated and hot. Increasing temperatures in the shrinking nebula vaporized most of the solid material that was originally present.

At the same time, the collapsing nebula began to rotate faster through the conservation of angular momentum (see the Orbits and Gravity and Earth, Moon, and Sky chapters). Like a figure skater pulling her arms in to spin faster, the shrinking cloud spun more quickly as time went on. Now, think about how a round object spins. Close to the poles, the spin rate is slow, and it gets faster as you get closer to the equator. In the same way, near the poles of the nebula, where orbits were slow, the nebular material fell directly into the center. Faster moving material, on the other hand, collapsed into a flat disk revolving around the central object (Figure 14.11). The existence of this disk-shaped rotating nebula explains the primary motions in the solar system that we discussed in the previous section. And since they formed from a rotating disk, the planets all orbit the same way.

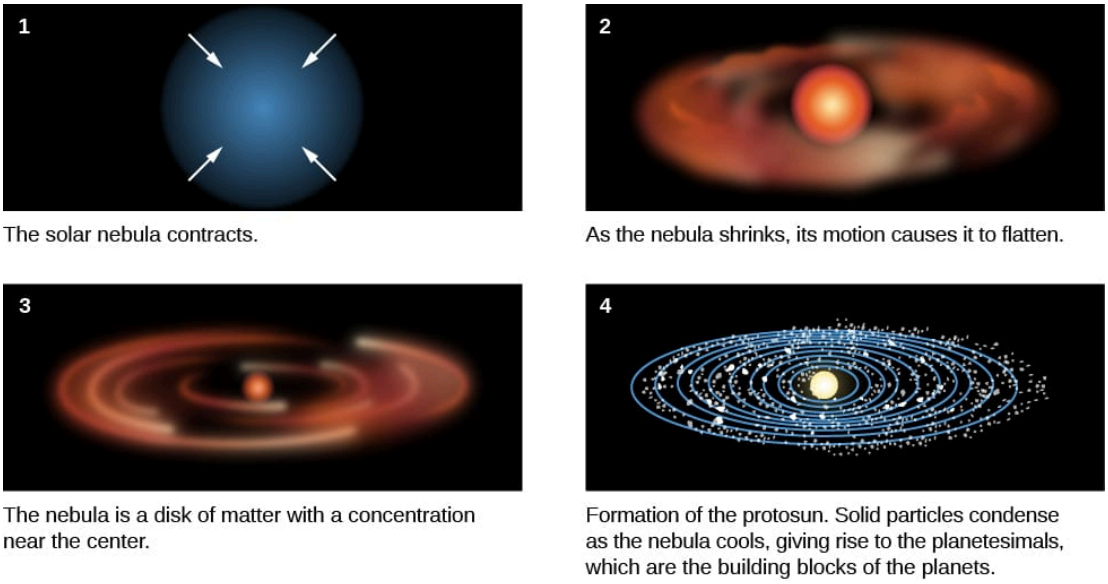


Figure 14.11 : Steps in Forming the Solar System. This illustration shows the steps in the formation of the solar system from the solar nebula. As the nebula shrinks, its rotation causes it to flatten into a disk. Much of the material is concentrated in the hot center, which will ultimately become a star. Away from the center, solid particles can condense as the nebula cools, giving rise to planetesimals, the building blocks of the planets and moons.

Picture the solar nebula at the end of the collapse phase, when it was at its hottest. With no more gravitational energy (from material falling in) to heat it, most of the nebula began to cool. The material in the center, however, where it was hottest and most crowded, formed a *star* that maintained high temperatures in its immediate neighborhood by producing its own energy. Turbulent motions and magnetic fields within the disk can drain away angular momentum, robbing the disk material of some of its spin. This allowed some material to continue to fall into the growing star, while the rest of the disk gradually stabilized.

The temperature within the disk decreased with increasing distance from the Sun, much as the planets' temperatures vary with position today. As the disk cooled, the gases interacted chemically to produce compounds; eventually these compounds condensed into liquid droplets or solid grains. This is similar to the process by which raindrops on Earth condense from moist air as it rises over a mountain.

Let's look in more detail at how material condensed at different places in the maturing disk (Figure 14.12). The first materials to form solid grains were the metals and various rock-forming silicates. As the temperature dropped, these were joined throughout much of the solar nebula by sulfur compounds and by carbon- and water-rich silicates, such as those now found abundantly among the asteroids. However, in the inner parts of the disk, the temperature never dropped low enough for such materials as ice or carbonaceous organic compounds to condense, so they were lacking on the innermost planets.

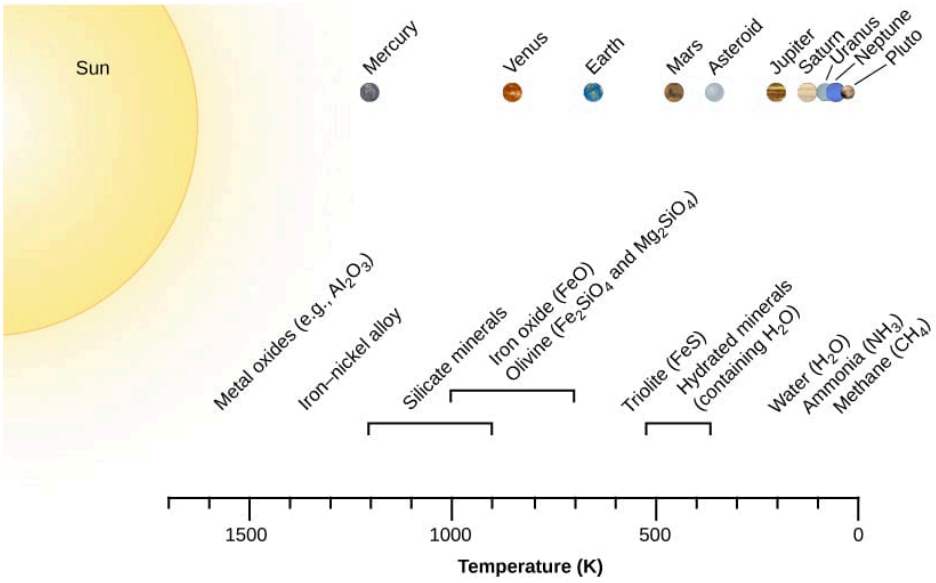


Figure 14.12 : Chemical Condensation Sequence in the Solar Nebula. The scale along the bottom shows temperature; above are the materials that would condense out at each temperature under the conditions expected to prevail in the nebula.

Far from the Sun, cooler temperatures allowed the oxygen to combine with hydrogen and condense in the form of water (H_2O) ice. Beyond the orbit of Saturn, carbon and nitrogen combined with hydrogen to make ices such as methane (CH_4) and ammonia (NH_3). This sequence of events explains the basic chemical composition differences among various regions of the solar system.

Example 14.1: Rotation of the Solar Nebula

We can use the concept of angular momentum to trace the evolution of the collapsing solar nebula. The angular momentum of an object is proportional to the square of its size (diameter) divided by its period of rotation (D2P)Appendix F tells us has a radius of about 40 AU?

Solution

We are given that the final diameter of the solar nebula is about 80 AU. Noting the initial state before the collapse and the final state at Pluto's orbit, then

PfinalPinitial=(DfinalDinitial)2=(8010,000)2=(0.008)2=0.000064PfinalPinitial=(DfinalDinitial)2=(8010,000)2=(0.008)2=0.00

With $P_{initial}$ equal to 1,000,000 years, P_{final}, the new rotation period, is 64 years. This is a lot shorter than the actual time Pluto takes to go around the Sun, but it gives you a sense of the kind of speeding up the conservation of angular momentum can produce. As we noted earlier, other mechanisms helped the material in the disk lose angular momentum before the planets fully formed.

Exercise 3.3.2.1

What would the rotation period of the nebula in our example be when it had shrunk to the size of Jupiter's orbit?

Answer available at the end of this section.

Formation of the Terrestrial Planets

The grains that condensed in the solar nebula rather quickly joined into larger and larger chunks, until most of the solid material was in the form of *planetesimals,* chunks a few kilometers to a few tens of kilometers in diameter. Some planetesimals still survive today as comets and asteroids. Others have left their imprint on the cratered surfaces of many of the worlds we studied in earlier

chapters. A substantial step up in size is required, however, to go from planetesimal to planet.

Some planetesimals were large enough to attract their neighbors gravitationally and thus to grow by the process called accretion. While the intermediate steps are not well understood, ultimately several dozen centers of accretion seem to have grown in the inner solar system. Each of these attracted surrounding planetesimals until it had acquired a mass similar to that of Mercury or Mars. At this stage, we may think of these objects as *protoplanets*—"not quite ready for prime time" planets.

Each of these protoplanets continued to grow by the accretion of planetesimals. Every incoming planetesimal was accelerated by the gravity of the protoplanet, striking with enough energy to melt both the projectile and a part of the impact area. Soon the entire protoplanet was heated to above the melting temperature of rocks. The result was *planetary differentiation*, with heavier metals sinking toward the core and lighter silicates rising toward the surface. As they were heated, the inner protoplanets lost some of their more volatile constituents (the lighter gases), leaving more of the heavier elements and compounds behind.

Formation of the Giant Planets

In the outer solar system, where the available raw materials included ices as well as rocks, the protoplanets grew to be much larger, with masses ten times greater than Earth. These protoplanets of the outer solar system were so large that they were able to attract and hold the surrounding gas. As the hydrogen and helium rapidly collapsed onto their cores, the giant planets were heated by the energy of contraction. But although these giant planets got hotter than their terrestrial siblings, they were far too small to raise their central temperatures and pressures to the point where nuclear reactions could begin (and it is such reactions that give us our definition of a star). After glowing dull red for a few thousand years, the giant planets gradually cooled to their present state (Figure 14.13).

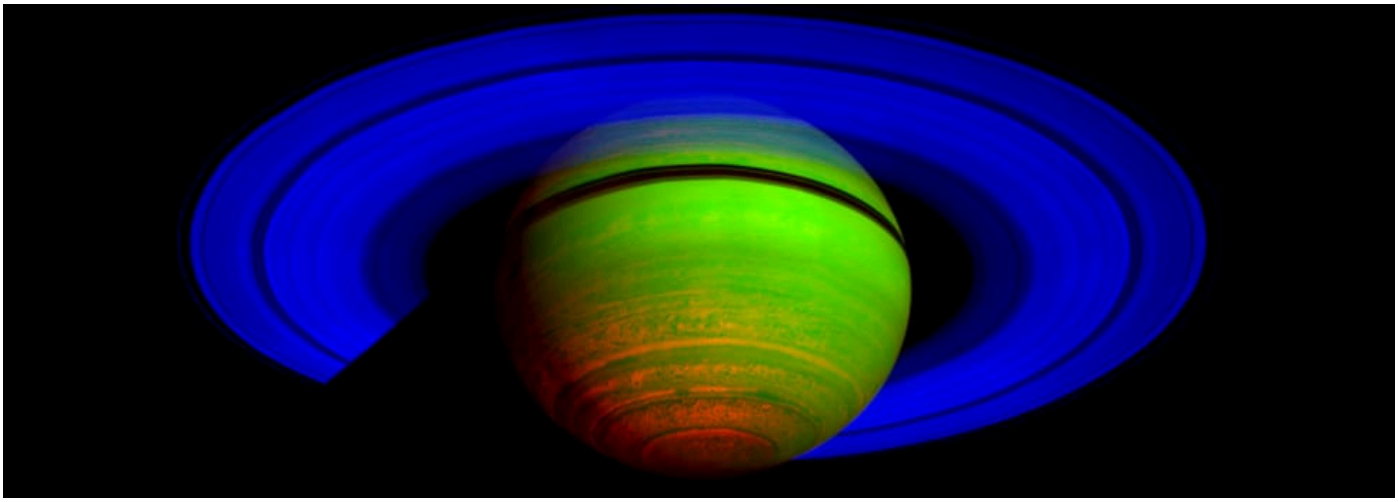


Figure 14.13 : Saturn Seen in Infrared. This image from the Cassini spacecraft is stitched together from 65 individual observations. Sunlight reflected at a wavelength of 2 micrometers is shown as blue, sunlight reflected at 3 micrometers is shown as green, and heat radiated from Saturn's interior at 5 micrometers is red. For example, Saturn's rings reflect sunlight at 2 micrometers, but not at 3 and 5 micrometers, so they appear blue. Saturn's south polar regions are seen glowing with internal heat. (credit: modification of work by NASA/JPL/University of Arizona)

The collapse of gas from the nebula onto the cores of the giant planets explains how these objects acquired nearly the same hydrogen-rich composition as the Sun. The process was most efficient for Jupiter and Saturn; hence, their compositions are most nearly "cosmic." Much less gas was captured by Uranus and Neptune, which is why these two planets have compositions dominated by the icy and rocky building blocks that made up their large cores rather than by hydrogen and helium. The initial formation period ended when much of the available raw material was used up and the solar wind (the flow of atomic particles) from the young Sun blew away the remaining supply of lighter gases.

Further Evolution of the System

All the processes we have just described, from the collapse of the solar nebula to the formation of protoplanets, took place within a few million years. However, the story of the formation of the solar system was not complete at this stage; there were many planetesimals and other debris that did not initially accumulate to form the planets. What was their fate?

The comets visible to us today are merely the tip of the cosmic iceberg (if you'll pardon the pun). Most comets are believed to be in the Oort cloud, far from the region of the planets. Additional comets and icy dwarf planets are in the Kuiper belt, which stretches beyond the orbit of Neptune. These icy pieces probably formed near the present orbits of Uranus and Neptune but were ejected from their initial orbits by the gravitational influence of the giant planets.

In the inner parts of the system, remnant planetesimals and perhaps several dozen protoplanets continued to whiz about. Over the vast span of time we are discussing, collisions among these objects were inevitable. Giant impacts at this stage may have stripped Mercury of part of its mantle and crust, reversed the rotation of Venus, and broke off part of Earth to create the Moon (all events we discussed in other chapters).

Smaller-scale impacts also added mass to the inner protoplanets. Because the gravity of the giant planets could "stir up" the orbits of the planetesimals, the material impacting on the inner protoplanets could have come from almost anywhere within the solar system. In contrast to the previous stage of accretion, therefore, this new material did not represent just a narrow range of compositions.

As a result, much of the debris striking the inner planets was ice-rich material that had condensed in the outer part of the solar nebula. As this comet-like bombardment progressed, Earth accumulated the water and various organic compounds that would later be critical to the formation of life. Mars and Venus probably also acquired abundant water and organic materials from the same source, as Mercury and the Moon are still doing to form their icy polar caps.

Gradually, as the planets swept up or ejected the remaining debris, most of the planetesimals disappeared. In two regions, however, stable orbits are possible where leftover planetesimals could avoid impacting the planets or being ejected from the system. These regions are the asteroid belt between Mars and Jupiter and the Kuiper belt beyond Neptune. The planetesimals (and their fragments) that survive in these special locations are what we now call asteroids, comets, and trans-neptunian objects.

Astronomers used to think that the solar system that emerged from this early evolution was similar to what we see today. Detailed recent studies of the orbits of the planets and asteroids, however, suggest that there were more violent events soon afterward, perhaps involving substantial changes in the orbits of Jupiter and Saturn. These two giant planets control, through their gravity, the distribution of asteroids. Working backward from our present solar system, it appears that orbital changes took place during the first few hundred million years. One consequence may have been scattering of asteroids into the inner solar system, causing the period of "heavy bombardment" recorded in the oldest lunar craters.

Answers

Exercise 3.3.2.1

The period of the rotating nebula is inversely proportional to D2D2. As we have just seen, PfinalPinitial=(DfinalDinitial)2.PfinalPinitial=(DfinalDinitial)2. Initially, we have $P_{initial} = 10^6$ yr and $D_{initial} = 10^4$ AU. Then, if D_{final} is in AU, P_{final} (in years) is given by Pfinal=0.01Dfinal2.Pfinal=0.01Dfinal2. If Jupiter's orbit has a radius of 5.2 AU, then the diameter is 10.4 AU. The period is then 1.08 years.

This page titled 3.3.2: Formation of the Solar System is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by OpenStax via source content that was edited to the style and standards of the LibreTexts platform.

3.3.3: Comparison with Other Planetary Systems

Learning Objectives

By the end of this section, you will be able to:

- Describe how the observations of protoplanetary disks provides evidence for the existence of other planetary systems
- Explain the two primary methods for detection of exoplanets
- Compare the main characteristics of other planetary systems with the features of the solar system

Until the middle 1990s, the practical study of the origin of planets focused on our single known example—the solar system. Although there had been a great deal of speculation about planets circling other stars, none had actually been detected. Logically enough, in the absence of data, most scientists assumed that our own system was likely to be typical. They were in for a big surprise.

Discovery of Other Planetary Systems

In The Birth of Stars and the Discovery of Planets outside the Solar System, we discuss the formation of stars and planets in some detail. Stars like our Sun are formed when dense regions in a molecular cloud (made of gas and dust) feel an extra gravitational force and begin to collapse. This is a runaway process: as the cloud collapses, the gravitational force gets stronger, concentrating material into a protostar. Roughly half of the time, the protostar will fragment or be gravitationally bound to other protostars, forming a binary or multiple star system—stars that are gravitationally bound and orbit each other. The rest of the time, the protostar collapses in isolation, as was the case for our Sun. In all cases, as we saw, conservation of angular momentum results in a spin-up of the collapsing protostar, with surrounding material flattened into a disk. Today, this kind of structure can actually be observed. The Hubble Space Telescope, as well as powerful new ground-based telescopes, enable astronomers to study directly the nearest of these *circumstellar disks* in regions of space where stars are being born today, such as the Orion Nebula (Figure 14.14) or the Taurus star-forming region.

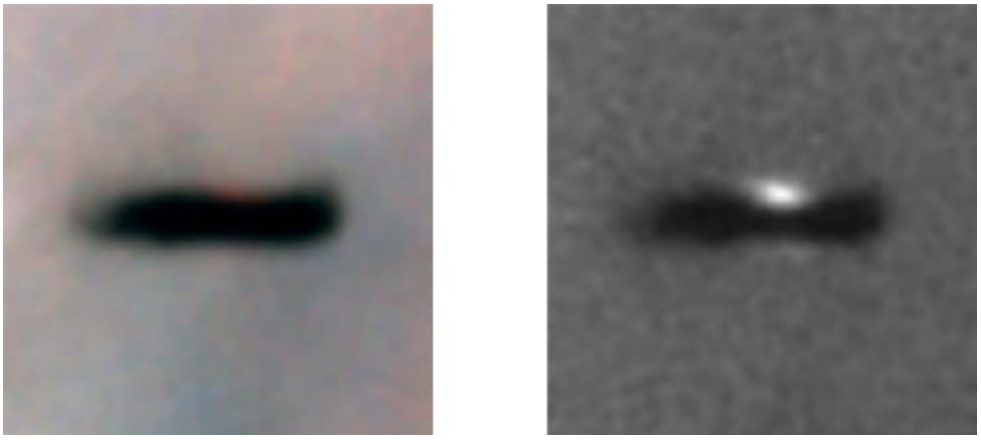
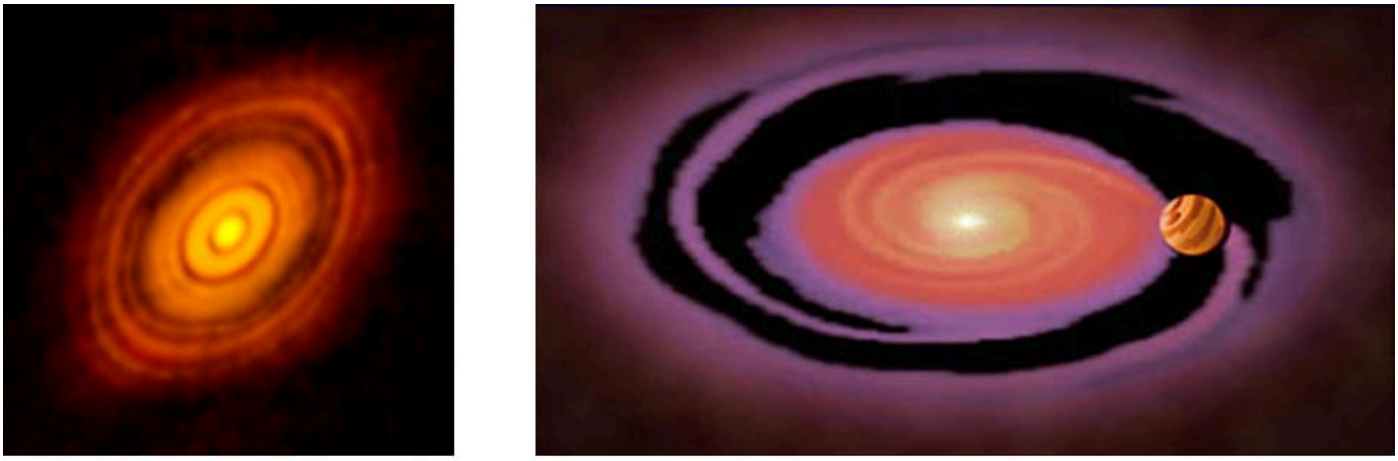


Figure 14.14 : Protoplanetary Disk in the Orion Nebula. The Hubble Space Telescope imaged this protoplanetary disk in the Orion Nebula, a region of active star formation, using two different filters. The disk, about 17 times the size of our solar system, is in an edge-on orientation to us, and the newly formed star is shining at the center of the flattened dust cloud. The dark areas indicate absorption, not an absence of material. In the left image we see the light of the nebula and the dark cloud; in the right image, a special filter was used to block the light of the background nebula. You can see gas above and below the disk set to glow by the light of the newborn star hidden by the disk. (credit: modification of work by Mark McCaughrean (Max-Planck-Institute for Astronomy), C. Robert O'Dell (Rice University), and NASA)

Many of the circumstellar disks we have discovered show internal structure. The disks appear to be donut-shaped, with gaps close to the star. Such gaps indicate that the gas and dust in the disk have already collapsed to form large planets (Figure 14.15). The newly born protoplanets are too small and faint to be seen directly, but the depletion of raw materials in the gaps hints at

the presence of something invisible in the inner part of the circumstellar disk—and that something is almost certainly one or more planets. Theoretical models of planet formation, like the one seen at right in Figure 14.15, have long supported the idea that planets would clear gaps as they form in disks.



(a) (b)

Figure 14.15 : Protoplanetary Disk around HL Tau. (a) This image of a protoplanetary disk around HL Tau was taken with the Atacama Large Millimeter/submillimeter Array (ALMA), which allows astronomers to construct radio images that rival those taken with visible light. (b) Newly formed planets that orbit the central star clear out dust lanes in their paths, just as our theoretical models predict. This computer simulation shows the empty lane and spiral density waves that result as a giant planet is forming within the disk. The planet is not shown to scale. (credit a: modification of work by ALMA (ESO/NAOJ/NRAO); credit b: modification of work by NASA/ESA and A. Feild (STScI))

Our figure shows HL Tau, a one-million-year-old "newborn" star in the Taurus star-forming region. The star is embedded in a shroud of dust and gas that obscures our visible-light view of a circumstellar disk around the star. In 2014 astronomers obtained a dramatic view of the HL Tau circumstellar disk using millimeter waves, which pierce the cocoon of dust around the star, showing dust lanes being carved out by several newly formed protoplanets. As the mass of the protoplanets increases, they travel in their orbits at speeds that are faster than the dust and gas in the circumstellar disk. As the protoplanets plow through the disk, their gravitational reach begins to exceed their cross-sectional area, and they become very efficient at sweeping up material and growing until they clear a gap in the disk. The image of Figure 14.15 shows us that a number of protoplanets are forming in the disk and that they were able to form faster than our earlier ideas had suggested—all in the first million years of star formation.

Link to Learning

For an explanation of ALMA's ground-breaking observations of HL Tau and what they reveal about plant formation, watch this videocast from the European Southern Observatory.

Discovering Exoplanets

You might think that with the advanced telescopes and detectors astronomers have today, they could directly image planets around nearby stars (which we call exoplanets). This has proved extremely difficult, however, not only because the exoplanets are faint, but also because they are generally lost in the brilliant glare of the star they orbit. As we discuss in more detail in The Birth of Stars and the Discovery of Planets outside the Solar System, the detection techniques that work best are indirect: they observe the effects of the planet on the star it orbits, rather than seeing the planet itself.

The first technique that yielded many planet detections is very high-resolution stellar spectroscopy. The *Doppler effect* lets astronomers measure the star's *radial velocity:* that is, the speed of the star, toward us or away from us, relative to the observer. If there is a massive planet in orbit around the star, the gravity of the planet causes the star to wobble, changing its radial velocity by a small but detectable amount. The distance of the star does not matter, as long as it is bright enough for us to take very high quality spectra.

Measurements of the variation in the star's radial velocity as the planet goes around the star can tell us the mass and orbital period of the planet. If there are several planets present, their effects on the radial velocity can be disentangled, so the entire planetary system can be deciphered—as long as the planets are massive enough to produce a measureable Doppler effect. This detection technique is most sensitive to large planets orbiting close to the star, since these produce the greatest wobble in their stars. It has been used on large ground-based telescopes to detect hundreds of planets, including one around Proxima Centauri, the nearest star to the Sun.

The second indirect technique is based on the slight dimming of a star when one of its planets *transits*, or crosses over the face of the star, as seen from Earth. Astronomers do not see the planet, but only detect its presence from careful measurements of a change in the brightness of the star over long periods of time. If the slight dips in brightness repeat at regular intervals, we can determine the orbital period of the planet. From the amount of starlight obscured, we can measure the planet's size.

While some transits have been measured from Earth, large-scale application of this transit technique requires a telescope in space, above the atmosphere and its distortions of the star images. It has been most successfully applied from the NASA Kepler space observatory, which was built for the sole purpose of "staring" for 5 years at a single part of the sky, continuously monitoring the light from more than 150,000 stars. The primary goal of Kepler was to determine the frequency of occurrence of exoplanets of different sizes around different classes of stars. Like the Doppler technique, the transit observations favor discovery of large planets and short-period orbits. The Kepler observations closed down in 2018 when the spacecraft ran out of the fuel needed to point the telescope.

Recent detection of exoplanets using both the Doppler and transit techniques has been incredibly successful. Within two decades, we went from no knowledge of other planetary systems to a catalog of *thousands* of exoplanets. Most of the exoplanets found so far are more massive than or larger in size than Earth. It is not that Earth analogs do not exist. Rather, the shortage of small rocky planets is an observational bias: smaller planets are more difficult to detect.

Analyses of the data to correct for such biases or selection effects indicate that small planets (like the terrestrial planets in our system) are actually much more common than giant planets. Also relatively common are "super Earths," planets with two to ten times the mass of our planet (Figure 14.16). We don't have any of these in our solar system, but nature seems to have no trouble making them elsewhere. Overall, the Kepler data suggest that approximately one quarter of stars have exoplanet systems, implying the existence of at least 50 billion planets in our Galaxy alone.

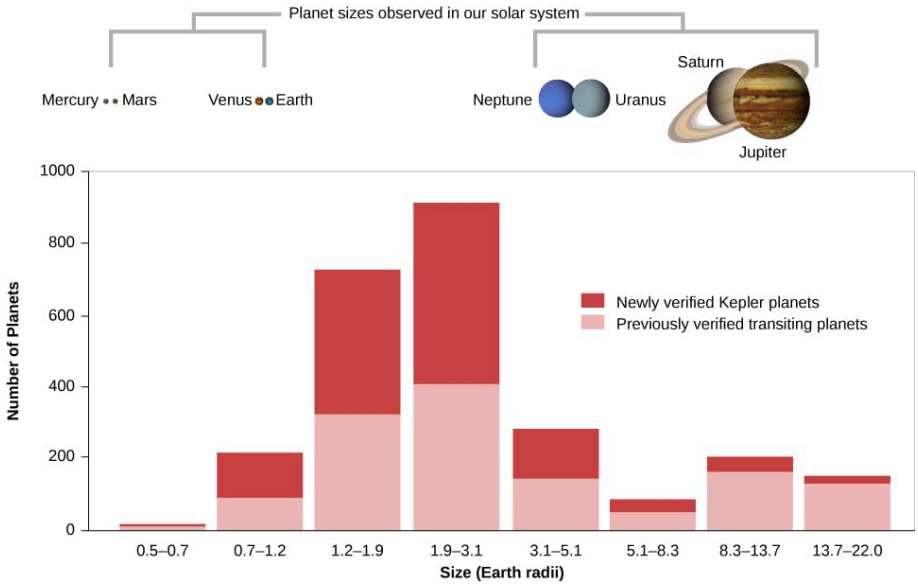


Figure 14.16 : Transiting Planets by Size. This bar graph shows the planets found so far using the transit method (the vast majority found by the Kepler mission). The orange parts of each bar indicate the planets announced by the Kepler team in May 2016. Note that the largest number of planets found so far are in two categories that we don't have in our own solar system—planets whose size is between Earth's and Neptune's. (credit: modification of work by NASA)

The Configurations of Other Planetary Systems

Let's look more closely at the progress in the detection of exoplanets. Figure 14.17 shows the planets that were discovered each

year by the two techniques we discussed. In the early years of exoplanet discovery, most of the planets were similar in mass to Jupiter. This is because, as mentioned above, the most massive planets were easiest to detect. In more recent years, planets smaller than Neptune and even close to the size of Earth have been detected.

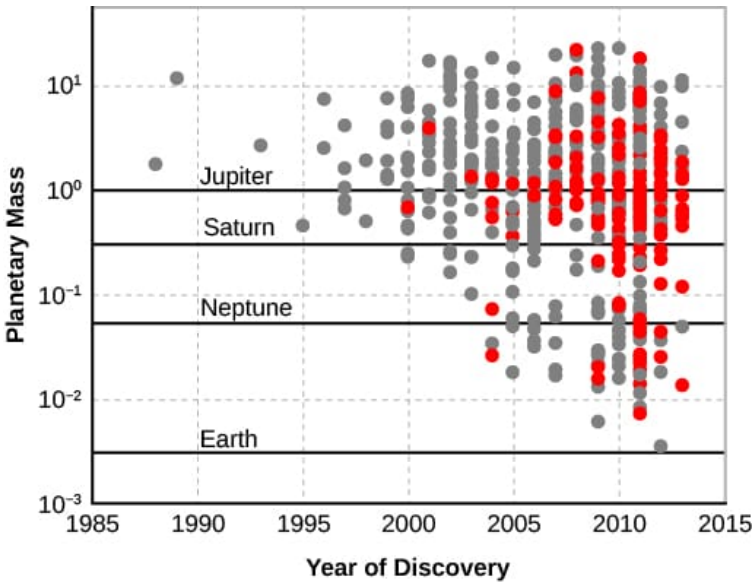


Figure 14.17 : Masses of Exoplanets Discovered by Year. Horizontal lines are drawn to reference the masses of Jupiter, Saturn, Neptune, and Earth. The gray dots indicate planets discovered by measuring the radial velocity of the star, and the red dots are for planets that transit their stars. In the early years, the only planets that could be detected were similar in mass to Jupiter. Improvements in technology and observing strategies enabled the detection of lower mass planets as time went on, and now even smaller worlds are being found. (Note that this tally ends in 2014, but the trend should be clear.)

We also know that many exoplanets are in multiplanet systems. This is one characteristic that our solar system shares with exosystems. Looking back at Figure 14.15 and seeing how such large disks can give rise to more than one center of condensation, it is not too surprising that multiplanet systems are a typical outcome of planet formation. Astronomers have tried to measure whether multiple planet systems all lie in the same plane using astrometry. This is a difficult measurement to make with current technology, but it is an important measurement that could help us understand the origin and evolution of planetary systems.

Comparison between Theory and Data

Many of the planetary systems discovered so far do not resemble our own solar system. Consequently, we have had to reassess some aspects of the “standard models” for the formation of planetary systems. Science sometimes works in this way, with new data contradicting our expectations. The press often talks about a scientist making experiments to “confirm” a theory. Indeed, it is comforting when new data support a hypothesis or theory and increase our confidence in an earlier result. But the most exciting and productive moments in science often come when new data *don’t* support existing theories, forcing scientists to rethink their position and develop new and deeper insights into the way nature works.

Nothing about the new planetary systems contradicts the basic idea that planets form from the aggregation (clumping) of material within circumstellar disks. However, the existence of “hot Jupiters”—planets of jovian mass that are closer to their stars than the orbit of Mercury—poses the biggest problem. As far as we know, a giant planet cannot be formed without the condensation of water ice, and water ice is not stable so close to the heat of a star. It seems likely that all the giant planets, “hot” or “normal,” formed at a distance of several astronomical units from the star, but we now see that they did not necessarily stay there. This discovery has led to a revision in our understanding of planet formation that now includes “planet migrations” within the protoplanetary disk, or later gravitational encounters between sibling planets that scatter one of the planets inward.

Many exoplanets have large orbital eccentricity (recall this means the orbits are not circular). High eccentricities were not expected for planets that form in a disk. This discovery provides further support for the scattering of planets when they interact gravitationally. When planets change each other’s motions, their orbits could become much more eccentric than the ones with which they began.

There are several suggestions for ways migration might have occurred. Most involve interactions between the giant planets and

the remnant material in the circumstellar disk from which they formed. These interactions would have taken place when the system was very young, while material still remained in the disk. In such cases, the planet travels at a faster velocity than the gas and dust and feels a kind of “headwind” (or friction) that causes it to lose energy and spiral inward. It is still unclear how the spiraling planet stops before it plunges into the star. Our best guess is that this plunge into the star is the fate for many protoplanets; however, clearly some migrating planets can stop their inward motions and escape this destruction, since we find hot Jupiters in many mature planetary systems.

This page titled 3.3.3: Comparison with Other Planetary Systems is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by OpenStax via source content that was edited to the style and standards of the LibreTexts platform.

4: LIGHT AND MATTER

4.1: The Behavior of Light

4.2: The Electromagnetic Spectrum

4.3: Spectroscopy in Astronomy

4.4: The Structure of the Atom

4.5: Formation of Spectral Lines

4.6: The Doppler Effect

Chapter 4: Light and Matter — Introduction

Almost everything we know about the universe beyond Earth comes from light. Astronomers cannot travel to stars and galaxies to collect samples or make measurements directly. Instead, they study the radiation that reaches us from these distant objects.

Light carries much more information than we can see with our eyes. By examining different wavelengths of electromagnetic radiation and the patterns produced when light interacts with matter, astronomers can determine properties such as an object's temperature, composition, and motion. To understand how this works, we also need to look at the structure of atoms and the ways that atoms absorb and emit energy.

In this chapter, we will explore the behavior of light, the electromagnetic spectrum, and the formation of spectra. These ideas provide some of the most important tools astronomers use to investigate objects throughout the universe and will form the foundation for our study of stars in the next chapter.

This page titled 4: Light and Matter is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Robert Wagner.

4.1: The Behavior of Light

Learning Objectives

By the end of this section, you will be able to:

- Explain the evidence for Maxwell's electromagnetic model of light
- Describe the relationship between wavelength, frequency, and speed of light
- Discuss the particle model of light and the definition of photon
- Explain how and why the amount of light we see from an object depends upon its distance

Coded into the light and other kinds of radiation that reach us from objects in the universe is a wide range of information about what those objects are like and how they work. If we can decipher this code and read the messages it contains, we can learn an enormous amount about the cosmos without ever having to leave Earth or its immediate environment.

The visible light and other radiation we receive from the stars and planets is generated by processes at the atomic level—by changes in the way the parts of an atom interact and move. Thus, to appreciate how light is generated, we must explore how atoms work. There is a bit of irony in the fact that in order to understand some of the largest structures in the universe, we must become acquainted with some of the smallest.

Notice that we have twice used the phrase "light and other radiation." One of the key ideas explored in this chapter is that visible light is not unique; it is merely the most familiar example of a much larger family of radiation that can carry information to us.

The word "radiation" will be used frequently in this book, so it is important to understand what it means. In everyday language, "radiation" is often used to describe certain kinds of energetic subatomic particles released by radioactive materials in our environment. (An example is the kind of radiation used to treat some cancers.) But this is not what we mean when we use the word "radiation" in an astronomy text. *Radiation*, as used in this book, is a general term for waves (including light waves) that *radiate* outward from a source.

As we saw in Orbits and Gravity, Newton's theory of gravity accounts for the motions of planets as well as objects on Earth. Application of this theory to a variety of problems dominated the work of scientists for nearly two centuries. In the nineteenth century, many physicists turned to the study of electricity and magnetism, which are intimately connected with the production of light.

The scientist who played a role in this field comparable to Newton's role in the study of gravity was physicist James Clerk Maxwell, born and educated in Scotland (Figure 5.2). Inspired by a number of ingenious experiments that showed an intimate relationship between electricity and magnetism, Maxwell developed a theory that describes both electricity and magnetism with only a small number of elegant equations. It is this theory that gives us important insights into the nature and behavior of light.



Figure 5.2 : James Clerk Maxwell (1831–1879). Maxwell unified the rules governing electricity and magnetism into a coherent theory.

Maxwell's Theory of Electromagnetism

We will look at the structure of the atom in more detail later, but we begin by noting that the typical atom consists of several types of particles, a number of which have not only mass but an additional property called electric charge. In the nucleus (central part) of every atom are *protons*, which are positively charged; outside the nucleus are electrons, which have a negative charge.

Maxwell's theory deals with these electric charges and their effects, especially when they are moving. In the vicinity of an electron charge, another charge feels a force of attraction or repulsion: opposite charges attract; like charges repel. When charges are not in motion, we observe only this electric attraction or repulsion. If charges are in motion, however (as they are inside every atom and in a wire carrying a current), then we measure another force called *magnetism*.

Magnetism was well known for much of recorded human history, but its cause was not understood until the nineteenth century. Experiments with electric charges demonstrated that magnetism was the result of moving charged particles. Sometimes, the motion is clear, as in the coils of heavy wire that make an industrial electromagnet. Other times, it is more subtle, as in the kind of magnet you buy in a hardware store, in which many of the electrons inside the atoms are spinning in roughly the same direction; it is the alignment of their motion that causes the material to become magnetic.

Physicists use the word *field* to describe the action of forces that one object exerts on other distant objects. For example, we say the Sun produces a *gravitational field* that controls Earth's orbit, even though the Sun and Earth do not come directly into contact. Using this terminology, we can say that stationary electric charges produce *electric fields*, and moving electric charges also produce *magnetic fields*.

Actually, the relationship between electric and magnetic phenomena is even more profound. Experiments showed that changing magnetic fields could produce electric currents (and thus changing electric fields), and changing electric currents could in turn produce changing magnetic fields. So once begun, electric and magnetic field changes could continue to trigger each other.

Maxwell analyzed what would happen if electric charges were oscillating (moving constantly back and forth) and found that the resulting pattern of electric and magnetic fields would spread out and travel rapidly through space. Something similar happens when a raindrop strikes the surface of water or a frog jumps into a pond. The disturbance moves outward and creates a pattern we call a *wave* in the water (Figure 5.3). You might, at first, think that there must be very few situations in nature where electric charges oscillate, but this is not at all the case. As we shall see, atoms and molecules (which consist of charged particles) oscillate back and forth all the time. The resulting electromagnetic disturbances are among the most common phenomena in the universe.



Figure 5.3 : Making Waves. An oscillation in a pool of water creates an expanding disturbance called a wave. (credit: modification of work by "vastateparksstaff"/Flickr)

Maxwell was able to calculate the speed at which an electromagnetic disturbance moves through space; he found that it is equal to the speed of light, which had been measured experimentally. On that basis, he speculated that light was one form of a family of possible electromagnetic disturbances called electromagnetic radiation, a conclusion that was again confirmed in laboratory experiments. When light (reflected from the pages of an astronomy textbook, for example) enters a human eye, its changing electric and magnetic fields stimulate nerve endings, which then transmit the information contained in these changing fields to the brain. The science of astronomy is primarily about analyzing radiation from distant objects to understand what they are and how they work.

The Wave-Like Characteristics of Light

The changing electric and magnetic fields in light are similar to the waves that can be set up in a quiet pool of water. In both cases, the disturbance travels rapidly outward from the point of origin and can use its energy to disturb other things farther away. (For example, in water, the expanding ripples moving away from our frog could disturb the peace of a dragonfly resting on a leaf in the same pool.) In the case of electromagnetic waves, the radiation generated by a transmitting antenna full of charged particles and moving electrons at your local radio station can, sometime later, disturb a group of electrons in your car radio antenna and bring you the news and weather while you are driving to class or work in the morning.

The waves generated by charged particles differ from water waves in some profound ways, however. Water waves require water to travel in. The sound waves we hear, to give another example, are pressure disturbances that require air to travel though. But electromagnetic waves do not require water or air: the fields generate each other and so can move through a vacuum (such as outer space). This was such a disturbing idea to nineteenth-century scientists that they actually made up a substance to fill all of space—one for which there was not a single shred of evidence—just so light waves could have something to travel through: they called it the *aether*. Today, we know that there is no aether and that electromagnetic waves have no trouble at all moving through empty space (as all the starlight visible on a clear night must surely be doing).

The other difference is that *all* electromagnetic waves move at the same speed in empty space (the speed of light—approximately 300,000 kilometers per second, or 300,000,000 meters per second, which can also be written as 3×108 m/s3×108 m/s), which turns out to be the fastest possible speed in the universe. No matter where electromagnetic waves are generated from and no matter what other properties they have, when they are moving (and not interacting with matter), they move at the speed of light. Yet you know from everyday experience that there are different kinds of light. For example, we perceive that light waves differ from one another in a property we call color. Let's see how we can denote the differences among the whole broad family of electromagnetic waves.

The nice thing about a wave is that it is a repeating phenomenon. Whether it is the up-and-down motion of a water wave or the changing electric and magnetic fields in a wave of light, the pattern of disturbance repeats in a cyclical way. Thus, any wave motion can be characterized by a series of crests and troughs (Figure 5.4). Moving from one crest through a trough to the next

crest completes one cycle. The horizontal length covered by one cycle is called the wavelength. Ocean waves provide an analogy: the wavelength is the distance that separates successive wave crests.

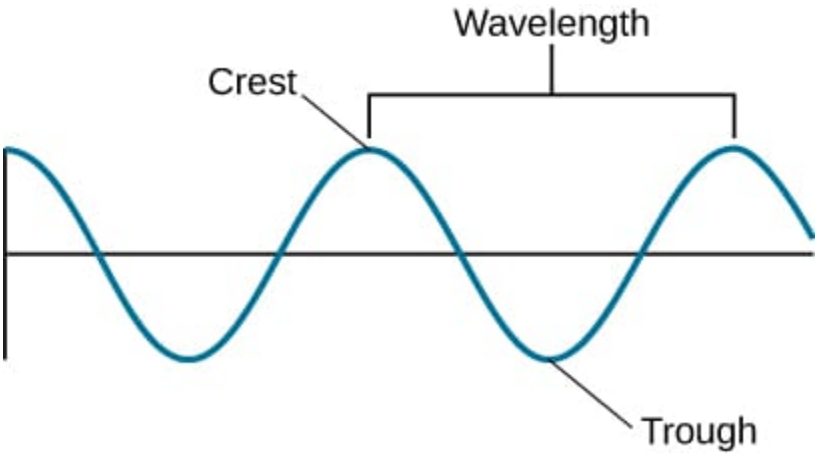


Figure 5.4 : Characterizing Waves. Electromagnetic radiation has wave-like characteristics. The wavelength (λ) is the distance between crests, the frequency (*f*) is the number of cycles per second, and the speed (*c*) is the distance the wave covers during a specified period of time (e.g., kilometers per second).

For visible light, our eyes perceive different wavelengths as different colors: red, for example, is the longest visible wavelength, and violet is the shortest. The main colors of visible light from longest to shortest wavelength can be remembered using the mnemonic ROY G BIV—for Red, Orange, Yellow, Green, Blue, Indigo, and Violet. Other invisible forms of electromagnetic radiation have different wavelengths, as we will see in the next section.

We can also characterize different waves by their frequency, the number of wave cycles that pass by per second. If you count 10 crests moving by each second, for example, then the frequency is 10 cycles per second (cps). In honor of Heinrich Hertz, the physicist who—inspired by Maxwell's work—discovered radio waves, a cps is also called a *hertz* (Hz). Take a look at your radio, for example, and you will see the channel assigned to each radio station is characterized by its frequency, usually in units of KHz (kilohertz, or thousands of hertz) or MHz (megahertz, or millions of hertz).

Wavelength (λ) and frequency (*f*) are related because all electromagnetic waves travel at the same speed. To see how this works, imagine a parade in which everyone is forced by prevailing traffic conditions to move at exactly the same speed. You stand on a corner and watch the waves of marchers come by. First you see row after row of miniature ponies. Because they are not very large and, therefore, have a shorter wavelength, a good number of the ponies can move past you each minute; we can say they have a high frequency. Next, however, come several rows of circus elephants. The elephants are large and marching at the same speed as the ponies, so far fewer of them can march past you per minute: Because they have a wider spacing (longer wavelength), they represent a lower frequency.

The formula for this relationship can be expressed as follows: for any wave motion, the speed at which a wave moves equals the frequency times the wavelength. Waves with longer wavelengths have lower frequencies. Mathematically, we can express this as

c=λfc=λf

where the Greek letter for "l"—lambda, λ—is used to denote wavelength and *c* is the scientific symbol for the speed of light. Solving for the wavelength, this is expressed as:

λ=cf.λ=cf.

Example 5.1: Deriving and Using the Wave Equation

The equation for the relationship between the speed and other characteristics of a wave can be derived from our basic understanding of motion. The average speed of anything that is moving is:

average speed=distancetimeaverage speed=distancetime

(So, for example, a car on the highway traveling at a speed of 100 km/h covers 100 km during the time of 1 h.) For an electromagnetic wave to travel the distance of one of its wavelengths, λ, at the speed of light, *c*, we have c=λtc=λt. The frequency of a wave is the number of cycles per second. If a wave has a frequency of a million cycles per second, then the time for each cycle to go by is a millionth of a second. So, in general, t=1ft=1f. Substituting into our wave equation, we get c=λ×fc=λ×f. Now let's use this to calculate an example. What is the wavelength of visible light that has a frequency of 5.66×1014 Hz5.66×1014 Hz?

Solution

Solving the wave equation for wavelength, we find:

λ=cfλ=cf

Substituting our values gives:

λ=3.00×108m/s5.66×1014Hz=5.30×10−7mλ=3.00×108m/s5.66×1014Hz=5.30×10−7m

This answer can also be written as 530 nm, which is in the yellow-green part of the visible spectrum (nm stands for nanometers, where the term "nano" means "billionths").

Exercise 4.1.1

"Tidal waves," or tsunamis, are waves caused by earthquakes that travel rapidly through the ocean. If a tsunami travels at the speed of 600 km/h and approaches a shore at a rate of one wave crest every 15 min (4 waves/h), what would be the distance between those wave crests at sea?

Answer available at the end of this section.

Light as a Photon

The electromagnetic wave model of light (as formulated by Maxwell) was one of the great triumphs of nineteenth-century science. In 1887, when Heinrich Hertz actually made invisible electromagnetic waves (what today are called radio waves) on one side of a room and detected them on the other side, it ushered in a new era that led to the modern age of telecommunications. His experiment ultimately led to the technologies of television, cell phones, and today's wireless networks around the globe.

However, by the beginning of the twentieth century, more sophisticated experiments had revealed that light behaves in certain ways that cannot be explained by the wave model. Reluctantly, physicists had to accept that sometimes light behaves more like a "particle"—or at least a self-contained packet of energy—than a wave. We call such a packet of electromagnetic energy a photon.

The fact that light behaves like a wave in certain experiments and like a particle in others was a very surprising and unlikely idea. After all, our common sense says that waves and particles are opposite concepts. On one hand, a wave is a repeating disturbance that, by its very nature, is not in only one place, but spreads out. A particle, on the other hand, is something that can be in only one place at any given time. Strange as it sounds, though, countless experiments now confirm that electromagnetic radiation can sometimes behave like a wave and at other times like a particle.

Then, again, perhaps we shouldn't be surprised that something that always travels at the "speed limit" of the universe and doesn't need a medium to travel through might not obey our everyday common sense ideas. The confusion that this wave-particle duality of light caused in physics was eventually resolved by the introduction of a more complicated theory of waves and particles, now called quantum mechanics. (This is one of the most interesting fields of modern science, but it is mostly beyond the scope of our book. If you are interested in it, see some of the suggested resources at the end of this chapter.)

In any case, you should now be prepared when scientists (or the authors of this book) sometimes discuss electromagnetic radiation as if it consisted of waves and at other times refer to it as a stream of photons. A photon (being a packet of energy) carries a specific amount of energy. We can use the idea of energy to connect the photon and wave models. How much energy a photon has depends on its frequency when you think about it as a wave. A low-energy radio wave has a low frequency as a wave, while a high-energy X-ray at your dentist's office is a high-frequency wave. Among the colors of visible light, violet-light photons

have the highest energy and red-light photons have the lowest.

Test whether the connection between photons and waves is clear to you. In the above example, which photon would have the longer wavelength as a wave: the radio wave or the X-ray? If you answered the radio wave, you are correct. Radio waves have a lower frequency, so the wave cycles are longer (they are elephants, not miniature ponies).

Propagation of Light

Let's think for a moment about how light from a lightbulb moves through space. As waves expand, they travel away from the bulb, not just toward your eyes but in all directions. They must therefore cover an ever-widening space. Yet the total amount of light available can't change once the light has left the bulb. This means that, as the same expanding shell of light covers a larger and larger area, there must be less and less of it in any given place. Light (and all other electromagnetic radiation) gets weaker and weaker as it gets farther from its source.

The increase in the area that the light must cover is proportional to the square of the distance that the light has traveled (Figure 5.5). If we stand twice as far from the source, our eyes will intercept two-squared (2 × 2), or four times less light. If we stand 10 times farther from the source, we get 10-squared, or 100 times less light. You can see how this weakening means trouble for sources of light at astronomical distances. One of the nearest stars, Alpha Centauri A, emits about the same total energy as the Sun. But it is about 270,000 times farther away, and so it appears about 73 billion times fainter. No wonder the stars, which close-up would look more or less like the Sun, look like faint pinpoints of light from far away.

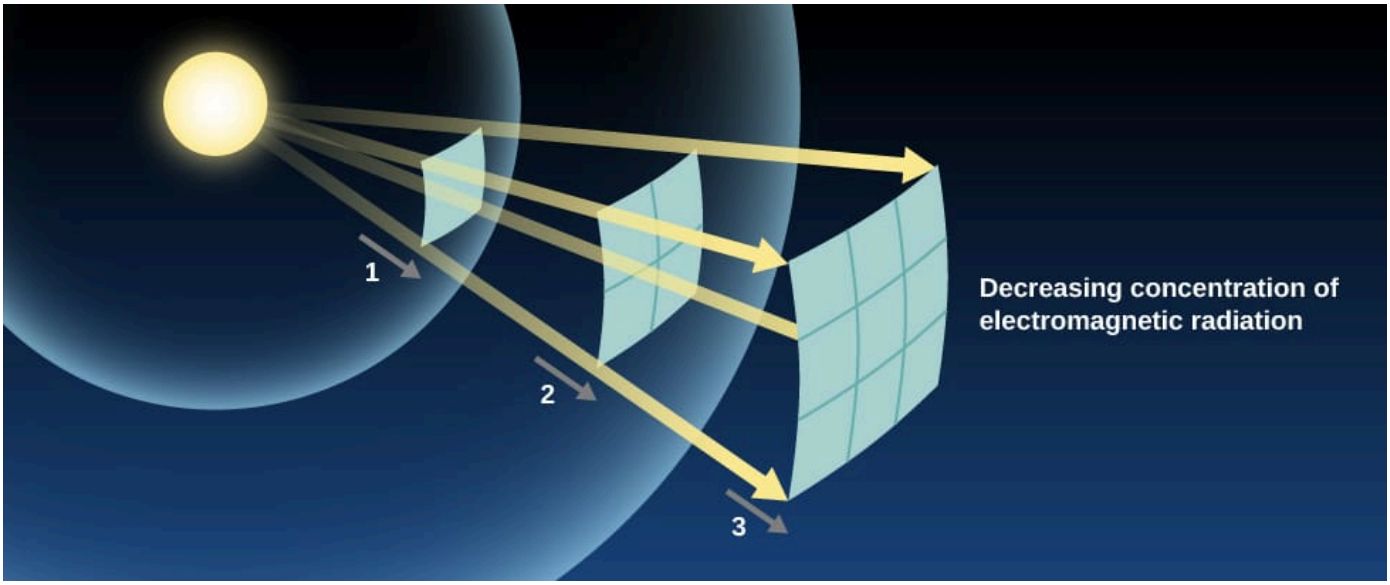


Figure 5.5 : Inverse Square Law for Light. As light radiates away from its source, it spreads out in such a way that the energy per unit area (the amount of energy passing through one of the small squares) decreases as the square of the distance from its source.

This idea—that the apparent brightness of a source (how bright it looks to us) gets weaker with distance in the way we have described—is known as the inverse square law for light propagation. In this respect, the propagation of light is similar to the effects of gravity. Remember that the force of gravity between two attracting masses is also inversely proportional to the square of their separation.

Example 5.2: The Inverse Square Law for Light

The intensity of a 120-W lightbulb observed from a distance 2 m away is 2.4 W/m^2. What would be the intensity if this distance was doubled?

Solution

If we move twice as far away, then the answer will change according to the inverse square of the distance, so the new intensity will be (12)2=14(12)2=14 of the original intensity, or 0.6 W/m^2.

Exercise 4.1.1

How many times brighter or fainter would a star appear if it were moved to:

a. twice its present distance?
b. ten times its present distance?
c. half its present distance?

Answer available at the end of this section.

Answers

Exercise 4.1.1

λ=600 km/h4 waves/h=150kmλ=600 km/h4 waves/h=150km

Exercise 4.1.1

a. (12)2=14(12)2=14; b. (110)2=1100(110)2=1100; c. (11/2)2=4(11/2)2=4

This page titled 4.1: The Behavior of Light is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by OpenStax via source content that was edited to the style and standards of the LibreTexts platform.

4.2: The Electromagnetic Spectrum

Learning Objectives

By the end of this section, you will be able to:

- Understand the bands of the electromagnetic spectrum and how they differ from one another
- Understand how each part of the spectrum interacts with Earth's atmosphere
- Explain how and why the light emitted by an object depends on its temperature

Objects in the universe send out an enormous range of electromagnetic radiation. Scientists call this range the electromagnetic spectrum, which they have divided into a number of categories. The spectrum is shown in Figure 5.6, with some information about the waves in each part or band.

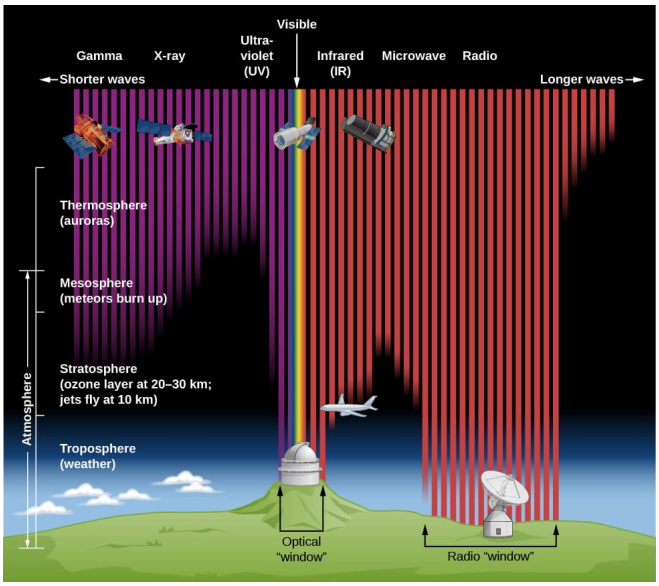


Figure 5.6 : Radiation and Earth's Atmosphere. This figure shows the bands of the electromagnetic spectrum and how well Earth's atmosphere transmits them. Note that high-frequency waves from space do not make it to the surface and must therefore be observed from space. Some infrared and microwaves are absorbed by water and thus are best observed from high altitudes. Low-frequency radio waves are blocked by Earth's ionosphere. (credit: modification of work by STScI/JHU/NASA)

Types of Electromagnetic Radiation

Electromagnetic radiation with the shortest wavelengths, no longer than 0.01 nanometer, is categorized as gamma rays (1 nanometer = 10^{-9} meters; see Appendix D). The name *gamma* comes from the third letter of the Greek alphabet: gamma rays were the third kind of radiation discovered coming from radioactive atoms when physicists first investigated their behavior. Because gamma rays carry a lot of energy, they can be dangerous for living tissues. Gamma radiation is generated deep in the interior of stars, as well as by some of the most violent phenomena in the universe, such as the deaths of stars and the merging of stellar corpses. Gamma rays coming to Earth are absorbed by our atmosphere before they reach the ground (which is a good thing for our health); thus, they can only be studied using instruments in space.

Electromagnetic radiation with wavelengths between 0.01 nanometer and 20 nanometers is referred to as X-rays. Being more energetic than visible light, X-rays are able to penetrate soft tissues but not bones, and so allow us to make images of the shadows of the bones inside us. While X-rays can penetrate a short length of human flesh, they are stopped by the large numbers of atoms in Earth's atmosphere with which they interact. Thus, X-ray astronomy (like gamma-ray astronomy) could not develop until we invented ways of sending instruments above our atmosphere (Figure 5.7).

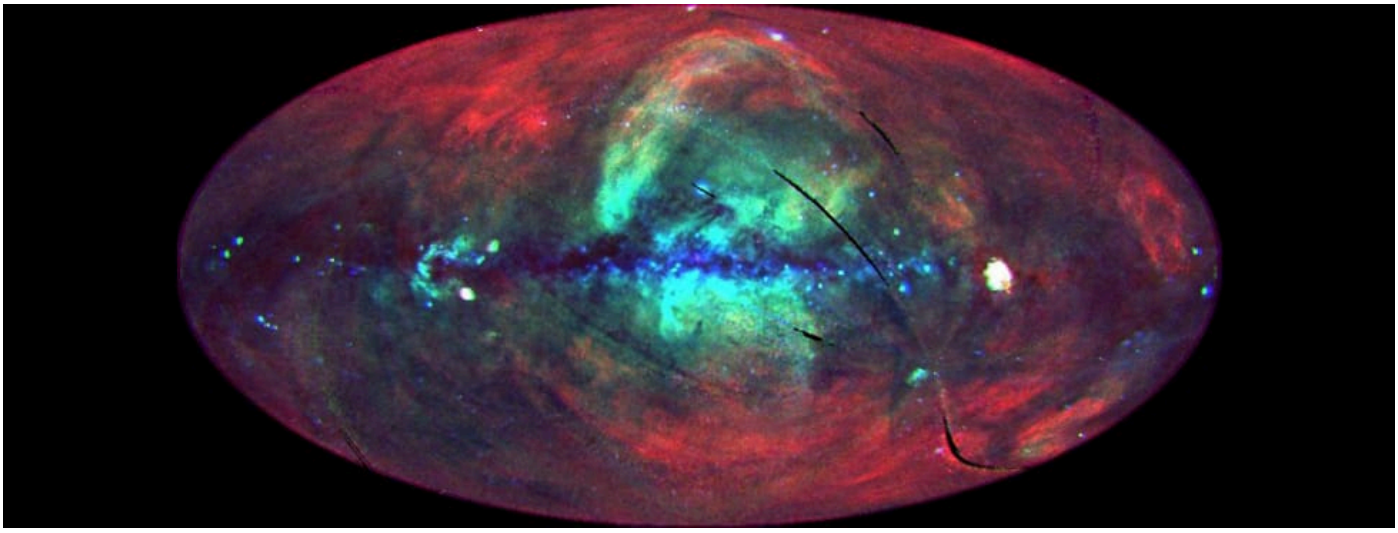


Figure 5.7 : X-Ray Sky. This is a map of the sky tuned to certain types of X-rays (seen from above Earth's atmosphere). The map tilts the sky so that the disk of our Milky Way Galaxy runs across its center. It was constructed and artificially colored from data gathered by the European ROSAT satellite. Each color (red, yellow, and blue) shows X-rays of different frequencies or energies. For example, red outlines the glow from a hot local bubble of gas all around us, blown by one or more exploding stars in our cosmic vicinity. Yellow and blue show more distant sources of X-rays, such as remnants of other exploded stars or the active center of our Galaxy (in the middle of the picture). (credit: modification of work by NASA)

Radiation intermediate between X-rays and visible light is ultraviolet (meaning higher energy than violet). Outside the world of science, ultraviolet light is sometimes called "black light" because our eyes cannot see it. Ultraviolet radiation is mostly blocked by the ozone layer of Earth's atmosphere, but a small fraction of ultraviolet rays from our Sun do penetrate to cause sunburn or, in extreme cases of overexposure, skin cancer in human beings. Ultraviolet astronomy is also best done from space.

Electromagnetic radiation with wavelengths between roughly 400 and 700 nm is called visible light because these are the waves that human vision can perceive. This is also the band of the electromagnetic spectrum that most readily reaches Earth's surface. These two observations are not coincidental: human eyes evolved to see the kinds of waves that arrive from the Sun most effectively. Visible light penetrates Earth's atmosphere effectively, except when it is temporarily blocked by clouds.

Between visible light and radio waves are the wavelengths of infrared or heat radiation. Astronomer William Herschel first discovered infrared in 1800 while trying to measure the temperatures of different colors of sunlight spread out into a spectrum. He noticed that when he accidently positioned his thermometer beyond the reddest color, it still registered heating due to some invisible energy coming from the Sun. This was the first hint about the existence of the other (invisible) bands of the electromagnetic spectrum, although it would take many decades for our full understanding to develop.

A heat lamp radiates mostly infrared radiation, and the nerve endings in our skin are sensitive to this band of the electromagnetic spectrum. Infrared waves are absorbed by water and carbon dioxide molecules, which are more concentrated low in Earth's atmosphere. For this reason, infrared astronomy is best done from high mountaintops, high-flying airplanes, and spacecraft.

After infrared comes the familiar microwave, used in short-wave communication and microwave ovens. (Wavelengths vary from 1 millimeter to 1 meter and are absorbed by water vapor, which makes them effective in heating foods.) The "micro-" prefix refers to the fact that microwaves are small in comparison to radio waves, the next on the spectrum. You may remember that tea—which is full of water—heats up quickly in your microwave oven, while a ceramic cup—from which water has been removed by baking—stays cool in comparison.

All electromagnetic waves longer than microwaves are called radio waves, but this is so broad a category that we generally divide it into several subsections. Among the most familiar of these are radar waves, which are used in radar guns by traffic officers to determine vehicle speeds, and AM radio waves, which were the first to be developed for broadcasting. The wavelengths of these different categories range from over a meter to hundreds of meters, and other radio radiation can have wavelengths as long as several kilometers.

With such a wide range of wavelengths, not all radio waves interact with Earth's atmosphere in the same way. FM and TV waves are not absorbed and can travel easily through our atmosphere. AM radio waves are absorbed or reflected by a layer in Earth's atmosphere called the ionosphere (the ionosphere is a layer of charged particles at the top of our atmosphere, produced by

interactions with sunlight and charged particles that are ejected from the Sun).

We hope this brief survey has left you with one strong impression: although visible light is what most people associate with astronomy, the light that our eyes can see is only a tiny fraction of the broad range of waves generated in the universe. Today, we understand that judging some astronomical phenomenon by using only the light we can see is like hiding under the table at a big dinner party and judging all the guests by nothing but their shoes. There's a lot more to each person than meets our eye under the table. It is very important for those who study astronomy today to avoid being "visible light chauvinists"—to respect only the information seen by their eyes while ignoring the information gathered by instruments sensitive to other bands of the electromagnetic spectrum.

Table 5.1 summarizes the bands of the electromagnetic spectrum and indicates the temperatures and typical astronomical objects that emit each kind of electromagnetic radiation. While at first, some of the types of radiation listed in the table may seem unfamiliar, you will get to know them better as your astronomy course continues. You can return to this table as you learn more about the types of objects astronomers study.

Table 5.1: Types of Electromagnetic Radiation

Type of Radiation	Wavelength Range (nm)	Radiated by Objects at This Temperature	Typical Sources
Gamma rays	Less than 0.01	More than 10^8 K	Produced in nuclear reactions; require very high-energy processes
X-rays	0.01–20	10^6–10^8 K	Gas in clusters of galaxies, supernova remnants, solar corona
Ultraviolet	20–400	10^4–10^6 K	Supernova remnants, very hot stars
Visible	400–700	10^3–10^4 K	Stars
Infrared	10^3–10^6	10–10^3 K	Cool clouds of dust and gas, planets, moons
Microwave	10^6–10^9	Less than 10 K	Active galaxies, pulsars, cosmic background radiation
Radio	More than 10^9	Less than 10 K	Supernova remnants, pulsars, cold gas

Radiation and Temperature

Some astronomical objects emit mostly infrared radiation, others mostly visible light, and still others mostly ultraviolet radiation. What determines the type of electromagnetic radiation emitted by the Sun, stars, and other dense astronomical objects? The answer often turns out to be their *temperature*.

At the microscopic level, everything in nature is in motion. A solid is composed of molecules and atoms in continuous vibration: they move back and forth in place, but their motion is much too small for our eyes to make out. A gas consists of atoms and/or molecules that are flying about freely at high speed, continually bumping into one another and bombarding the surrounding matter. The hotter the solid or gas, the more rapid the motion of its molecules or atoms. The temperature of something is thus a measure of the average motion energy of the particles that make it up.

This motion at the microscopic level is responsible for much of the electromagnetic radiation on Earth and in the universe. As atoms and molecules move about and collide, or vibrate in place, their electrons give off electromagnetic radiation. The characteristics of this radiation are determined by the temperature of those atoms and molecules. In a hot material, for example, the individual particles vibrate in place or move rapidly from collisions, so the emitted waves are, on average, more energetic. And recall that higher energy waves have a higher frequency. In very cool material, the particles have low-energy atomic and molecular motions and thus generate lower-energy waves.

Link to Learning

Check out the NASA briefing or NASA's 5-minute introductory video to learn more about the electromagnetic spectrum.

Radiation Laws

To understand, in more quantitative detail, the relationship between temperature and electromagnetic radiation, we imagine an idealized object called a blackbody. Such an object (unlike your sweater or your astronomy instructor's head) does not reflect or scatter any radiation, but absorbs all the electromagnetic energy that falls onto it. The energy that is absorbed causes the atoms and molecules in it to vibrate or move around at increasing speeds. As it gets hotter, this object will radiate electromagnetic waves until absorption and radiation are in balance. We want to discuss such an idealized object because, as you will see, stars behave in very nearly the same way.

The radiation from a blackbody has several characteristics, as illustrated in Figure 5.8. The graph shows the power emitted at each wavelength by objects of different temperatures. In science, the word *power* means the energy coming off per second (and it is typically measured in *watts*, which you are probably familiar with from buying lightbulbs).

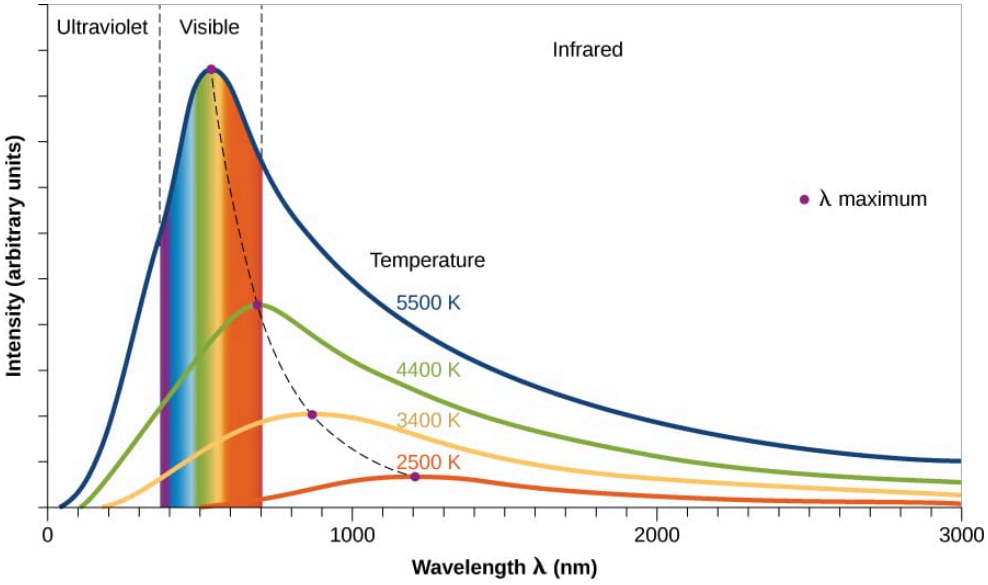


Figure 5.8 : Radiation Laws Illustrated. This graph shows in arbitrary units how many photons are given off at each wavelength for objects at four different temperatures. The wavelengths corresponding to visible light are shown by the colored bands. Note that at hotter temperatures, more energy (in the form of photons) is emitted at all wavelengths. The higher the temperature, the shorter the wavelength at which the peak amount of energy is radiated (this is known as Wien's law).

First of all, notice that the curves show that, at each temperature, our blackbody object emits radiation (photons) at all wavelengths (all colors). This is because in any solid or denser gas, some molecules or atoms vibrate or move between collisions slower than average and some move faster than average. So when we look at the electromagnetic waves emitted, we find a broad range, or spectrum, of energies and wavelengths. More energy is emitted at the average vibration or motion rate (the highest part of each curve), but if we have a large number of atoms or molecules, some energy will be detected at each wavelength.

Second, note that an object at a higher temperature emits more power at all wavelengths than does a cooler one. In a hot gas (the taller curves in Figure 5.8), for example, the atoms have more collisions and give off more energy. In the real world of stars, this means that hotter stars give off more energy at every wavelength than do cooler stars.

Third, the graph shows us that the higher the temperature, the shorter the wavelength at which the maximum power is emitted. Remember that a shorter wavelength means a higher frequency and energy. It makes sense, then, that hot objects give off a larger fraction of their energy at shorter wavelengths (higher energies) than do cool objects. You may have observed examples of this rule in everyday life. When a burner on an electric stove is turned on low, it emits only heat, which is infrared radiation, but does not glow with visible light. If the burner is set to a higher temperature, it starts to glow a dull red. At a still-higher setting, it glows a brighter orange-red (shorter wavelength). At even higher temperatures, which cannot be reached with ordinary stoves, metal can appear brilliant yellow or even blue-white.

We can use these ideas to come up with a rough sort of "thermometer" for measuring the temperatures of stars. Because many stars give off most of their energy in visible light, the color of light that dominates a star's appearance is a rough indicator of its temperature. If one star looks red and another looks blue, which one has the higher temperature? Because blue is the shorter-

wavelength color, it is the sign of a hotter star. (Note that the temperatures we associate with different colors in science are not the same as the ones artists use. In art, red is often called a “hot” color and blue a “cool” color. Likewise, we commonly see red on faucet or air conditioning controls to indicate hot temperatures and blue to indicate cold temperatures. Although these are common uses to us in daily life, in nature, it’s the other way around.)

We can develop a more precise star thermometer by measuring how much energy a star gives off at each wavelength and by constructing diagrams like Figure 5.8. The location of the peak (or maximum) in the power curve of each star can tell us its temperature. The average temperature at the surface of the Sun, which is where the radiation that we see is emitted, turns out to be 5800 K. (Throughout this text, we use the kelvin or absolute temperature scale. On this scale, water freezes at 273 K and boils at 373 K. All molecular motion ceases at 0 K. The various temperature scales are described in Appendix D.) There are stars cooler than the Sun and stars hotter than the Sun.

The wavelength at which maximum power is emitted can be calculated according to the equation

λmax=3×106Tλmax=3×106T

where the wavelength is in nanometers (one billionth of a meter) and the temperature is in K (the constant 3×10^6 has units of nm×KTable 5.1.

Example 5.3: Calculating the Temperature of a Blackbody

We can use Wien’s law to calculate the temperature of a star provided we know the wavelength of peak intensity for its spectrum. If the emitted radiation from a red dwarf star has a wavelength of maximum power at 1200 nm, what is the temperature of this star, assuming it is a blackbody?

Solution

Solving Wien’s law for temperature gives:

T=3×106nm Kλmax=3×106nm K1200 nm=2500KT=3×106nm Kλmax=3×106nm K1200 nm=2500K

Exercise 4.2.1

What is the temperature of a star whose maximum light is emitted at a much shorter wavelength of 290 nm?

Answer available at the end of this section.

Since this star has a peak wavelength that is at a shorter wavelength (in the ultraviolet part of the spectrum) than that of our Sun (in the visible part of the spectrum), it should come as no surprise that its surface temperature is much hotter than our Sun’s.

We can also describe our observation that hotter objects radiate more power at all wavelengths in a mathematical form. If we sum up the contributions from all parts of the electromagnetic spectrum, we obtain the total energy emitted by a blackbody. What we usually measure from a large object like a star is the energy flux, the power emitted per square meter. The word *flux* means “flow” here: we are interested in the flow of power into an area (like the area of a telescope mirror). It turns out that the energy flux from a blackbody at temperature *T* is proportional to the fourth power of its absolute temperature. This relationship is known as the Stefan-Boltzmann law and can be written in the form of an equation as

F=σT4F=σT4

where *F* stands for the energy flux (in units of watts per square meter), T is given in Kelvins, and σ (Greek letter sigma) is a constant number (5.67×10-8)(5.67×10-8).

Notice how impressive this result is. Increasing the temperature of a star would have a tremendous effect on the power it radiates. If the Sun, for example, were twice as hot—that is, if it had a temperature of 11,600 K—it would radiate 2^4, or 16 times more power than it does now. Tripling the temperature would raise the power output 81 times. Hot stars really shine away a tremendous amount of energy.

Example 5.4: Calculating the Power of a Star

While energy flux tells us how much power a star emits per square meter, we would often like to know how much total power is emitted by the star. We can determine that by multiplying the energy flux by the number of square meters on the surface of the star. Stars are mostly spherical, so we can use the formula 4πR24πR2 for the surface area, where R is the radius of the star. The total power emitted by the star (which we call the star's "absolute luminosity") can be found by multiplying the formula for energy flux and the formula for the surface area:

L=4πR2σT4L=4πR2σT4

Two stars have the same size and are the same distance from us. Star A has a surface temperature of 6000 K, and star B has a surface temperature twice as high, 12,000 K. How much more luminous is star B compared to star A?

Solution

LA=4πRA2σTA4andLB=4πRB2σTB4LA=4πRA2σTA4andLB=4πRB2σTB4

Take the ratio of the luminosity of Star A to Star B:

LBLA=4πRB2σTB44πRA2σTA4=RB2TB4RA2TA4LBLA=4πRB2σTB44πRA2σTA4=RB2TB4RA2TA4

Because the two stars are the same size, $R_A = R_B$, leaving

TB4TA4=(12,000K)4(6000K)4=24=16TB4TA4=(12,000K)4(6000K)4=24=16

Exercise 4.2.1

Two stars with identical diameters are the same distance away. One has a temperature of 8700 K and the other has a temperature of 2900 K. Which is brighter? How much brighter is it?

Answer available at the end of this section.

Link to Learning

You can use the Blackbody Spectrum simulator to explore the relationship between an object's temperature, peak wavelength, overall brightness, and apparent color.

Answers

Exercise 4.2.1

T=3×106nm Kλmax=3×106nm K290nm=10,300KT=3×106nm Kλmax=3×106nm K290nm=10,300K

Exercise 4.2.1

The 8700 K star has triple the temperature, so it is 3^4 = 81 times brighter.

This page titled 4.2: The Electromagnetic Spectrum is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by OpenStax via source content that was edited to the style and standards of the LibreTexts platform.

4.3: Spectroscopy in Astronomy

Learning Objectives

By the end of this section, you will be able to:

- Describe the properties of light
- Explain how astronomers learn the composition of a gas by examining its spectral lines
- Discuss the various types of spectra

Electromagnetic radiation carries a lot of information about the nature of stars and other astronomical objects. To extract this information, however, astronomers must be able to study the amounts of energy we receive at different wavelengths of light in fine detail. Let's examine how we can do this and what we can learn.

Properties of Light

Light exhibits certain behaviors that are important to the design of telescopes and other instruments. For example, light can be *reflected* from a surface. If the surface is smooth and shiny, as with a mirror, the direction of the reflected light beam can be calculated accurately from knowledge of the shape of the reflecting surface. Light is also bent, or *refracted*, when it passes from one kind of transparent material into another—say, from the air into a glass lens.

Reflection and refraction of light are the basic properties that make possible all *optical* instruments (devices that help us to see things better)—from eyeglasses to giant astronomical telescopes. Such instruments are generally combinations of glass lenses, which bend light according to the principles of refraction, and curved mirrors, which depend on the properties of reflection. Small optical devices, such as eyeglasses or binoculars, generally use lenses, whereas large telescopes depend almost entirely on mirrors for their main optical elements. We will discuss astronomical instruments and their uses more fully in Astronomical Instruments. For now, we turn to another behavior of light, one that is essential for the decoding of light.

In 1672, in the first paper that he submitted to the Royal Society, Sir Isaac Newton described an experiment in which he permitted sunlight to pass through a small hole and then through a prism. Newton found that sunlight, which looks white to us, is actually made up of a mixture of all the colors of the rainbow (Figure 5.9).

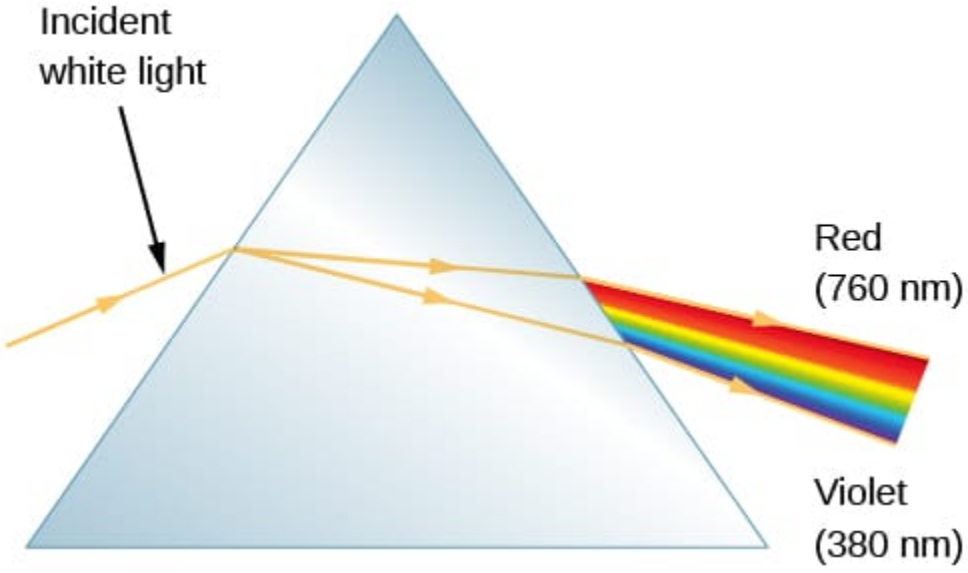


Figure 5.9 : Action of a Prism. When we pass a beam of white sunlight through a prism, we see a rainbow-colored band of light that we call a continuous spectrum.

Figure 5.9 shows how light is separated into different colors with a prism—a piece of glass in the shape of a triangle with refracting

surfaces. Upon entering one face of the prism, the path of the light is refracted (bent), but not all of the colors are bent by the same amount. The bending of the beam depends on the wavelength of the light as well as the properties of the material, and as a result, different wavelengths (or colors of light) are bent by different amounts and therefore follow slightly different paths through the prism. The violet light is bent more than the red. This phenomenon is called dispersion and explains Newton's rainbow experiment.

Upon leaving the opposite face of the prism, the light is bent again and further dispersed. If the light leaving the prism is focused on a screen, the different wavelengths or colors that make up white light are lined up side by side just like a rainbow (Figure 5.10). (In fact, a rainbow is formed by the dispersion of light though raindrops; see The Rainbow feature box.) Because this array of colors is a spectrum of light, the instrument used to disperse the light and form the spectrum is called a spectrometer.

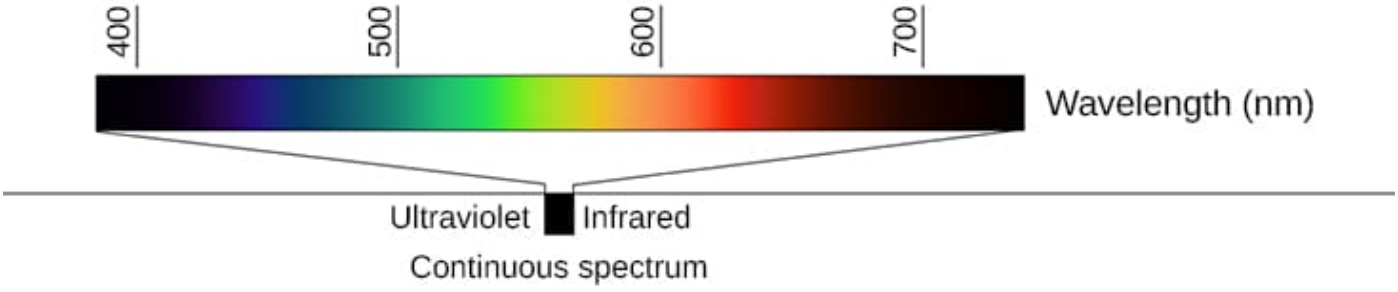


Figure 5.10 : Continuous Spectrum. When white light passes through a prism, it is dispersed and forms a continuous spectrum of all the colors. Although it is hard to see in this printed version, in a well-dispersed spectrum, many subtle gradations in color are visible as your eye scans from one end (violet) to the other (red).

The Value of Stellar Spectra

When Newton described the laws of refraction and dispersion in optics, and observed the solar spectrum, all he could see was a continuous band of colors. If the spectrum of the white light from the Sun and stars were simply a continuous rainbow of colors, astronomers would have little interest in the detailed study of a star's spectrum once they had learned its average surface temperature. In 1802, however, William Wollaston built an improved spectrometer that included a lens to focus the Sun's spectrum on a screen. With this device, Wollaston saw that the colors were not spread out uniformly, but instead, some ranges of color were missing, appearing as dark bands in the solar spectrum. He mistakenly attributed these lines to natural boundaries between the colors. In 1815, German physicist Joseph Fraunhofer, upon a more careful examination of the solar spectrum, found about 600 such dark lines (missing colors), which led scientists to rule out the boundary hypothesis (Figure 5.11).

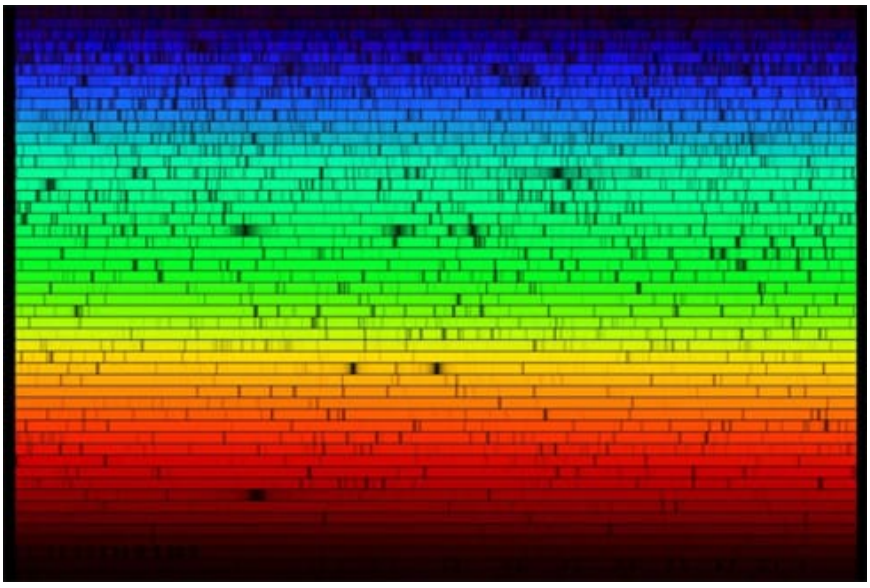


Figure 5.11 : Visible Spectrum of the Sun. Our star's spectrum is crossed by dark lines produced by atoms in the solar atmosphere that absorb light at certain wavelengths. (credit: modification of work by Nigel Sharp, NOAO/National Solar Observatory at Kitt Peak/AURA, and the National Science

Foundation)

Later, researchers found that similar dark lines could be produced in the spectra ("spectra" is the plural of "spectrum") of artificial light sources. They did this by passing their light through various apparently transparent substances—usually containers with just a bit of thin gas in them.

These gases turned out not to be transparent at *all* colors: they were quite opaque at a few sharply defined wavelengths. Something in each gas had to be absorbing just a few colors of light and no others. All gases did this, but each different element absorbed a different set of colors and thus showed different dark lines. If the gas in a container consisted of two elements, then light passing through it was missing the colors (showing dark lines) for both of the elements. So it became clear that certain lines in the spectrum "go with" certain elements. This discovery was one of the most important steps forward in the history of astronomy.

What would happen if there were no continuous spectrum for our gases to remove light from? What if, instead, we heated the same thin gases until they were hot enough to glow with their own light? When the gases were heated, a spectrometer revealed no continuous spectrum, but several separate bright lines. That is, these hot gases emitted light only at certain specific wavelengths or colors.

When the gas was pure hydrogen, it would emit one pattern of colors; when it was pure sodium, it would emit a different pattern. A mixture of hydrogen and sodium emitted both sets of spectral lines. The colors the gases emitted when they were heated were the very same colors as those they had absorbed when a continuous source of light was behind them. From such experiments, scientists began to see that different substances showed distinctive *spectral signatures* by which their presence could be detected (Figure 5.12). Just as your signature allows the bank to identify you, the unique pattern of colors for each type of atom (its spectrum) can help us identify which element or elements are in a gas.

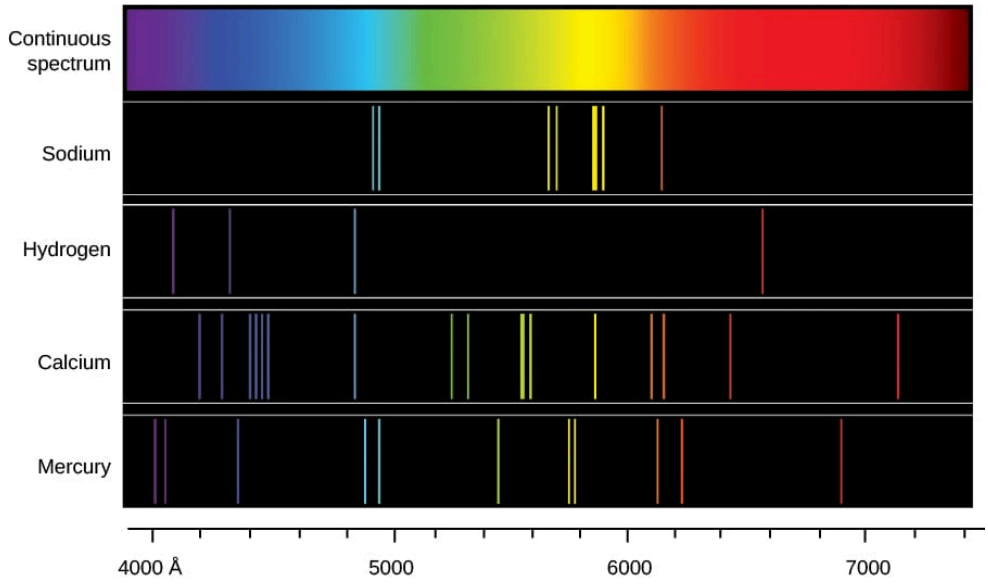


Figure 5.12 : Continuous Spectrum and Line Spectra from Different Elements. Each type of glowing gas (each element) produces its own unique pattern of lines, so the composition of a gas can be identified by its spectrum. The spectra of sodium, hydrogen, calcium, and mercury gases are shown here.

Types of Spectra

In these experiments, then, there were three different types of spectra. A continuous spectrum (formed when a solid or very dense gas gives off radiation) is an array of all wavelengths or colors of the rainbow. A continuous spectrum can serve as a backdrop from which the atoms of much less dense gas can absorb light. A dark line, or absorption spectrum, consists of a series or pattern of dark lines—missing colors—superimposed upon the continuous spectrum of a source. A bright line, or emission spectrum, appears as a pattern or series of bright lines; it consists of light in which only certain discrete wavelengths are present. (Figure 5.11 shows an absorption spectrum, whereas Figure 5.12 shows the emission spectrum of a number of common elements along with an example of a continuous spectrum.)

When we have a hot, thin gas, each particular chemical element or compound produces its own characteristic pattern of spectral lines—its spectral signature. No two types of atoms or molecules give the same patterns. In other words, each particular gas can absorb or emit only certain wavelengths of the light peculiar to that gas. In contrast, absorption spectra occur when passing white light through a cool, thin gas. The temperature and other conditions determine whether the lines are bright or dark (whether light is absorbed or emitted), but the wavelengths of the lines for any element are the same in either case. It is the precise pattern of wavelengths that makes the signature of each element unique. Liquids and solids can also generate spectral lines or bands, but they are broader and less well defined—and hence, more difficult to interpret. Spectral analysis, however, can be quite useful. It can, for example, be applied to light reflected off the surface of a nearby asteroid as well as to light from a distant galaxy.

The dark lines in the solar spectrum thus give evidence of certain chemical elements between us and the Sun absorbing those wavelengths of sunlight. Because the space between us and the Sun is pretty empty, astronomers realized that the atoms doing the absorbing must be in a thin atmosphere of cooler gas around the Sun. This outer atmosphere is not all that different from the rest of the Sun, just thinner and cooler. Thus, we can use what we learn about its composition as an indicator of what the whole Sun is made of. Similarly, we can use the presence of absorption and emission lines to analyze the composition of other stars and clouds of gas in space.

Such analysis of spectra is the key to modern astronomy. Only in this way can we “sample” the stars, which are too far away for us to visit. Encoded in the electromagnetic radiation from celestial objects is clear information about the chemical makeup of these objects. Only by understanding what the stars were made of could astronomers begin to form theories about what made them shine and how they evolved.

In 1860, German physicist Gustav Kirchhoff became the first person to use spectroscopy to identify an element in the Sun when he found the spectral signature of sodium gas. In the years that followed, astronomers found many other chemical elements in the Sun and stars. In fact, the element helium was found first in the Sun from its spectrum and only later identified on Earth. (The word “helium” comes from *helios*, the Greek name for the Sun.)

Why are there specific lines for each element? The answer to that question was not found until the twentieth century; it required the development of a model for the atom. We therefore turn next to a closer examination of the atoms that make up all matter.

Link to Learning

You can use the Spectrum Constructor to explore the three types of spectra and how they appear through a diffraction grating when the brightness of each wavelength is measured with a spectrometer.

Making Connections: The Rainbow

Rainbows are an excellent illustration of the dispersion of sunlight. You have a good chance of seeing a rainbow any time you are between the Sun and a rain shower, as illustrated in Figure 5.13. The raindrops act like little prisms and break white light into the spectrum of colors. Suppose a ray of sunlight encounters a raindrop and passes into it. The light changes direction—is refracted—when it passes from air to water; the blue and violet light are refracted more than the red. Some of the light is then reflected at the backside of the drop and reemerges from the front, where it is again refracted. As a result, the white light is spread out into a rainbow of colors.



Figure 5.13 : Rainbow Refraction. (a) This diagram shows how light from the Sun, which is located behind the observer, can be refracted by raindrops to produce (b) a rainbow. (c) Refraction separates white light into its component colors.

Note that violet light lies above the red light after it emerges from the raindrop. When you look at a rainbow, however, the red light is higher in the sky. Why? Look again at Figure 5.13. If the observer looks at a raindrop that is high in the sky, the violet light passes over her head and the red light enters her eye. Similarly, if the observer looks at a raindrop that is low in the sky, the violet light reaches her eye and the drop appears violet, whereas the red light from that same drop strikes the ground and is not seen. Colors of intermediate wavelengths are refracted to the eye by drops that are intermediate in altitude between the drops that appear violet and the ones that appear red. Thus, a single rainbow always has red on the outside and violet on the inside.

This page titled 4.3: Spectroscopy in Astronomy is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by OpenStax via source content that was edited to the style and standards of the LibreTexts platform.

4.4: The Structure of the Atom

Learning Objectives

By the end of this section, you will be able to:

- Describe the structure of atoms and the components of nuclei
- The idea that matter is composed of tiny particles called atoms is at least 25 centuries old. It took until the twentieth century, however, for scientists to invent instruments that permitted them to probe inside an atom and find that it is not, as had been thought, hard and indivisible. Instead, the atom is a complex structure composed of still smaller particles.

Probing the Atom

The first of these smaller particles was discovered by British physicist James (J. J.) Thomson in 1897. Named the *electron*, this particle is negatively charged. (It is the flow of these particles that produces currents of electricity, whether in lightning bolts or in the wires leading to your lamp.) Because an atom in its normal state is electrically neutral, each electron in an atom must be balanced by the same amount of positive charge.

The next step was to determine where in the atom the positive and negative charges are located. In 1911, British physicist Ernest Rutherford devised an experiment that provided part of the answer to this question. He bombarded an extremely thin piece of gold foil, only about 400 atoms thick, with a beam of alpha particles (Figure 5.14). *Alpha particles* (α particles) are helium atoms that have lost their electrons and thus are positively charged. Most of these particles passed though the gold foil just as if it and the atoms in it were nearly empty space. About 1 in 8000 of the alpha particles, however, completely reversed direction and bounced backward from the foil. Rutherford wrote, “It was quite the most incredible event that has ever happened to me in my life. It was almost as incredible as if you fired a 15-inch shell at a piece of tissue paper and it came back and hit you.”

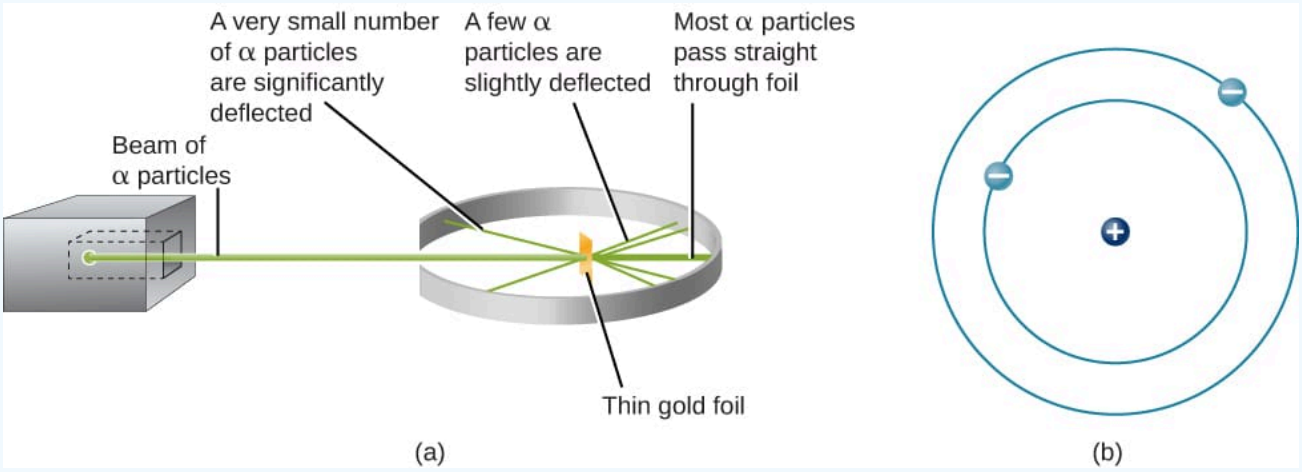


Figure 5.14 : Rutherford’s Experiment. (a) When Rutherford allowed α particles from a radioactive source to strike a target of gold foil, he found that, although most of them went straight through, some rebounded back in the direction from which they came. (b) From this experiment, he concluded that the atom must be constructed like a miniature solar system, with the positive charge concentrated in the nucleus and the negative charge orbiting in the large volume around the nucleus. Note that this drawing is not to scale; the electron orbits are much larger relative to the size of the nucleus.

The only way to account for the particles that reversed direction when they hit the gold foil was to assume that nearly all of the mass, as well as all of the positive charge in each individual gold atom, is concentrated in a tiny center or nucleus. When a positively charged alpha particle strikes a nucleus, it reverses direction, much as a cue ball reverses direction when it strikes another billiard ball. Rutherford’s model placed the other type of charge—the negative electrons—in orbit around this nucleus.

Rutherford's model required that the electrons be in motion. Positive and negative charges attract each other, so stationary electrons would fall into the positive nucleus. Also, because both the electrons and the nucleus are extremely small, most of the atom is empty, which is why nearly all of Rutherford's particles were able to pass right through the gold foil without colliding with anything. Rutherford's model was a very successful explanation of the experiments he conducted, although eventually scientists would discover that even the nucleus itself has structure.

The Atomic Nucleus

The simplest possible atom (and the most common one in the Sun and stars) is hydrogen. The nucleus of ordinary hydrogen contains a single proton. Moving around this proton is a single electron. The mass of an electron is nearly 2000 times smaller than the mass of a proton; the electron carries an amount of charge exactly equal to that of the proton but opposite in sign (Figure 5.15). Opposite charges attract each other, so it is an electromagnetic force that holds the proton and electron together, just as gravity is the force that keeps planets in orbit around the Sun.

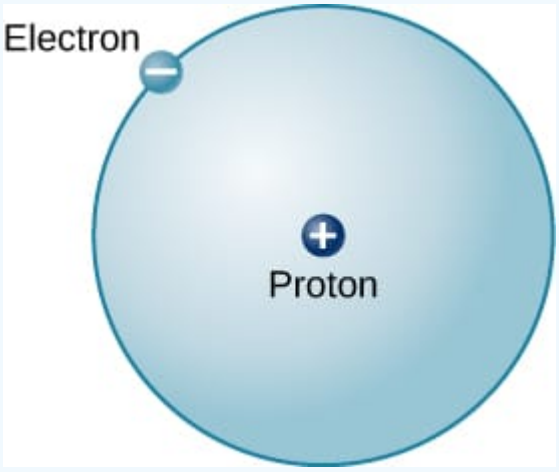


Figure 5.15 : Hydrogen Atom. This is a schematic diagram of a hydrogen atom in its lowest energy state, also called the ground state. The proton and electron have equal but opposite charges, which exert an electromagnetic force that binds the hydrogen atom together. In the illustration, the size of the particles is exaggerated so that you can see them; they are not to scale. They are also shown much closer than they would actually be as it would take more than an entire page to show their actual distance to scale.

There are many other types of atoms in nature. Helium, for example, is the second-most abundant element in the Sun. Helium has two protons in its nucleus instead of the single proton that characterizes hydrogen. In addition, the helium nucleus contains two neutrons, particles with a mass comparable to that of the proton but with no electric charge. Moving around this nucleus are two electrons, so the total net charge of the helium atom is also zero (Figure 5.16).

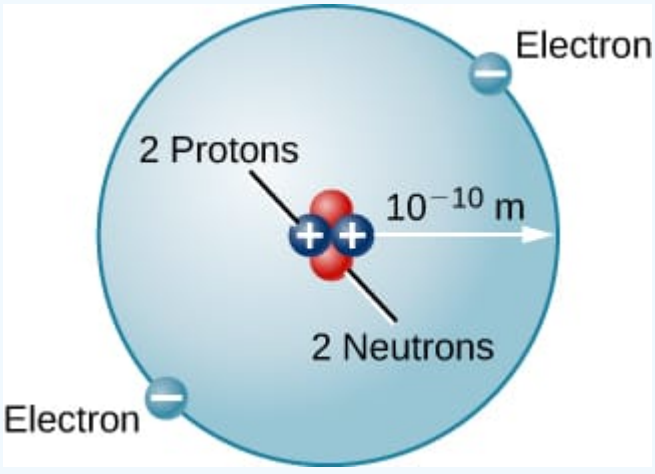


Figure 5.16 : Helium Atom. Here we see a schematic diagram of a helium atom in its lowest energy state. Two protons are present in the nucleus of all helium atoms. In the most common variety of helium, the nucleus also contains two neutrons, which have nearly the same mass as the proton but carry no charge. Two electrons orbit the nucleus.

From this description of hydrogen and helium, perhaps you have guessed the pattern for building up all the elements (different types of atoms) that we find in the universe. The type of element is determined by the number of protons in the nucleus of the atom. For example, any atom with six protons is the element carbon, with eight protons is oxygen, with 26 is iron, and with 92 is uranium. On Earth, a typical atom has the same number of electrons as protons, and these electrons follow complex orbital patterns around the nucleus. Deep inside stars, however, it is so hot that the electrons get loose from the nucleus and (as we shall see) lead separate yet productive lives.

The ratio of neutrons to protons increases as the number of protons increases, but each element is unique. The number of neutrons is not necessarily the same for all atoms of a given element. For example, most hydrogen atoms contain no neutrons at all. There are, however, hydrogen atoms that contain one proton and one neutron, and others that contain one proton and two neutrons. The various types of hydrogen nuclei with different numbers of neutrons are called isotopes of hydrogen (Figure 5.17), and all other elements have isotopes as well. You can think of isotopes as siblings in the same element "family"—closely related but with different characteristics and behaviors.

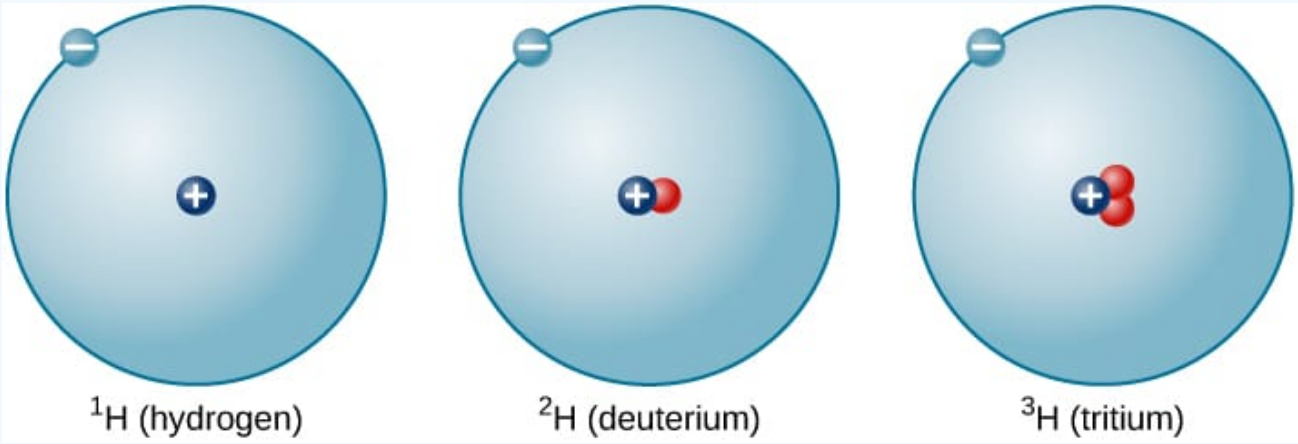


Figure 5.17 : Isotopes of Hydrogen. A single proton in the nucleus defines the atom to be hydrogen, but there may be zero, one, or two neutrons. The most common isotope of hydrogen is the one with only a single proton and no neutrons.

Link to Learning

To explore the structure of atoms, go to the PhET Build an Atom website where you can add protons, neutrons, or electrons to a model and the name of the element you have created will appear. You can also see the net charge, the mass number, whether it is stable or unstable, and whether it is an ion or a neutral atom.

The Bohr Atom

Rutherford's model for atoms has one serious problem. Maxwell's theory of electromagnetic radiation says that when electrons change either speed or the direction of motion, they must emit energy. Orbiting electrons constantly change their direction of motion, so they should emit a constant stream of energy. Applying Maxwell's theory to Rutherford's model, all electrons should spiral into the nucleus of the atom as they lose energy, and this collapse should happen very quickly—in about 10^{-16} seconds.

It was Danish physicist Niels Bohr (1885–1962) who solved the mystery of how electrons remain in orbit. He was trying to develop a model of the atom that would also explain certain regularities observed in the spectrum of hydrogen. He suggested that the spectrum of hydrogen can be understood if we assume that orbits of only certain sizes are possible for the electron. Bohr further assumed that as long as the electron moves in only one of these allowed orbits, it radiates no energy: its energy would change only if it moved from one orbit to another.

This suggestion, in the words of science historian Abraham Pais, was "one of the most audacious hypotheses ever introduced in physics." If something equivalent were at work in the everyday world, you might find that, as you went for a walk after astronomy class, nature permitted you to walk two steps per minute, five steps per minute, and 12 steps per minute, but no speeds in between. No matter how you tried to move your legs, only certain walking speeds would be permitted. To make things more bizarre, it would take no effort to walk at any one of the allowed speeds, but it would

be difficult to change from one speed to another. Luckily, no such rules apply at the level of human behavior. But at the microscopic level of the atom, experiment after experiment has confirmed the validity of Bohr's strange idea. Bohr's suggestions became one of the foundations of the new (and much more sophisticated) model of the subatomic world called quantum mechanics.

In Bohr's model, if the electron moves from one orbit to another closer to the atomic nucleus, it must give up some energy in the form of electromagnetic radiation. If the electron goes from an inner orbit to one farther from the nucleus, however, it requires some additional energy. One way to obtain the necessary energy is to absorb electromagnetic radiation that may be streaming past the atom from an outside source.

A key feature of Bohr's model is that each of the permitted electron orbits around a given atom has a certain energy value; we therefore can think of each orbit as an energy level. To move from one orbit to another (which will have its own specific energy value) requires a change in the electron's energy—a change determined by the difference between the two energy values. If the electron goes to a lower level, the energy difference will be given off; if the electron goes to a higher level, the energy difference must be obtained from somewhere else. Each jump (or transition) to a different level has a fixed and definite energy change associated with it.

A crude analogy for this situation might be life in a tower of luxury apartments where the rent is determined by the quality of the view. Such a building has certain, definite numbered levels or floors on which apartments are located. No one can live on floor 5.37 or 22.5. In addition, the rent gets higher as you go up to higher floors. If you want to exchange an apartment on the twentieth floor for one on the second floor, you will not owe as much rent. However, if you want to move from the third floor to the twenty-fifth floor, your rent will increase. In an atom, too, the "cheapest" place for an electron to live is the lowest possible level, and energy is required to move to a higher level.

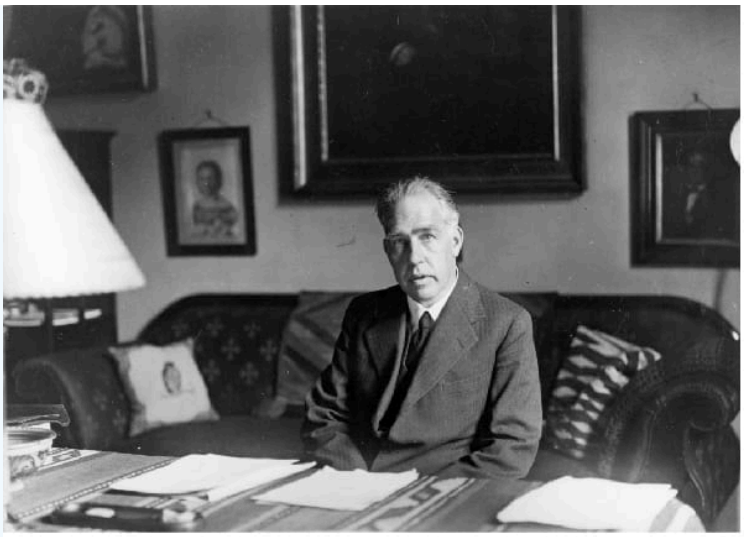
Here we have one of the situations where it is easier to think of electromagnetic radiation as particles (photons) rather than as waves. As electrons move from one level to another, they give off or absorb little packets of energy. When an electron moves to a higher level, it absorbs a photon of just the right energy (provided one is available). When it moves to a lower level, it emits a photon with the exact amount of energy it no longer needs in its "lower-cost living situation."

The photon and wave perspectives must be equivalent: light is light, no matter how we look at it. Thus, each photon carries a certain amount of energy that is proportional to the frequency (f) of the wave it represents. The value of its energy (E) is given by the formula

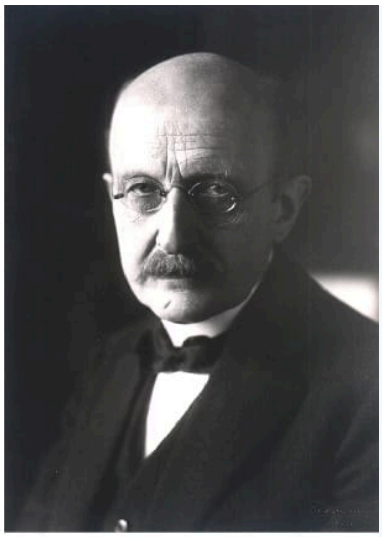
E=hfE=hf

where the constant of proportionality, h, is called Planck's constant.

The constant is named for Max Planck, the German physicist who was one of the originators of the quantum theory (Figure 5.18). If metric units are used (that is, if energy is measured in joules and frequency in hertz), then Planck's constant has the value $h = 6.626 \times 10^{-34}$ joule-seconds (J-s). Higher-energy photons correspond to higher-frequency waves (which have a shorter wavelength); lower-energy photons are waves of lower frequency.



(a)



(b)

Figure 5.18 : Niels Bohr (1885–1962) and Max Planck (1858–1947). (a) Bohr, shown at his desk in this 1935 photograph, and (b) Planck helped us understand the energy behavior of photons.

To take a specific example, consider a calcium atom inside the Sun's atmosphere in which an electron jumps from a lower level to a higher level. To do this, it needs about 5×10-195×10-19 joules of energy, which it can conveniently obtain by absorbing a passing photon of that energy coming from deeper inside the Sun. This photon is equivalent to a wave of light whose frequency is about 7.5×10147.5×1014 hertz and whose wavelength is about 3.9×10-73.9×10-7 meters (393 nanometers), in the deep violet part of the visible light spectrum. Although it may seem strange at first to switch from picturing light as a photon (or energy packet) to picturing it as a wave, such switching has become second nature to astronomers and can be a handy tool for doing calculations about spectra.

Example 5.5: The Energy of a Photon

Now that we know how to calculate the wavelength and frequency of a photon, we can use this information, along with Planck's constant, to determine how much energy each photon carries. How much energy does a red photon of wavelength 630 nm have?

Solution

First, as we learned earlier, we can find the frequency of the photon:

f=cλ=3×108m/s630×10−9m=4.8×1014Hzf=cλ=3×108m/s630×10−9m=4.8×1014Hz

Next, we can use Planck's constant to determine the energy (remember that a Hz is the same as 1/s):

E=hf=(6.626×10−34J-s)(4.8×1014Hz(1/s))=3.2×10−19JE=hf=(6.626×10−34J-s)(4.8×1014Hz(1/s))=3.2×10−19J

Exercise 4.4.1

What is the energy of a yellow photon with a frequency of 5.5×10145.5×1014 Hz?

Answer available at the end of this section.

Answers

Exercise 4.4.1

E=hf=(6.626×10−34J-s)(5.5×1014Hz)=3.6×10−19JE=hf=(6.626×10−34J-s)(5.5×1014Hz)=3.6×10−19J

This page titled 4.4: The Structure of the Atom is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by OpenStax via source content that was edited to the style and standards of the LibreTexts platform.

4.5: Formation of Spectral Lines

Learning Objectives

By the end of this section, you will be able to:

- Explain how emission line spectra and absorption line spectra are formed
- Describe what ions are and how they are formed
- Explain how spectral lines and ionization levels in a gas can help us determine its temperature

We can use Bohr's model of the atom to understand how spectral lines are formed. The concept of energy levels for the electron orbits in an atom leads naturally to an explanation of why atoms absorb or emit only specific energies or wavelengths of light.

The Hydrogen Spectrum

Let's look at the hydrogen atom from the perspective of the Bohr model. Suppose a beam of white light (which consists of photons of all visible wavelengths) shines through a gas of atomic hydrogen. A photon of wavelength 656 nanometers has just the right energy to raise an electron in a hydrogen atom from the second to the third orbit. Thus, as all the photons of different energies (or wavelengths or colors) stream by the hydrogen atoms, photons with *this* particular wavelength can be absorbed by those atoms whose electrons are orbiting on the second level. When they are absorbed, the electrons on the second level will move to the third level, and a number of the photons of this wavelength and energy will be missing from the general stream of white light.

Other photons will have the right energies to raise electrons from the second to the fourth orbit, or from the first to the fifth orbit, and so on. Only photons with these exact energies can be absorbed. All of the other photons will stream past the atoms untouched. Thus, hydrogen atoms absorb light at only certain wavelengths and produce dark lines at those wavelengths in the spectrum we see.

Suppose we have a container of hydrogen gas through which a whole series of photons is passing, allowing many electrons to move up to higher levels. When we turn off the light source, these electrons "fall" back down from larger to smaller orbits and emit photons of light—but, again, only light of those energies or wavelengths that correspond to the energy difference between permissible orbits. The orbital changes of hydrogen electrons that give rise to some spectral lines are shown in Figure 5.19.

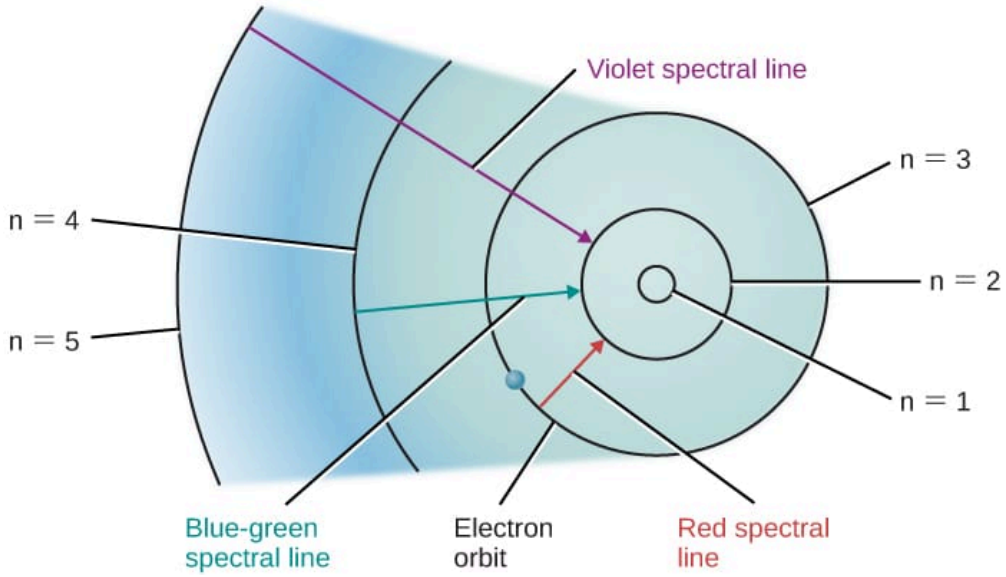


Figure 5.19 : Bohr Model for Hydrogen. In this simplified model of a hydrogen atom, the concentric circles shown represent permitted orbits or energy levels. An electron in a hydrogen atom can only exist in one of these energy levels (or states). The closer the electron is to the nucleus, the more tightly bound the electron is to the nucleus. By absorbing energy, the electron can move to energy levels farther from

the nucleus (and even escape if enough energy is absorbed).

Similar pictures can be drawn for atoms other than hydrogen. However, because these other atoms ordinarily have more than one electron each, the orbits of their electrons are much more complicated, and the spectra are more complex as well. For our purposes, the key conclusion is this: *each type of atom has its own unique pattern of electron orbits, and no two sets of orbits are exactly alike*. This means that each type of atom shows its own unique set of spectral lines, produced by electrons moving between its unique set of orbits.

Astronomers and physicists have worked hard to learn the lines that go with each element by studying the way atoms absorb and emit light in laboratories here on Earth. Then they can use this knowledge to identify the elements in celestial bodies. In this way, we now know the chemical makeup of not just any star, but even galaxies of stars so distant that their light started on its way to us long before Earth had even formed.

Energy Levels and Excitation

Bohr's model of the hydrogen atom was a great step forward in our understanding of the atom. However, we know today that atoms cannot be represented by quite so simple a picture. For example, the concept of sharply defined electron orbits is not really correct; however, at the level of this introductory course, the notion that only certain discrete energies are allowable for an atom is very useful. The energy levels we have been discussing can be thought of as representing certain average distances of the electron's possible orbits from the atomic nucleus.

Ordinarily, an atom is in the state of lowest possible energy, its ground state. In the Bohr model of the hydrogen atom, the ground state corresponds to the electron being in the innermost orbit. An atom can absorb energy, which raises it to a higher energy level (corresponding, in the simple Bohr picture, to an electron's movement to a larger orbit)—this is referred to as excitation. The atom is then said to be in an *excited state*. Generally, an atom remains excited for only a very brief time. After a short interval, typically a hundred-millionth of a second or so, it drops back spontaneously to its ground state, with the simultaneous emission of light. The atom may return to its lowest state in one jump, or it may make the transition in steps of two or more jumps, stopping at intermediate levels on the way down. With each jump, it emits a photon of the wavelength that corresponds to the energy difference between the levels at the beginning and end of that jump.

An energy-level diagram for a hydrogen atom and several possible atomic transitions are shown in Figure 5.20. When we measure the energies involved as the atom jumps between levels, we find that the transitions to or from the ground state, called the *Lyman series* of lines, result in the emission or absorption of ultraviolet photons. But the transitions to or from the first excited state (labeled n = 2 in part (a) of Figure 5.20), called the Balmer series, produce emission or absorption in visible light. In fact, it was to explain this Balmer series that Bohr first suggested his model of the atom.

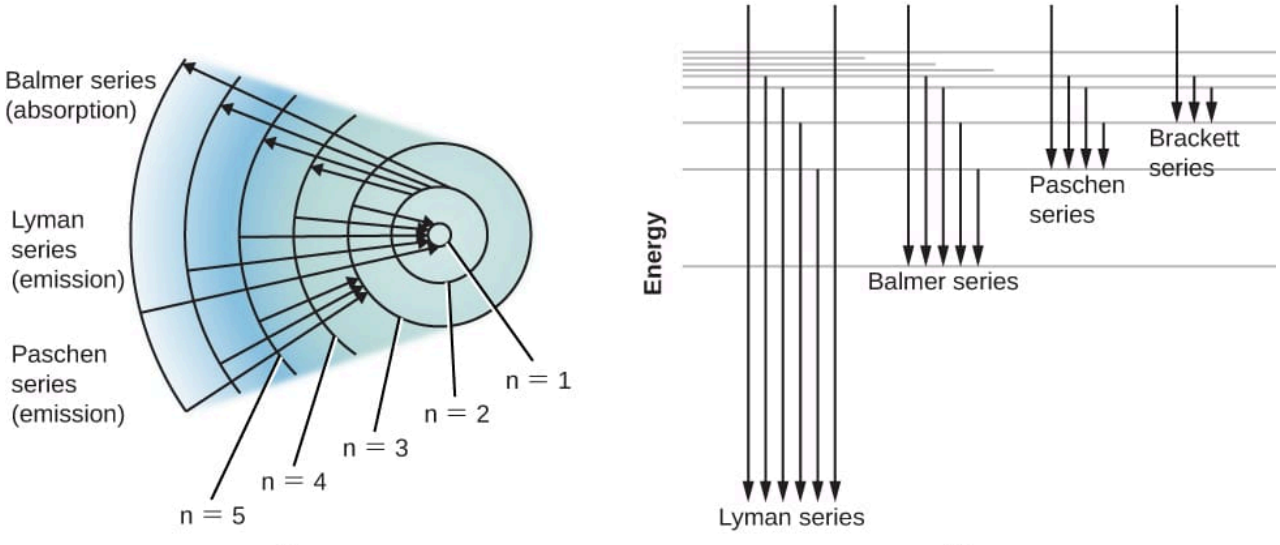


Figure 5.20 : Energy-Level Diagrams for Hydrogen. (a) Here we follow the emission or absorption of photons by a hydrogen atom according to the Bohr model. Several different series of spectral lines are shown, corresponding to transitions of electrons from or to certain allowed orbits. Each series of lines that terminates on a specific inner orbit is named for the physicist who studied

it. At the top, for example, you see the Balmer series, and arrows show electrons jumping from the second orbit (n = 2) to the third, fourth, fifth, and sixth orbits. Each time a "poor" electron from a lower level wants to rise to a higher position in life, it must absorb energy to do so. It can absorb the energy it needs from passing waves (or photons) of light. The next set of arrows (Lyman series) show electrons falling down to the first orbit from different (higher) levels. Each time a "rich" electron goes downward toward the nucleus, it can afford to give off (emit) some energy it no longer needs. (In this idealized diagram, the energy levels are shown equally spaced; in real life, they are not.) (b) At higher and higher energy levels, the levels become more and more crowded together, approaching a limit. The region above the top line represents energies at which the atom is ionized (the electron is no longer attached to the atom). Each series of arrows represents electrons falling from higher levels to lower ones, releasing photons or waves of energy in the process.

Atoms that have absorbed specific photons from a passing beam of white light and have thus become excited generally de-excite themselves and emit that light again in a very short time. You might wonder, then, why *dark* spectral lines are ever produced. In other words, why doesn't this reemitted light quickly "fill in" the darker absorption lines?

Imagine a beam of white light coming toward you through some cooler gas. Some of the reemitted light *is* actually returned to the beam of white light you see, but this fills in the absorption lines only to a slight extent. The reason is that the atoms in the gas reemit light *in all directions*, and only a small fraction of the reemitted light is in the direction of the original beam (toward you). In a star, much of the reemitted light actually goes in directions leading back into the star, which does observers outside the star no good whatsoever.

Figure 5.21 summarizes the different kinds of spectra we have discussed. An incandescent lightbulb produces a continuous spectrum. When that continuous spectrum is viewed through a thinner cloud of gas, an absorption line spectrum can be seen superimposed on the continuous spectrum. If we look only at a cloud of excited gas atoms (with no continuous source seen behind it), we see that the excited atoms give off an emission line spectrum.

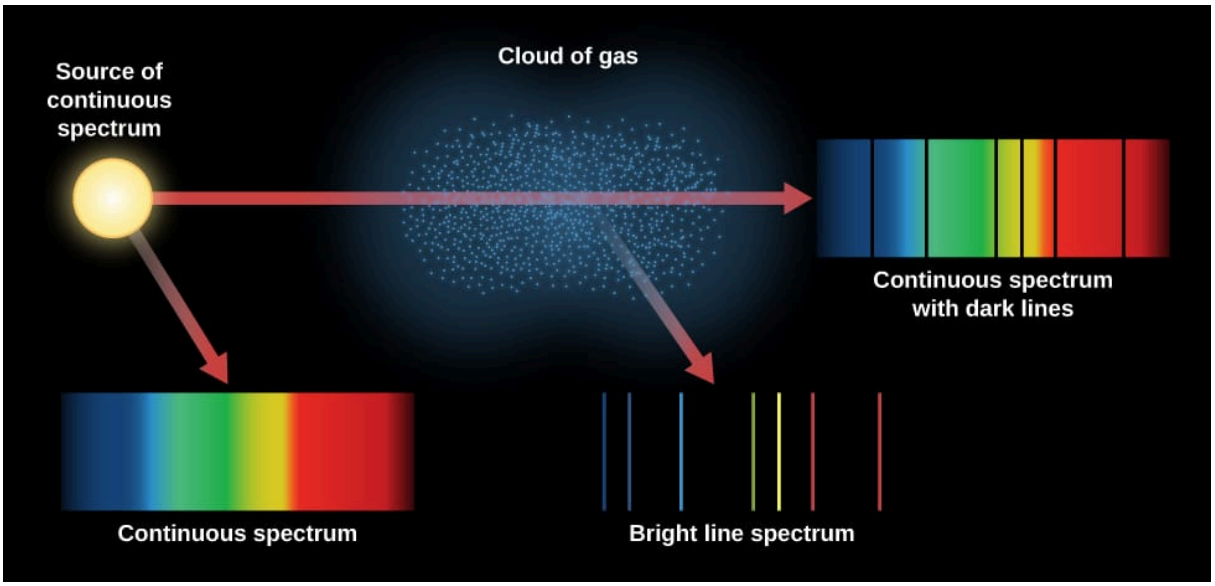


Figure 5.21 : Three Kinds of Spectra. When we see a lightbulb or other source of continuous radiation, all the colors are present. When the continuous spectrum is seen through a thinner gas cloud, the cloud's atoms produce absorption lines in the continuous spectrum. When the excited cloud is seen without the continuous source behind it, its atoms produce emission lines. We can learn which types of atoms are in the gas cloud from the pattern of absorption or emission lines.

Atoms in a hot gas are moving at high speeds and continually colliding with one another and with any loose electrons. They can be excited (electrons moving to a higher level) and de-excited (electrons moving to a lower level) by these collisions as well as by absorbing and emitting light. The speed of atoms in a gas depends on the temperature. When the temperature is higher, so are the speed and energy of the collisions. The hotter the gas, therefore, the more likely that electrons will occupy the outermost orbits, which correspond to the highest energy levels. This means that the level where electrons *start* their upward jumps in a gas can serve as an indicator of how hot that gas is. In this way, the absorption lines in a spectrum give astronomers information about the temperature of the regions where the lines originate.

Link to Learning

Use this simulation to play with a hydrogen atom and see what happens when electrons move to higher levels and then give off photons as they go to a lower level.

Ionization

We have described how certain discrete amounts of energy can be absorbed by an atom, raising it to an excited state and moving one of its electrons farther from its nucleus. If enough energy is absorbed, the electron can be completely removed from the atom—this is called ionization. The atom is then said to be ionized. The minimum amount of energy required to remove one electron from an atom in its ground state is called its ionization energy.

Still-greater amounts of energy must be absorbed by the now-ionized atom (called an ion) to remove an additional electron deeper in the structure of the atom. Successively greater energies are needed to remove the third, fourth, fifth—and so on—electrons from the atom. If enough energy is available, an atom can become completely ionized, losing all of its electrons. A hydrogen atom, having only one electron to lose, can be ionized only once; a helium atom can be ionized twice; and an oxygen atom up to eight times. When we examine regions of the cosmos where there is a great deal of energetic radiation, such as the neighborhoods where hot young stars have recently formed, we see a lot of ionization going on.

An atom that has become positively ionized has lost a negative charge—the missing electron—and thus is left with a net positive charge. It therefore exerts a strong attraction on any free electron. Eventually, one or more electrons will be captured and the atom will become neutral (or ionized to one less degree) again. During the electron-capture process, the atom emits one or more photons. Which photons are emitted depends on whether the electron is captured at once to the lowest energy level of the atom or stops at one or more intermediate levels on its way to the lowest available level.

Just as the excitation of an atom can result from a collision with another atom, ion, or electron (collisions with electrons are usually most important), so can ionization. The rate at which such collisional ionizations occur depends on the speeds of the atoms and hence on the temperature of the gas—the hotter the gas, the more of its atoms will be ionized.

The rate at which ions and electrons recombine also depends on their relative speeds—that is, on the temperature. In addition, it depends on the density of the gas: the higher the density, the greater the chance for recapture, because the different kinds of particles are crowded more closely together. From a knowledge of the temperature and density of a gas, it is possible to calculate the fraction of atoms that have been ionized once, twice, and so on. In the Sun, for example, we find that most of the hydrogen and helium atoms in its atmosphere are neutral, whereas most of the calcium atoms, as well as many other heavier atoms, are ionized once.

The energy levels of an ionized atom are entirely different from those of the same atom when it is neutral. Each time an electron is removed from the atom, the energy levels of the ion, and thus the wavelengths of the spectral lines it can produce, change. This helps astronomers differentiate the ions of a given element. Ionized hydrogen, having no electron, can produce no absorption lines.

This page titled 4.5: Formation of Spectral Lines is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by OpenStax via source content that was edited to the style and standards of the LibreTexts platform.

4.6: The Doppler Effect

Learning Objectives

By the end of this section, you will be able to:

- Explain why the spectral lines of photons we observe from an object will change as a result of the object's motion toward or away from us
- Describe how we can use the Doppler effect to deduce how fast astronomical objects are moving through space

The last two sections introduced you to many new concepts, and we hope that through those, you have seen one major idea emerge. Astronomers can learn about the elements in stars and galaxies by decoding the information in their spectral lines. There is a complicating factor in learning how to decode the message of starlight, however. If a star is moving toward or away from us, its lines will be in a slightly different place in the spectrum from where they would be in a star at rest. And most objects in the universe do have some motion relative to the Sun.

Motion Affects Waves

In 1842, Christian Doppler first measured the effect of motion on waves by hiring a group of musicians to play on an open railroad car as it was moving along the track. He then applied what he learned to all waves, including light, and pointed out that if a light source is approaching or receding from the observer, the light waves will be, respectively, crowded more closely together or spread out. The general principle, now known as the Doppler effect, is illustrated in Figure 5.22.

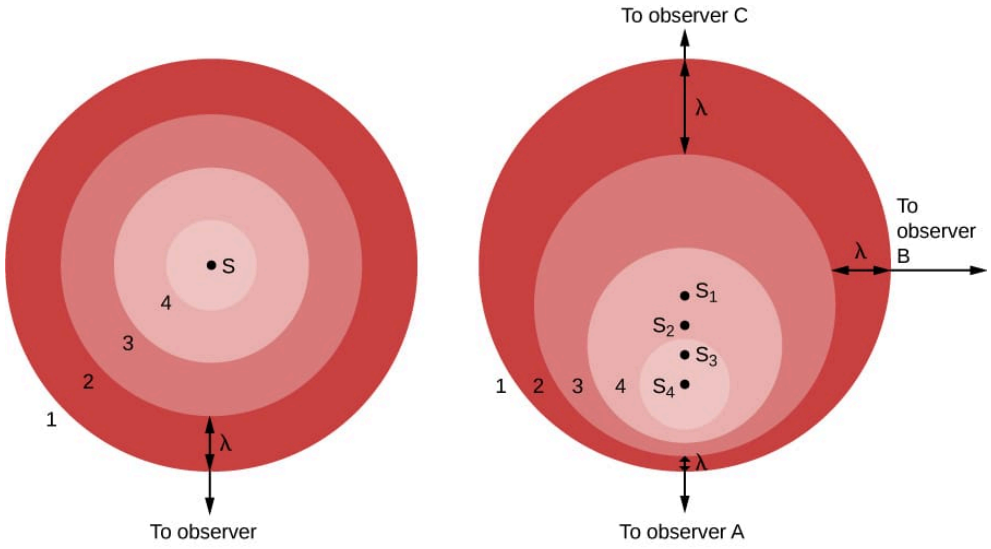


Figure 5.22 : Doppler Effect. (a) A source, S, makes waves whose numbered crests (1, 2, 3, and 4) wash over a stationary observer. (b) The source S now moves toward observer *A* and away from observer *C*. Wave crest 1 was emitted when the source was at position S1, crest 2 at position S2, and so forth. Observer *A* sees waves compressed by this motion and sees a blueshift (if the waves are light). Observer *C* sees the waves stretched out by the motion and sees a redshift. Observer *B*, whose line of sight is perpendicular to the source's motion, sees no change in the waves (and feels left out).

In part (a) of the figure, the light source (S) is at rest with respect to the observer. The source gives off a series of waves, whose crests we have labeled 1, 2, 3, and 4. The light waves spread out evenly in all directions, like the ripples from a splash in a pond. The crests are separated by a distance, λ, where λ is the wavelength. The observer, who happens to be located in the direction of the bottom of the image, sees the light waves coming nice and evenly, one wavelength apart. Observers located anywhere else would see the same thing.

On the other hand, if the source of light is moving with respect to the observer, as seen in part (b), the situation is more complicated. Between the time one crest is emitted and the next one is ready to come out, the source has moved a bit, toward the bottom of the page. From the point of view of observer *A*, this motion of the source has decreased the distance between crests—it's squeezing the crests together, this observer might say.

In part (b), we show the situation from the perspective of three observers. The source is seen in four positions, S_1, S_2, S_3, and S_4, each corresponding to the emission of one wave crest. To observer *A*, the waves seem to follow one another more closely, at a decreased wavelength and thus increased frequency. (Remember, all light waves travel at the speed of light through empty space, no matter what. This means that motion cannot affect the speed, but only the wavelength and the frequency. As the wavelength decreases, the frequency must increase. If the waves are shorter, more will be able to move by during each second.)

The situation is not the same for other observers. Let's look at the situation from the point of view of observer *C*, located opposite observer *A* in the figure. For her, the source is moving away from her location. As a result, the waves are not squeezed together but instead are spread out by the motion of the source. The crests arrive with an increased wavelength and decreased frequency. To observer *B*, in a direction at right angles to the motion of the source, no effect is observed. The wavelength and frequency remain the same as they were in part (a) of the figure.

We can see from this illustration that the Doppler effect is produced only by a motion toward or away from the observer, a motion called radial velocity. Sideways motion does not produce such an effect. Observers between *A* and *B* would observe some shortening of the light waves for that part of the motion of the source that is along their line of sight. Observers between *B* and *C* would observe lengthening of the light waves that are along their line of sight.

You may have heard the Doppler effect with sound waves. When a train whistle or police siren approaches you and then moves away, you will notice a decrease in the pitch (which is how human senses interpret sound wave frequency) of the sound waves. Compared to the waves at rest, they have changed from slightly more frequent when coming toward you, to slightly less frequent when moving away from you.

Link to Learning

A nice example of this change in the sound of a train whistle can be heard at the end of the classic Beach Boys song "Caroline, No" on their album *Pet Sounds*. To hear this sound, go to this YouTube version of the song. The sound of the train begins at approximately 2:20.

Color Shifts

When the source of waves moves toward you, the wavelength decreases a bit. If the waves involved are visible light, then the colors of the light change slightly. As wavelength decreases, they shift toward the blue end of the spectrum: astronomers call this a *blueshift* (since the end of the spectrum is really violet, the term should probably be *violetshift*, but blue is a more common color). When the source moves away from you and the wavelength gets longer, we call the change in colors a *redshift*. Because the Doppler effect was first used with visible light in astronomy, the terms "blueshift" and "redshift" became well established. Today, astronomers use these words to describe changes in the wavelengths of radio waves or X-rays as comfortably as they use them to describe changes in visible light.

The greater the motion toward or away from us, the greater the Doppler shift. If the relative motion is entirely along the line of sight, the formula for the Doppler shift of light is

$$\Delta\lambda\lambda=vc\Delta\lambda\lambda=vc$$

where λ is the wavelength emitted by the source, $\Delta\lambda$ is the difference between λ and the wavelength measured by the observer, *c* is the speed of light, and *v* is the relative speed of the observer and the source in the line of sight. The variable *v* is counted as positive if the velocity is one of recession, and negative if it is one of approach. Solving this equation for the velocity, we find $v = c \times \Delta\lambda/\lambda$.

Link to Learning

Click-and-drag the object emitting waves or the object receiving waves in this simulator to experiment with the Doppler effect yourself. The plots along the top show how the detected wavelengths change if the objects are approaching each other or moving further apart.

If a star approaches or recedes from us, the wavelengths of light in its continuous spectrum appear shortened or lengthened, respectively, as do those of the dark lines. However, unless its speed is tens of thousands of kilometers per second, the star does not appear noticeably bluer or redder than normal. The Doppler shift is thus not easily detected in a continuous spectrum and cannot be measured accurately in such a spectrum. The wavelengths of the absorption lines can be measured accurately, however, and their Doppler shift is relatively simple to detect.

Example 5.6: The Doppler Effect

We can use the Doppler effect equation to calculate the radial velocity of an object if we know three things: the speed of light, the original (unshifted) wavelength of the light emitted, and the difference between the wavelength of the emitted light and the wavelength we observe. For particular absorption or emission lines, we usually know exactly what wavelength the line has in our laboratories on Earth, where the source of light is not moving. We can measure the new wavelength with our instruments at the telescope, and so we know the difference in wavelength due to Doppler shifting. Since the speed of light is a universal constant, we can then calculate the radial velocity of the star.

A particular emission line of hydrogen is originally emitted with a wavelength of 656.3 nm from a gas cloud. At our telescope, we observe the wavelength of the emission line to be 656.6 nm. How fast is this gas cloud moving toward or away from Earth?

Solution

Because the light is shifted to a longer wavelength (redshifted), we know this gas cloud is moving away from us. The speed can be calculated using the Doppler shift formula:

$$v = c \times \frac{\Delta\lambda}{\lambda} = (3.0 \times 10^8 \text{ m/s})\left(\frac{0.3 \text{ nm}}{656.3 \text{ nm}}\right) = (3.0 \times 10^8 \text{ m/s})\left(\frac{0.3 \times 10^{-9} \text{ m}}{656.3 \times 10^{-9} \text{ m}}\right) = 140{,}000 \text{ m/s} = 140 \text{ km/s}$$

Exercise 4.6.1

Suppose a spectral line of hydrogen, normally at 500 nm, is observed in the spectrum of a star to be at 500.1 nm. How fast is the star moving toward or away from Earth?

Answer available at the end of this section.

You may now be asking: if all the stars are moving and motion changes the wavelength of each spectral line, won't this be a disaster for astronomers trying to figure out what elements are present in the stars? After all, it is the precise wavelength (or color) that tells astronomers which lines belong to which element. And we first measure these wavelengths in containers of gas in our laboratories, which are not moving. If every line in a star's spectrum is now shifted by its motion to a different wavelength (color), how can we be sure which lines and which elements we are looking at in a star whose speed we do not know?

Take heart. This situation sounds worse than it really is. Astronomers rarely judge the presence of an element in an astronomical object by a single line. It is the *pattern* of lines unique to hydrogen or calcium that enables us to determine that those elements are part of the star or galaxy we are observing. The Doppler effect does not change the pattern of lines from a given element—it only shifts the whole pattern slightly toward redder or bluer wavelengths. The shifted pattern is still quite easy to recognize. Best of all, when we do recognize a familiar element's pattern, we get a bonus: the amount the pattern is shifted can enable us to determine the speed of the objects in our line of sight.

The training of astronomers includes much work on learning to decode light (and other electromagnetic radiation). A skillful "decoder" can learn the temperature of a star, what elements are in it, and even its speed in a direction toward us or away from us. That's really an impressive amount of information for stars that are light-years away.

Answers

Exercise 4.6.1

Because the light is shifted to a longer wavelength, the star is moving away from us:

$v = c \times \frac{\Delta\lambda}{\lambda} = (3.0 \times 10^8 \text{m/s})(0.1\text{nm}500\text{nm}) = (3.0 \times 10^8 \text{m/s})(0.1 \times 10^{-9}\text{m}500 \times 10^{-9}\text{m}) = 60{,}000\text{m/s}$.$v = c \times \frac{\Delta\lambda}{\lambda} = (3.0 \times 10^8 \text{m/s})(0.1\text{nm}500\text{nm}) = (3.0 \times 10^8 \text{m/s})(0.1 \times 10^{-9}\text{m}500 \times 10^{-9}\text{m}) = 60{,}000\text{m/s}$. Its speed is
60,000 m/s.

This page titled 4.6: The Doppler Effect is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by OpenStax via source content that was edited to the style and standards of the LibreTexts platform.

5: STARS

5.1: The Structure and Composition of the Sun

5.2: Understanding Stars

5.2.1: The Brightness of Stars

5.2.2: Colors of Stars

5.2.3: The Spectra of Stars (and Brown Dwarfs)

5.2.4: Using Spectra to Measure Stellar Radius, Composition, and Motion

5.2.5: The H-R Diagram

5.3: Stellar Evolution

5.3.1: Star Formation

5.3.2: The H-R Diagram and the Study of Stellar Evolution

5.3.3: Evolution from the Main Sequence to Red Giants

5.3.4: The Death of Low-Mass Stars

5.3.5: Evolution of Massive Stars-An Explosive Finish

Chapter 5: Stars — Introduction

The Sun is the star we know best. Because it is close enough for astronomers to study in detail, it provides an important starting point for understanding the billions of other stars throughout our galaxy. Although stars appear as tiny points of light in the night sky, observations of their light allow astronomers to determine surprisingly detailed information about their physical properties.

In this chapter, we will begin with the structure and composition of the Sun before turning our attention to other stars. We will examine how astronomers use brightness, color, and spectra to determine properties such as stellar temperature, composition, size, and motion. These characteristics also allow us to organize stars using the Hertzsprung-Russell (H-R) diagram, one of the most useful tools for understanding stellar populations and evolution.

Stars are not permanent objects. They form within clouds of gas and dust and change throughout their lives as conditions within their interiors evolve. A star's mass plays a particularly important role in determining the path it follows. Some stars eventually shed their outer layers and leave behind dense stellar remnants, while more massive stars can end their lives in tremendous explosions.

By following stars from their formation through their final stages, we can see how observations of individual stars fit into a much larger story of stellar evolution. That story will also help us understand how stars contribute to the changing composition of galaxies and the universe.

This page titled 5: Stars is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Robert Wagner.

5.1: The Structure and Composition of the Sun

Learning Objectives

By the end of this section, you will be able to:

- Explain how the composition of the Sun differs from that of Earth
- Describe the various layers of the Sun and their functions
- Explain what happens in the different parts of the Sun's atmosphere

The Sun, like all stars, is an enormous ball of extremely hot, largely ionized gas, shining under its own power. And we do mean enormous. The Sun could fit 109 Earths side-by-side across its diameter, and it has enough volume (takes up enough space) to hold about 1.3 million Earths.

The Sun does not have a solid surface or continents like Earth, nor does it have a solid core (Figure 15.2). However, it does have a lot of structure and can be discussed as a series of layers, not unlike an onion. In this section, we describe the huge changes that occur in the Sun's extensive interior and atmosphere, and the dynamic and violent eruptions that occur daily in its outer layers.

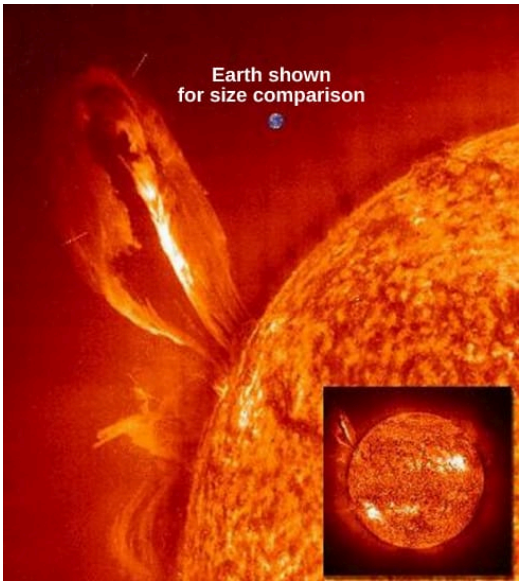


Figure 15.2 : Earth and the Sun. Here, Earth is shown to scale with part of the Sun and a giant loop of hot gas erupting from its surface. The inset shows the entire Sun, smaller. (credit: modification of work by SOHO/EIT/ESA)

Some of the basic characteristics of the Sun are listed in Table 15.1. Although some of the terms in that table may be unfamiliar to you right now, you will get to know them as you read further.

Table 15.1: Characteristics of the Sun

Characteristic	How Found	Value
Mean distance	Radar reflection from planets	1 AU (149,597,892 km)
Maximum distance from Earth		1.521×10^8 km
Minimum distance from Earth		1.471×10^8 km
Mass	Orbit of Earth	333,400 Earth masses (1.99×10^{30} kg)

Characteristic	How Found	Value
Mean angular diameter	Direct measure	31′59′′.3
Diameter of photosphere	Angular size and distance	109.3 × Earth diameter (1.39×10^6 km)
Mean density	Mass/volume	1.41 g/cm^3 (1400 kg/m^3)
Gravitational acceleration at photosphere (surface gravity)	GM/R^2	27.9 × Earth surface gravity = 273 m/s^2
Solar constant	Instrument sensitive to radiation at all wavelengths	1370 W/m^2
Luminosity	Solar constant × area of spherical surface 1 AU in radius	3.8×10^{26} W
Spectral class	Spectrum	G2V
Effective temperature	Derived from luminosity and radius of the Sun	5800 K
Rotation period at equator	Sunspots and Doppler shift in spectra taken at the edge of the Sun	24 days 16 hours
Inclination of equator to ecliptic	Motions of sunspots	7°10′.5

Composition of the Sun's Atmosphere

Let's begin by asking what the solar atmosphere is made of. As explained in Radiation and Spectra, we can use a star's *absorption line spectrum* to determine what elements are present. It turns out that the Sun contains the same elements as Earth but *not* in the same proportions. About 73% of the Sun's mass is hydrogen, and another 25% is helium. All the other chemical elements (including those we know and love in our own bodies, such as carbon, oxygen, and nitrogen) make up only 2% of our star. The 10 most abundant gases in the Sun's visible surface layer are listed in Table 15.2. Examine that table and notice that the composition of the Sun's outer layer is very different from Earth's crust, where we live. (In our planet's crust, the three most abundant elements are oxygen, silicon, and aluminum.) Although not like our planet's, the makeup of the Sun is quite typical of stars in general.

Table 15.2: The Abundance of Elements in the Sun

Element	Percentage by Number of Atoms	Percentage By Mass
Hydrogen	92.0	73.4
Helium	7.8	25.0
Carbon	0.02	0.20
Nitrogen	0.008	0.09
Oxygen	0.06	0.80
Neon	0.01	0.16
Magnesium	0.003	0.06
Silicon	0.004	0.09
Sulfur	0.002	0.05
Iron	0.003	0.14

The fact that our Sun and the stars all have similar compositions and are made up of mostly hydrogen and helium was first shown in a brilliant thesis in 1925 by Cecilia Payne-Gaposchkin, the first woman to get a PhD in astronomy in the United States (Figure 15.3). However, the idea that the simplest light gases—hydrogen and helium—were the most abundant elements in stars was so unexpected and so shocking that she was persuaded her analysis of the data must be wrong. At the time, she wrote, "The

enormous abundance derived for these elements in the stellar atmosphere is almost certainly not real." Even scientists sometimes find it hard to accept new ideas that do not agree with what everyone "knows" to be right.



Figure 15.3 : Cecilia Payne-Gaposchkin (1900–1979). Her 1925 doctoral thesis laid the foundations for understanding the composition of the Sun and the stars. Yet, being a woman, she was not given a formal appointment at Harvard, where she worked, until 1938 and was not appointed a professor until 1956. (credit: Smithsonian Institution)

Before Payne-Gaposchkin's work, everyone assumed that the composition of the Sun and stars would be much like that of Earth. It was 3 years after her thesis that other studies proved beyond a doubt that the enormous abundance of hydrogen and helium in the Sun is indeed real. (And, as we will see, the composition of the Sun and the stars is much more typical of the makeup of the universe than the odd concentration of heavier elements that characterizes our planet.)

Most of the elements found in the Sun are in the form of atoms, with a small number of molecules, all in the form of gases: the Sun is so hot that no matter can survive as a liquid or a solid. In fact, the Sun is so hot that many of the atoms in it are *ionized*, that is, stripped of one or more of their electrons. This removal of electrons from their atoms means that there is a large quantity of free electrons and positively charged ions in the Sun, making it an electrically charged environment—quite different from the neutral one in which you are reading this text. (Scientists call such a hot ionized gas a plasma.)

In the nineteenth century, scientists observed a spectral line at 530.3 nanometers in the Sun's outer atmosphere, called the corona (a layer we will discuss in a minute.) This line had never been seen before, and so it was assumed that this line was the result of a new element found in the corona, quickly named coronium. It was not until 60 years later that astronomers discovered that this emission was in fact due to highly ionized iron—iron with 13 of its electrons stripped off. This is how we first discovered that the Sun's atmosphere had a temperature of more than a million degrees.

The Layers of the Sun beneath the Visible Surface

Figure 15.4 shows what the Sun would look like if we could see all parts of it from the center to its outer atmosphere; the terms in the figure will become familiar to you as you read on.

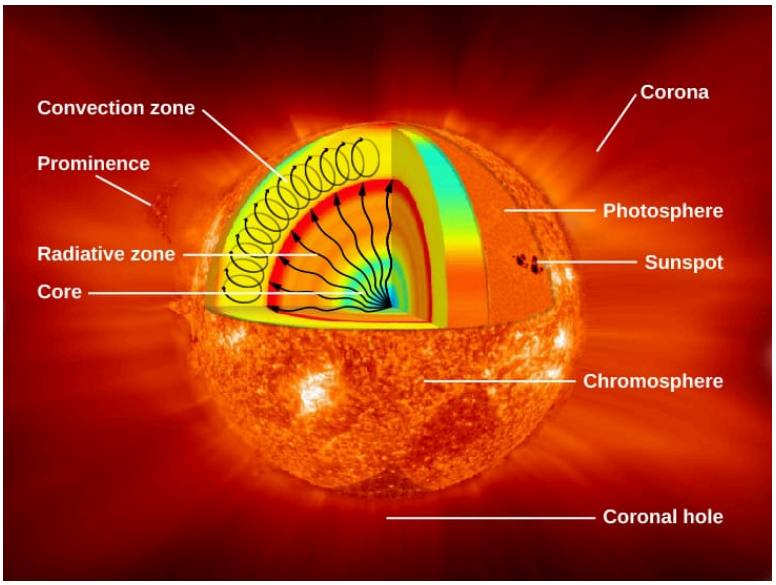


Figure 15.4 : Parts of the Sun. This illustration shows the different parts of the Sun, from the hot core where the energy is generated through regions where energy is transported outward, first by radiation, then by convection, and then out through the solar atmosphere. The parts of the atmosphere are also labeled the photosphere, chromosphere, and corona. Some typical features in the atmosphere are shown, such as coronal holes and prominences. (credit: modification of work by NASA/Goddard)

The Sun's layers are different from each other, and each plays a part in producing the energy that the Sun ultimately emits. We will begin with the core and work our way out through the layers. The Sun's *core* is extremely dense and is the source of all of its energy. Inside the core, nuclear energy is being released (in ways we will discuss in The Sun: A Nuclear Powerhouse). The core is approximately 20% of the size of the solar interior and is thought to have a temperature of approximately 15 million K, making it the hottest part of the Sun.

Above the core is a region known as the *radiative zone*—named for the primary mode of transporting energy across it. This region starts at about 25% of the distance to the solar surface and extends up to about 70% of the way to the surface. The light generated in the core is transported through the radiative zone very slowly, since the high density of matter in this region means a photon cannot travel too far without encountering a particle, causing it to change direction and lose some energy.

The *convective zone* is the outermost layer of the solar interior. It is a thick layer approximately 200,000 kilometers deep that transports energy from the edge of the radiative zone to the surface through giant convection cells, similar to a pot of boiling oatmeal. The plasma at the bottom of the convective zone is extremely hot, and it bubbles to the surface where it loses its heat to space. Once the plasma cools, it sinks back to the bottom of the convective zone.

Now that we have given a quick overview of the structure of the whole Sun, in this section, we will embark on a journey through the visible layers of the Sun, beginning with the photosphere—the visible surface.

The Solar Photosphere

Earth's air is generally transparent. But on a smoggy day in many cities, it can become opaque, which prevents us from seeing through it past a certain point. Something similar happens in the Sun. Its outer atmosphere is transparent, allowing us to look a short distance through it. But when we try to look through the atmosphere deeper into the Sun, our view is blocked. The photosphere is the layer where the Sun becomes opaque and marks the boundary past which we cannot see (Figure 15.5).

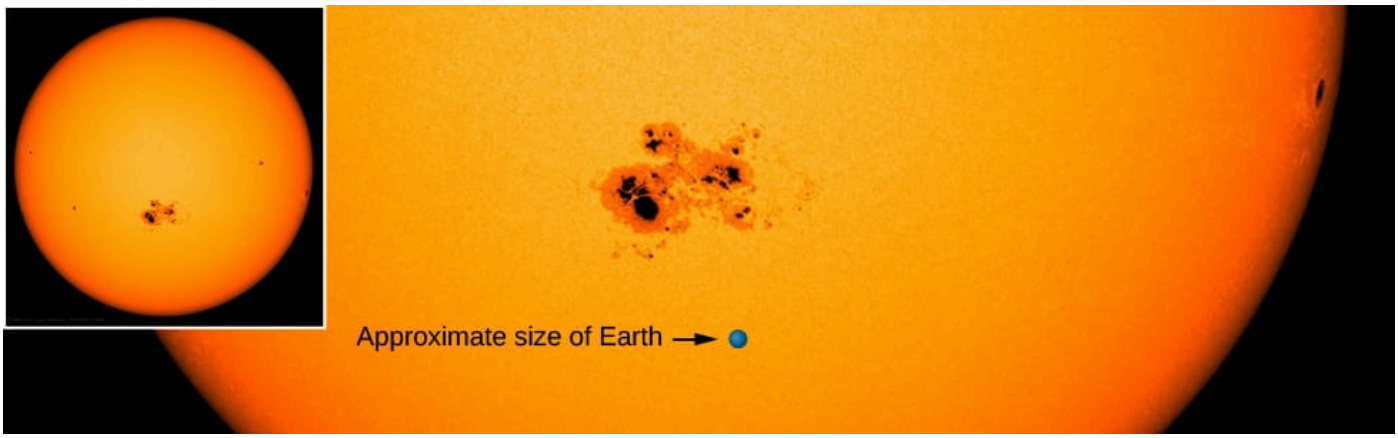


Figure 15.5 : Solar Photosphere plus Sunspots. This photograph shows the photosphere—the visible surface of the Sun. Also shown is an enlarged image of a group of sunspots; the size of Earth is shown for comparison. Sunspots appear darker because they are cooler than their surroundings. The typical temperature at the center of a large sunspot is about 3800 K, whereas the photosphere has a temperature of about 5800 K. (credit: modification of work by NASA/SDO)

As we saw, the energy that emerges from the photosphere was originally generated deep inside the Sun (more on this in The Sun: A Nuclear Powerhouse). This energy is in the form of photons, which make their way slowly toward the solar surface. Outside the Sun, we can observe *only* those photons that are emitted into the solar photosphere, where the density of atoms is sufficiently low and the photons can finally escape from the Sun without colliding with another atom or ion.

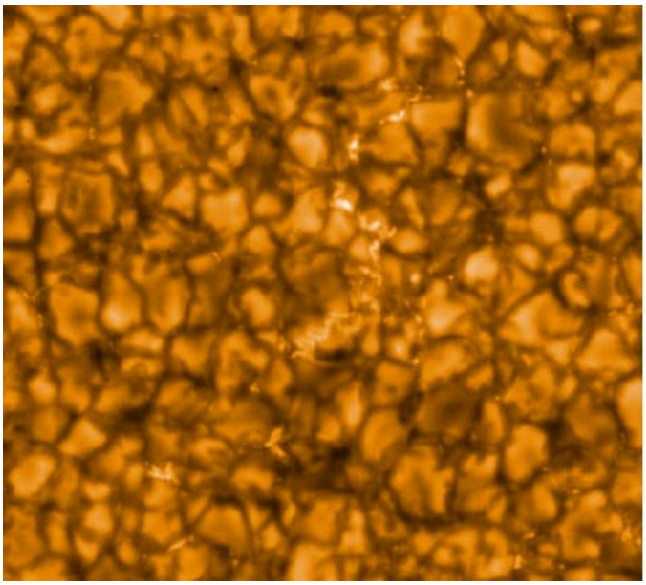
As an analogy, imagine that you are attending a big campus rally and have found a prime spot near the center of the action. Your friend arrives late and calls you on your cell phone to ask you to join her at the edge of the crowd. You decide that friendship is worth more than a prime spot, and so you work your way out through the dense crowd to meet her. You can move only a short distance before bumping into someone, changing direction, and trying again, making your way slowly to the outside edge of the crowd. All this while, your efforts are not visible to your waiting friend at the edge. Your friend can't see you until you get very close to the edge because of all the bodies in the way. So too photons making their way through the Sun are constantly bumping into atoms, changing direction, working their way slowly outward, and becoming visible only when they reach the atmosphere of the Sun where the density of atoms is too low to block their outward progress.

Astronomers have found that the solar atmosphere changes from almost perfectly transparent to almost completely opaque in a distance of just over 400 kilometers; it is this thin region that we call the *photosphere*, a word that comes from the Greek for "light sphere." When astronomers speak of the "diameter" of the Sun, they mean the size of the region surrounded by the photosphere.

The photosphere looks sharp only from a distance. If you were falling into the Sun, you would not feel any surface but would just sense a gradual increase in the density of the gas surrounding you. It is much the same as falling through a cloud while skydiving. From far away, the cloud looks as if it has a sharp surface, but you do not feel a surface as you fall into it. (One big difference between these two scenarios, however, is temperature. The Sun is so hot that you would be vaporized long before you reached the photosphere. Skydiving in Earth's atmosphere is much safer.)

We might note that the atmosphere of the Sun is not a very dense layer compared to the air in the room where you are reading this text. At a typical point in the photosphere, the pressure is less than 10% of Earth's pressure at sea level, and the density is about one ten-thousandth of Earth's atmospheric density at sea level.

Observations with telescopes show that the photosphere has a mottled appearance, resembling grains of rice spilled on a dark tablecloth or a pot of boiling oatmeal. This structure of the photosphere is called granulation (see Figure 15.6). Granules, which are typically 700 to 1000 kilometers in diameter (about the width of Texas), appear as bright areas surrounded by narrow, darker (cooler) regions. The lifetime of an individual granule is only 5 to 10 minutes. Even larger are supergranules, which are about 35,000 kilometers across (about the size of two Earths) and last about 24 hours.



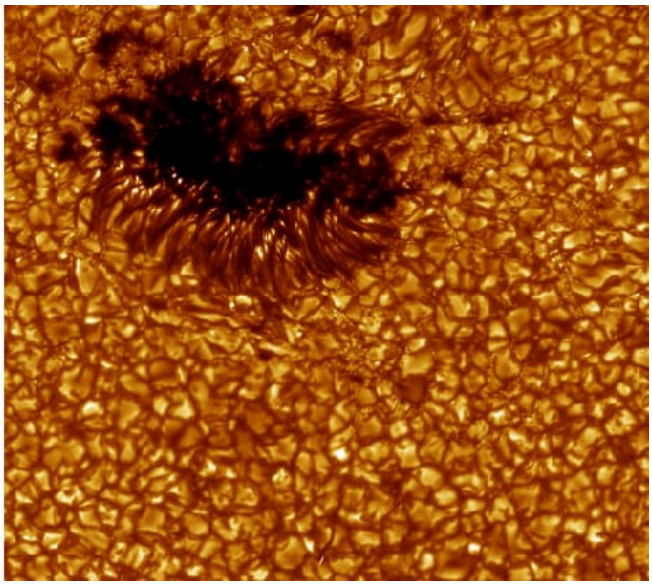


Figure 15.6 : Granulation Pattern. The surface markings of the convection cells create a granulation pattern on this dramatic image (left) taken from the Japanese Hinode spacecraft. You can see the same pattern when you heat up miso soup. The right image shows an irregular-shaped sunspot and granules on the Sun's surface, seen with the Swedish Solar Telescope on August 22, 2003. (credit left: modification of work by Hinode JAXA/NASA/PPARC; credit right: ISP/SST/Oddbjorn Engvold, Jun Elin Wiik, Luc Rouppe van der Voort)

The motions of the granules can be studied by examining the Doppler shifts in the spectra of gases just above them (see The Doppler Effect). The bright granules are columns of hotter gases rising at speeds of 2 to 3 kilometers per second from below the photosphere. As this rising gas reaches the photosphere, it spreads out, cools, and sinks down again into the darker regions between the granules. Measurements show that the centers of the granules are hotter than the intergranular regions by 50 to 100 K.

Link to Learning

See the "boiling" action of granulation in this 30-second time-lapse video from the Swedish Institute for Solar Physics.

The Chromosphere

The Sun's outer gases extend far beyond the photosphere (Figure 15.7). Because they are transparent to most visible radiation and emit only a small amount of light, these outer layers are difficult to observe. The region of the Sun's atmosphere that lies immediately above the photosphere is called the chromosphere. Until this century, the chromosphere was visible only when the photosphere was concealed by the Moon during a total solar eclipse (see the chapter on Earth, Moon, and Sky). In the seventeenth century, several observers described what appeared to them as a narrow red "streak" or "fringe" around the edge of the Moon during a brief instant after the Sun's photosphere had been covered. The name *chromosphere*, from the Greek for "colored sphere," was given to this red streak.

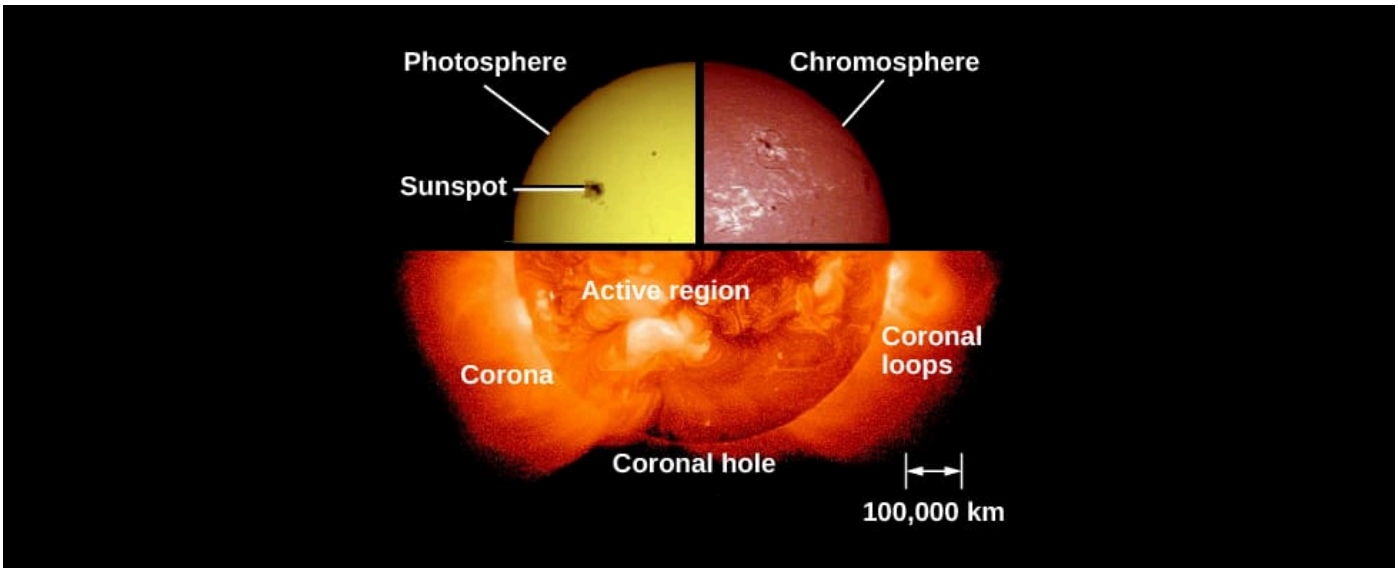


Figure 15.7 : The Sun's Atmosphere. Composite image showing the three components of the solar atmosphere: the photosphere or surface of the Sun taken in ordinary light; the chromosphere, imaged in the light of the strong red spectral line of hydrogen (H-alpha); and the corona as seen with X-rays. (credit: modification of work by NASA)

Observations made during eclipses show that the chromosphere is about 2000 to 3000 kilometers thick, and its spectrum consists of bright emission lines, indicating that this layer is composed of hot gases emitting light at discrete wavelengths. The reddish color of the chromosphere arises from one of the strongest emission lines in the visible part of its spectrum—the bright red line caused by hydrogen, the element that, as we have already seen, dominates the composition of the Sun.

In 1868, observations of the chromospheric spectrum revealed a yellow emission line that did not correspond to any previously known element on Earth. Scientists quickly realized they had found a new element and named it *helium* (after *helios*, the Greek word for "Sun"). It took until 1895 for helium to be discovered on our planet. Today, students are probably most familiar with it as the light gas used to inflate balloons, although it turns out to be the second-most abundant element in the universe.

The temperature of the chromosphere is about 10,000 K. This means that the chromosphere is hotter than the photosphere, which should seem surprising. In all the situations we are familiar with, temperatures fall as one moves away from the source of heat, and the chromosphere is farther from the center of the Sun than the photosphere is.

The Transition Region

The increase in temperature does not stop with the chromosphere. Above it is a region in the solar atmosphere where the temperature changes from 10,000 K (typical of the chromosphere) to nearly a million degrees. The hottest part of the solar atmosphere, which has a temperature of a million degrees or more, is called the corona. Appropriately, the part of the Sun where the rapid temperature rise occurs is called the transition region. It is probably only a few tens of kilometers thick. Figure 15.8 summarizes how the temperature of the solar atmosphere changes from the photosphere outward.

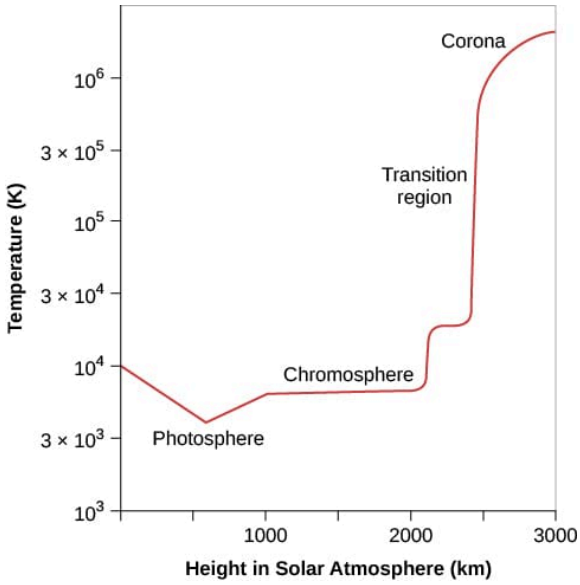


Figure 15.8 : Temperatures in the Solar Atmosphere. On this graph, temperature is shown increasing upward, and height above the photosphere is shown increasing to the right. Note the very rapid increase in temperature over a very short distance in the transition region between the chromosphere and the corona.

In 2013, NASA launched the Interface Region Imaging Spectrograph (IRIS) to study the transition region to understand better how and why this sharp temperature increase occurs. IRIS is the first space mission that is able to obtain high spatial resolution images of the different features produced over this wide temperature range and to see how they change with time and location (Figure 15.9).

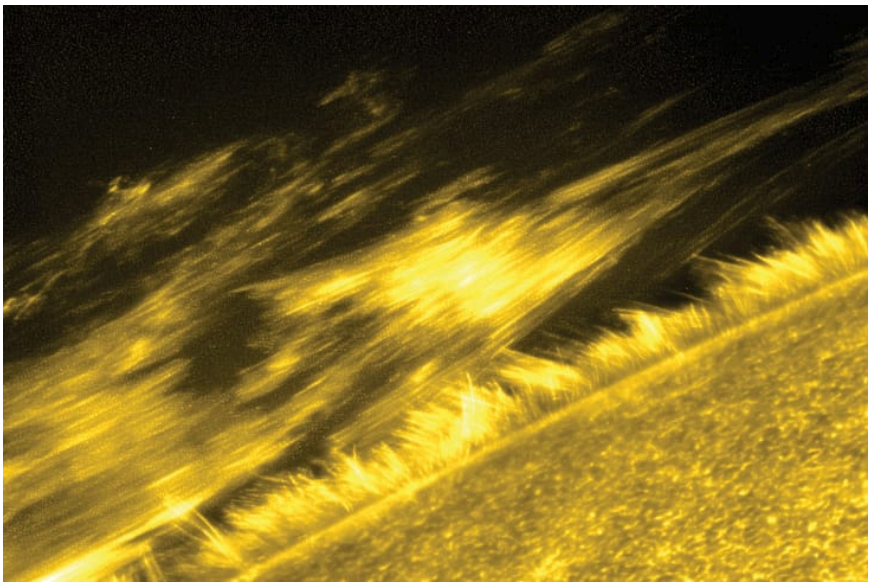


Figure 15.9 : Portion of the Transition Region. This image shows a giant ribbon of relatively cool gas threading through the lower portion of the hot corona. This ribbon (the technical term is filament) is made up of many individual threads. Time-lapse movies of this filament showed that it gradually heated as it moved through the corona. Scientists study events like this in order to try to understand what heats the chromosphere and corona to high temperatures. The “whiskers” at the edge of the Sun are spicules, jets of gas that shoot material up from the Sun’s surface and disappear after only a few minutes. This single image gives a hint of just how complicated it is to construct a model of the all the different structures and heating mechanisms in the solar atmosphere. (credit: JAXA/NASA/Hinode)

Figure 15.4 and the red graph in Figure 15.8 make the Sun seem rather like an onion, with smooth spherical shells, each one with a different temperature. For a long time, astronomers did indeed think of the Sun this way. However, we now know that while this idea of layers—photosphere, chromosphere, transition region, corona—describes the big picture fairly well, the Sun’s atmosphere is really more complicated, with hot and cool regions intermixed. For example, clouds of carbon monoxide gas with temperatures colder than 4000 K have now been found at the same height above the photosphere as the much hotter gas of the chromosphere.

The Corona

The outermost part of the Sun's atmosphere is called the *corona*. Like the chromosphere, the corona was first observed during total eclipses (Figure 15.10). Unlike the chromosphere, the corona has been known for many centuries: it was referred to by the Roman historian Plutarch and was discussed in some detail by Kepler.

The corona extends millions of kilometers above the photosphere and emits about half as much light as the full moon. The reason we don't see this light until an eclipse occurs is the overpowering brilliance of the photosphere. Just as bright city lights make it difficult to see faint starlight, so too does the intense light from the photosphere hide the faint light from the corona. While the best time to see the corona from Earth is during a total solar eclipse, it can be observed easily from orbiting spacecraft. Its brighter parts can now be photographed with a special instrument—a coronagraph—that removes the Sun's glare from the image with an occulting disk (a circular piece of material held so it is just in front of the Sun).

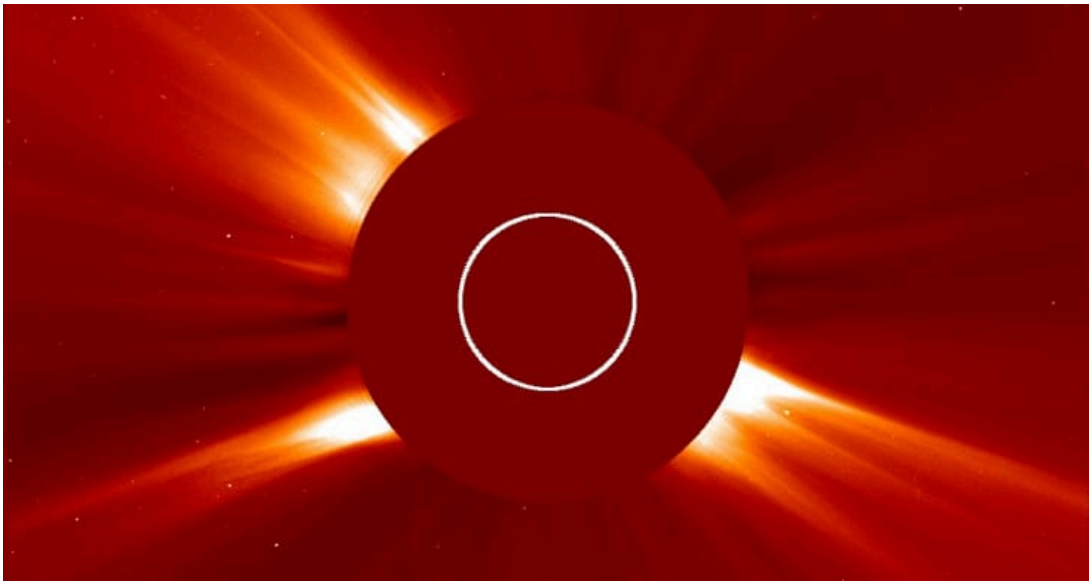


Figure 15.10 : Coronagraph. This image of the Sun was taken March 2, 2016. The larger dark circle in the center is the disk the blocks the Sun's glare, allowing us to see the corona. The smaller inner circle is where the Sun would be if it were visible in this image. (credit: modification of work by NASA/SOHO)

Studies of its spectrum show the corona to be very low in density. At the bottom of the corona, there are only about 10^9 atoms per cubic centimeter, compared with about 10^{16} atoms per cubic centimeter in the upper photosphere and 10^{19} molecules per cubic centimeter at sea level in Earth's atmosphere. The corona thins out very rapidly at greater heights, where it corresponds to a high vacuum by Earth laboratory standards. The corona extends so far into space—far past Earth—that here on our planet, we are technically living in the Sun's atmosphere.

The Solar Wind

One of the most remarkable discoveries about the Sun's atmosphere is that it produces a stream of charged particles (mainly protons and electrons) that we call the solar wind. These particles flow outward from the Sun into the solar system at a speed of about 400 kilometers per second (almost 1 million miles per hour)! The solar wind exists because the gases in the corona are so hot and moving so rapidly that they cannot be held back by solar gravity. (This wind was actually discovered by its effects on the charged tails of comets; in a sense, we can see the comet tails blow in the solar breeze the way wind socks at an airport or curtains in an open window flutter on Earth.)

Although the solar wind material is very, very rarified (i.e., *extremely* low density), the Sun has an enormous surface area. Astronomers estimate that the Sun is losing about 1–2 million tons of material each second through this wind. Although this sounds like a lot, it's so trivial compared to the enormous mass of the Sun that it can be neglected as we study the Sun.

From where in the Sun does the solar wind emerge? In visible photographs, the solar corona appears fairly uniform and smooth. X-ray and extreme ultraviolet pictures, however, show that the corona has loops, plumes, and both bright and dark regions. Large

dark regions of the corona that are relatively cool and quiet are called coronal holes (Figure 15.11). In these regions, magnetic field lines stretch far out into space away from the Sun, rather than looping back to the surface. The solar wind comes predominantly from coronal holes, where gas can stream away from the Sun into space unhindered by magnetic fields. Hot coronal gas, on the other hand, is present mainly where magnetic fields have trapped and concentrated it.

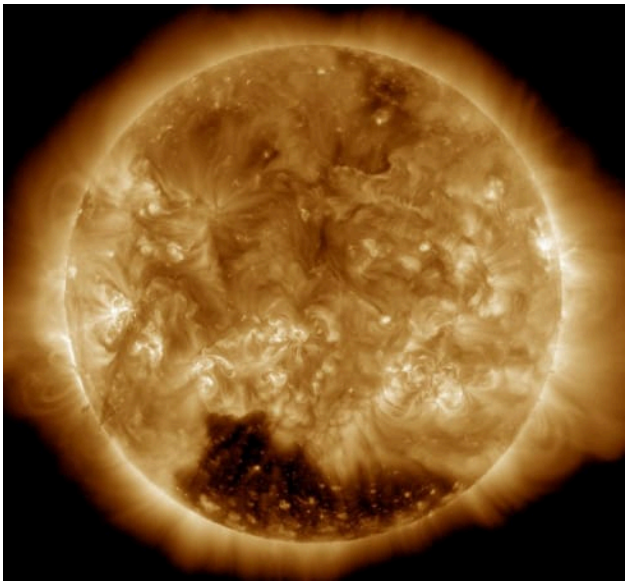


Figure 15.11 Coronal Hole. The dark area visible near the Sun's south pole on this Solar Dynamics Observer spacecraft image is a coronal hole. (credit: modification of work by NASA/SDO)

At the surface of Earth, we are protected to some degree from the solar wind by our atmosphere and Earth's magnetic field (see Earth as a Planet). However, the magnetic field lines come into Earth at the north and south magnetic poles. Here, charged particles accelerated by the solar wind can follow the field down into our atmosphere. As the particles strike molecules of air, they cause them to glow, producing beautiful curtains of light called the auroras, or the northern and southern lights (Figure 15.12).



Figure 15.12 : Aurora. The colorful glow in the sky results from charged particles in a solar wind interacting with Earth's magnetic fields. The stunning display captured here occurred over Jokulsarlon Lake in Iceland in 2013. (credit: Moyan Brenn)

Link to Learning

This NASA video explains and demonstrates the nature of the auroras and their relationship to Earth's magnetic field.

This page titled 5.1: The Structure and Composition of the Sun is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by OpenStax via source content that was edited to the style and standards of the LibreTexts platform.

5.2: Understanding Stars

This section examines how astronomers determine the physical properties of stars from the light they observe. It explores stellar brightness, color, and spectra and shows how these observations can reveal properties such as temperature, composition, motion, and size. These characteristics are brought together in the H-R diagram, an important tool for comparing stars and recognizing patterns among stellar populations.

This page titled 5.2: Understanding Stars is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Robert Wagner.

5.2.1: The Brightness of Stars

Learning Objectives

By the end of this section, you will be able to:

- Explain the difference between luminosity and apparent brightness
- Understand how astronomers specify brightness with magnitudes

Luminosity

Perhaps the most important characteristic of a star is its luminosity—the total amount of energy at all wavelengths that it emits per second. Earlier, we saw that the Sun puts out a tremendous amount of energy every second. (And there are stars far more luminous than the Sun out there.) To make the comparison among stars easy, astronomers express the luminosity of other stars in terms of the Sun's luminosity. For example, the luminosity of Sirius is about 25 times that of the Sun. We use the symbol L_{Sun} to denote the Sun's luminosity; hence, that of Sirius can be written as 25 L_{Sun}. In a later chapter, we will see that if we can measure how much energy a star emits and we also know its mass, then we can calculate how long it can continue to shine before it exhausts its nuclear energy and begins to die.

Apparent Brightness

Astronomers are careful to distinguish between the luminosity of the star (the total energy output) and the amount of energy that happens to reach our eyes or a telescope on Earth. Stars are democratic in how they produce radiation; they emit the same amount of energy in every direction in space. Consequently, only a minuscule fraction of the energy given off by a star actually reaches an observer on Earth. We call the amount of a star's energy that reaches a given area (say, one square meter) each second here on Earth its apparent brightness. If you look at the night sky, you see a wide range of apparent brightnesses among the stars. Most stars, in fact, are so dim that you need a telescope to detect them.

If all stars were the same luminosity—if they were like standard bulbs with the same light output—we could use the difference in their apparent brightnesses to tell us something we very much want to know: how far away they are. Imagine you are in a big concert hall or ballroom that is dark except for a few dozen 25-watt bulbs placed in fixtures around the walls. Since they are all 25-watt bulbs, their luminosity (energy output) is the same. But from where you are standing in one corner, they do *not* have the same apparent brightness. Those close to you appear brighter (more of their light reaches your eye), whereas those far away appear dimmer (their light has spread out more before reaching you). In this way, you can tell which bulbs are closest to you. In the same way, if all the stars had the same luminosity, we could immediately infer that the brightest-appearing stars were close by and the dimmest-appearing ones were far away.

To pin down this idea more precisely, recall from the Radiation and Spectra chapter that we know exactly how light fades with increasing distance. The energy we receive is inversely proportional to the square of the distance. If, for example, we have two stars of the same luminosity and one is twice as far away as the other, it will look four times dimmer than the closer one. If it is three times farther away, it will look nine (three squared) times dimmer, and so forth.

Alas, the stars do not all have the same luminosity. (Actually, we are pretty glad about that because having many different types of stars makes the universe a much more interesting place.) But this means that if a star looks dim in the sky, we cannot tell whether it appears dim because it has a low luminosity but is relatively nearby, or because it has a high luminosity but is very far away. To measure the luminosities of stars, we must first compensate for the dimming effects of distance on light, and to do that, we must know how far away they are. Distance is among the most difficult of all astronomical measurements. We will return to how it is determined after we have learned more about the stars. For now, we will describe how astronomers specify the apparent brightness of stars.

The Magnitude Scale

The process of measuring the apparent brightness of stars is called *photometry* (from the Greek *photo* meaning "light" and *–metry* meaning "to measure"). As we saw Observing the Sky: The Birth of Astronomy, astronomical photometry began with Hipparchus. Around 150 B.C.E., he erected an observatory on the island of Rhodes in the Mediterranean. There he prepared a catalog of nearly 1000 stars that included not only their positions but also estimates of their apparent brightnesses.

Hipparchus did not have a telescope or any instrument that could measure apparent brightness accurately, so he simply made estimates with his eyes. He sorted the stars into six brightness categories, each of which he called a magnitude. He referred to the brightest stars in his catalog as first-magnitudes stars, whereas those so faint he could barely see them were sixth-magnitude stars. During the nineteenth century, astronomers attempted to make the scale more precise by establishing exactly how much the apparent brightness of a sixth-magnitude star differs from that of a first-magnitude star. Measurements showed that we receive about 100 times more light from a first-magnitude star than from a sixth-magnitude star. Based on this measurement, astronomers then defined an accurate magnitude system in which a difference of five magnitudes corresponds exactly to a brightness ratio of 100:1. In addition, the magnitudes of stars are decimalized; for example, a star isn't just a "second-magnitude star," it has a magnitude of 2.0 (or 2.1, 2.3, and so forth). So what number is it that, when multiplied together five times, gives you this factor of 100? Play on your calculator and see if you can get it. The answer turns out to be about 2.5, which is the fifth root of 100. This means that a magnitude 1.0 star and a magnitude 2.0 star differ in brightness by a factor of about 2.5. Likewise, we receive about 2.5 times as much light from a magnitude 2.0 star as from a magnitude 3.0 star. What about the difference between a magnitude 1.0 star and a magnitude 3.0 star? Since the difference is 2.5 times for each "step" of magnitude, the total difference in brightness is 2.5 × 2.5 = 6.25 times.

Here are a few rules of thumb that might help those new to this system. If two stars differ by 0.75 magnitudes, they differ by a factor of about 2 in brightness. If they are 2.5 magnitudes apart, they differ in brightness by a factor of 10, and a 4-magnitude difference corresponds to a difference in brightness of a factor of 40.You might be saying to yourself at this point, "Why do astronomers continue to use this complicated system from more than 2000 years ago?" That's an excellent question and, as we shall discuss, astronomers today can use other ways of expressing how bright a star looks. But because this system is still used in many books, star charts, and computer apps, we felt we had to introduce students to it (even though we were very tempted to leave it out.)

The brightest stars, those that were traditionally referred to as first-magnitude stars, actually turned out (when measured accurately) not to be identical in brightness. For example, the brightest star in the sky, Sirius, sends us about 10 times as much light as the average first-magnitude star. On the modern magnitude scale, Sirius, the star with the brightest apparent magnitude, has been assigned a magnitude of –1.5. Other objects in the sky can appear even brighter. Venus at its brightest is of magnitude –4.4, while the Sun has a magnitude of –26.8. Figure 17.2 shows the range of observed magnitudes from the brightest to the faintest, along with the actual magnitudes of several well-known objects. The important fact to remember when using magnitude is that the system goes backward: the *larger* the magnitude, the *fainter* the object you are observing.

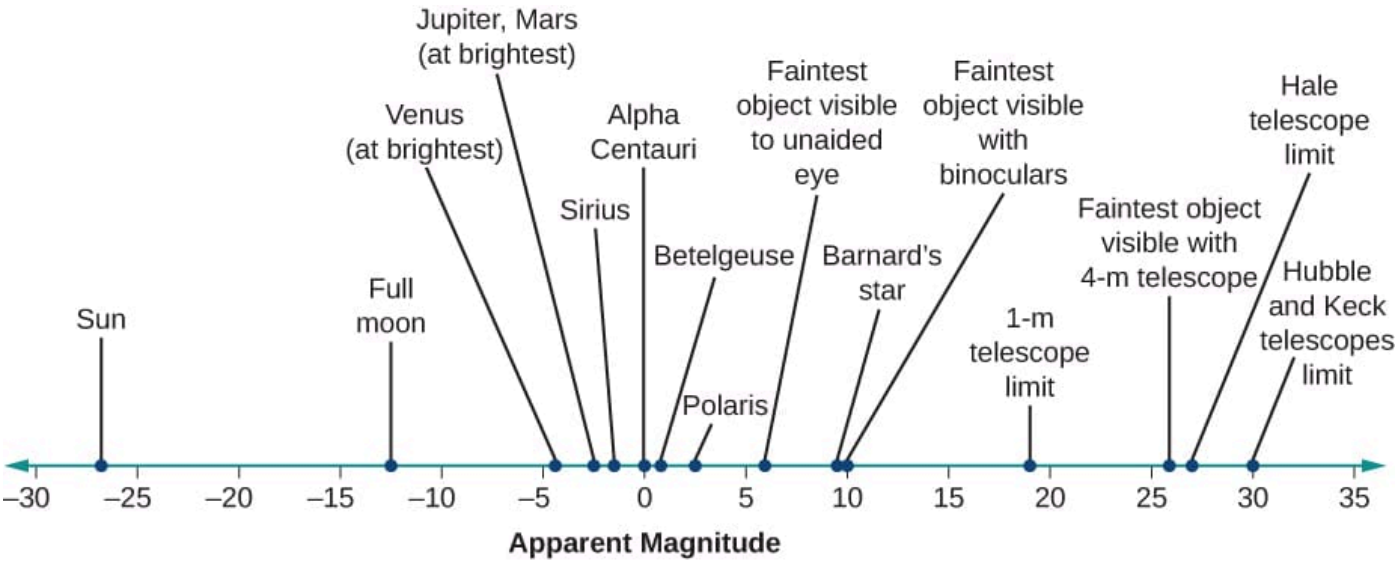


Figure 17.2 : Apparent Magnitudes of Well-Known Objects. The faintest magnitudes that can be detected by the unaided eye, binoculars, and large telescopes are also shown.

Example 17.1: The Magnitude Equation

Even scientists can't calculate fifth roots in their heads, so astronomers have summarized the above discussion in an equation to help calculate the difference in brightness for stars with different magnitudes. If m_1 and m_2 are the magnitudes of two stars, then we can calculate the ratio of their brightness (b2b1)(b2b1) using this equation:

m1−m2=2.5log(b2b1)orb2b1=2.5m1−m2m1−m2=2.5log(b2b1)orb2b1=2.5m1−m2

Here is another way to write this equation:

b2b1=(1000.2)m1−m2b2b1=(1000.2)m1−m2

Let's do a real example, just to show how this works. Imagine that an astronomer has discovered something special about a dim star (magnitude 8.5), and she wants to tell her students how much dimmer the star is than Sirius. Star 1 in the equation will be our dim star and star 2 will be Sirius.

Solution

Remember, Sirius has a magnitude of -1.5-1.5. In that case:

b2b1=(1000.2)8.5−(−1.5)=(1000.2)10=(100)2=100×100=10,000b2b1=(1000.2)8.5−(−1.5)=(1000.2)10=(100)2=100×100=10,000

Exercise 5.2.1.1

It is a common misconception that Polaris (magnitude 2.0) is the brightest star in the sky, but, as we saw, that distinction actually belongs to Sirius (magnitude –1.5). How does Sirius' apparent brightness compare to that of Polaris?

Answer available at the end of this section.

Other Units of Brightness

Although the magnitude scale is still used for visual astronomy, it is not used at all in newer branches of the field. In radio astronomy, for example, no equivalent of the magnitude system has been defined. Rather, radio astronomers measure the amount of energy being collected each second by each square meter of a radio telescope and express the brightness of each source in terms of, for example, watts per square meter.

Similarly, most researchers in the fields of infrared, X-ray, and gamma-ray astronomy use energy per area per second rather than magnitudes to express the results of their measurements. Nevertheless, astronomers in all fields are careful to distinguish between the *luminosity* of the source (even when that luminosity is all in X-rays) and the amount of energy that happens to reach us on Earth. After all, the luminosity is a really important characteristic that tells us a lot about the object in question, whereas the energy that reaches Earth is an accident of cosmic geography.

To make the comparison among stars easy, in this text, we avoid the use of magnitudes as much as possible and will express the luminosity of other stars in terms of the Sun's luminosity. For example, the luminosity of Sirius is 25 times that of the Sun. We use the symbol L_{Sun} to denote the Sun's luminosity; hence, that of Sirius can be written as 25 L_{Sun}.

Answers

Exercise 5.2.1.1

bSiriusbPolaris=(1000.2)2.0−(−1.5)=(1000.2)3.5=1000.7=25bSiriusbPolaris=(1000.2)2.0−(−1.5)=(1000.2)3.5=1000.7=25

(Hint: If you only have a basic calculator, you may wonder how to take 100 to the 0.7th power. But this is something you can ask Google to do. Google now accepts mathematical questions and will answer them. So try it for yourself. Ask Google, "What is 100 to the 0.7th power?")

Our calculation shows that Sirius' apparent brightness is 25 times greater than Polaris' apparent brightness.

This page titled 5.2.1: The Brightness of Stars is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by OpenStax via source content that was edited to the style and standards of the LibreTexts platform.

5.2.2: Colors of Stars

Learning Objectives

By the end of this section, you will be able to:

- Compare the relative temperatures of stars based on their colors
- Understand how astronomers use color indexes to measure the temperatures of stars

Look at the beautiful picture of the stars in the Sagittarius Star Cloud shown in Figure 17.3. The stars show a multitude of colors, including red, orange, yellow, white, and blue. As we have seen, stars are not all the same color because they do not all have identical temperatures. To define *color* precisely, astronomers have devised quantitative methods for characterizing the color of a star and then using those colors to determine stellar temperatures. In the chapters that follow, we will provide the temperature of the stars we are describing, and this section tells you how those temperatures are determined from the colors of light the stars give off.

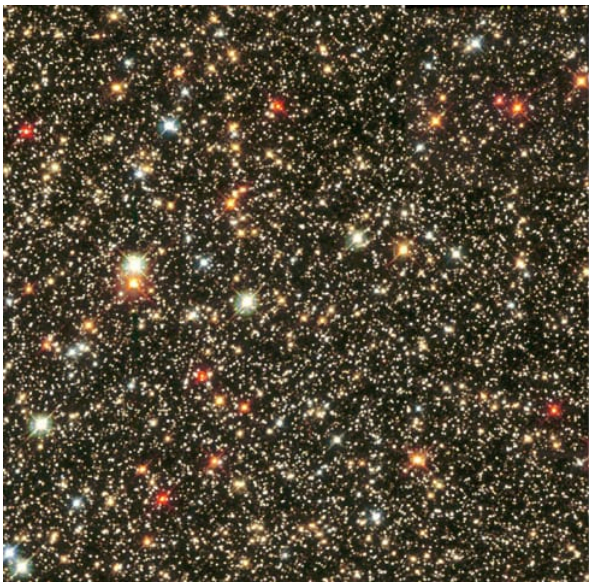


Figure 17.3 : Sagittarius Star Cloud. This image, which was taken by the Hubble Space Telescope, shows stars in the direction toward the center of the Milky Way Galaxy. The bright stars glitter like colored jewels on a black velvet background. The color of a star indicates its temperature. Blue-white stars are much hotter than the Sun, whereas red stars are cooler. On average, the stars in this field are at a distance of about 25,000 light-years (which means it takes light 25,000 years to traverse the distance from them to us) and the width of the field is about 13.3 light-years. (credit: Hubble Heritage Team (AURA/STScI/NASA))

Color and Temperature

As we learned in The Electromagnetic Spectrum section, Wien's law relates stellar color to stellar temperature. Blue colors dominate the visible light output of very hot stars (with much additional radiation in the ultraviolet). On the other hand, cool stars emit most of their visible light energy at red wavelengths (with more radiation coming off in the infrared) (Table 17.1). The color of a star therefore provides a measure of its intrinsic or true surface temperature (apart from the effects of reddening by interstellar dust, which will be discussed in Between the Stars: Gas and Dust in Space). Color does not depend on the distance to the object. This should be familiar to you from everyday experience. The color of a traffic signal, for example, appears the same no matter how far away it is. If we could somehow take a star, observe it, and then move it much farther away, its apparent brightness (magnitude) would change. But this change in brightness is the same for all wavelengths, and so its color would remain the same.

Table 17.1: Example Star Colors and Corresponding Approximate Temperatures

Star Color	Approximate Temperature	Example
Blue	25,000 K	Spica
White	10,000 K	Vega
Yellow	6000 K	Sun
Orange	4000 K	Aldebaran
Red	3000 K	Betelgeuse

Link to Learning

Go to this interactive simulation from the University of Colorado to see the color of a star changing as the temperature is changed.

The hottest stars have temperatures of over 40,000 K, and the coolest stars have temperatures of about 2000 K. Our Sun's surface temperature is about 6000 K; its peak wavelength color is a slightly greenish-yellow. In space, the Sun would look white, shining with about equal amounts of reddish and bluish wavelengths of light. It looks somewhat yellow as seen from Earth's surface because our planet's nitrogen molecules scatter some of the shorter (i.e., blue) wavelengths out of the beams of sunlight that reach us, leaving more long wavelength light behind. This also explains why the sky is blue: the blue sky is sunlight scattered by Earth's atmosphere.

Color Indices

In order to specify the exact color of a star, astronomers normally measure a star's apparent brightness through filters, each of which transmits only the light from a particular narrow band of wavelengths (colors). A crude example of a filter in everyday life is a green-colored, plastic, soft drink bottle, which, when held in front of your eyes, lets only the green colors of light through.

One commonly used set of filters in astronomy measures stellar brightness at three wavelengths corresponding to ultraviolet, blue, and yellow light. The filters are named: U (ultraviolet), B (blue), and V (visual, for yellow). These filters transmit light near the wavelengths of 360 nanometers (nm), 420 nm, and 540 nm, respectively. The brightness measured through each filter is usually expressed in magnitudes. The difference between any two of these magnitudes—say, between the blue and the visual magnitudes (B–V)—is called a color index.

By agreement among astronomers, the ultraviolet, blue, and visual magnitudes of the UBV system are adjusted to give a color index of 0 to a star with a surface temperature of about 10,000 K, such as Vega. The B–V color indexes of stars range from –0.4 for the bluest stars, with temperatures of about 40,000 K, to +2.0 for the reddest stars, with temperatures of about 2000 K. The B–V index for the Sun is about +0.65. Note that, by convention, the B–V index is always the "bluer" minus the "redder" color.

Why use a color index if it ultimately implies temperature? Because the brightness of a star through a filter is what astronomers actually measure, and we are always more comfortable when our statements have to do with measurable quantities.

This page titled 5.2.2: Colors of Stars is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by OpenStax via source content that was edited to the style and standards of the LibreTexts platform.

5.2.3: The Spectra of Stars (and Brown Dwarfs)

Learning Objectives

By the end of this section, you will be able to:

- Describe how astronomers use spectral classes to characterize stars
- Explain the difference between a star and a brown dwarf

Measuring colors is only one way of analyzing starlight. Another way is to use a spectrograph to spread out the light into a spectrum (see the Radiation and Spectra and the Astronomical Instruments chapters). In 1814, the German physicist Joseph Fraunhofer observed that the spectrum of the Sun shows dark lines crossing a continuous band of colors. In the 1860s, English astronomers Sir William Huggins and Lady Margaret Huggins (Figure 17.4) succeeded in identifying some of the lines in stellar spectra as those of known elements on Earth, showing that the same chemical elements found in the Sun and planets exist in the stars. Since then, astronomers have worked hard to perfect experimental techniques for obtaining and measuring spectra, and they have developed a theoretical understanding of what can be learned from spectra. Today, spectroscopic analysis is one of the cornerstones of astronomical research.

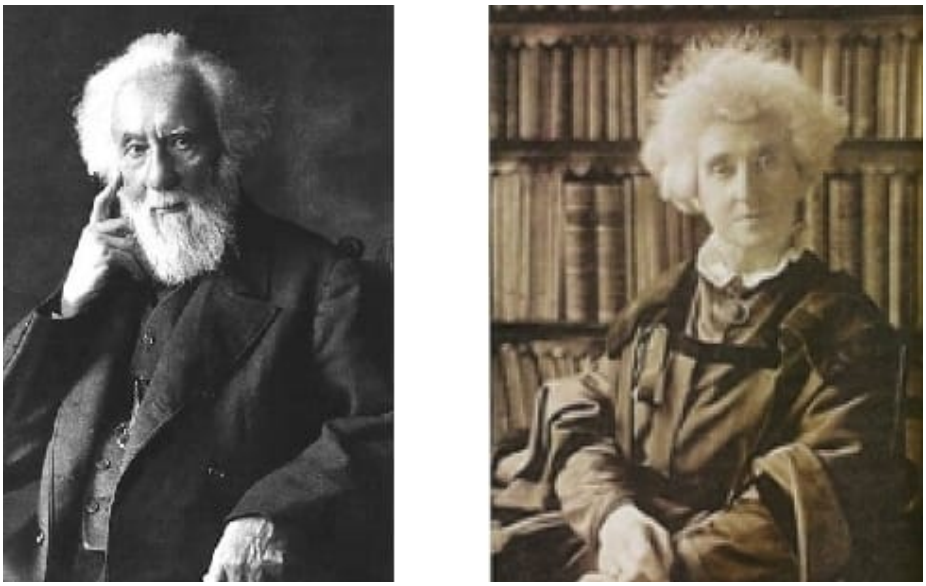


Figure 17.4 : William Huggins (1824–1910) and Margaret Huggins (1848–1915). William and Margaret Huggins were the first to identify the lines in the spectrum of a star other than the Sun; they also took the first spectrogram, or photograph of a stellar spectrum.

Formation of Stellar Spectra

When the spectra of different stars were first observed, astronomers found that they were not all identical. Since the dark lines are produced by the chemical elements present in the stars, astronomers first thought that the spectra differ from one another because stars are not all made of the same chemical elements. This hypothesis turned out to be wrong. *The primary reason that stellar spectra look different is because the stars have different temperatures*. Most stars have nearly the same composition as the Sun, with only a few exceptions.

Hydrogen, for example, is by far the most abundant element in most stars. However, lines of hydrogen are not seen in the spectra of the hottest and the coolest stars. In the atmospheres of the hottest stars, hydrogen atoms are completely ionized. Because the electron and the proton are separated, ionized hydrogen cannot produce absorption lines. (Recall from the Formation of Spectral Lines section, the lines are the result of electrons in orbit around a nucleus changing energy levels.)

In the atmospheres of the coolest stars, hydrogen atoms have their electrons attached and can switch energy levels to produce

lines. However, practically all of the hydrogen atoms are in the lowest energy state (unexcited) in these stars and thus can absorb only those photons able to lift an electron from that first energy level to a higher level. Photons with enough energy to do this lie in the ultraviolet part of the electromagnetic spectrum, and there are very few ultraviolet photons in the radiation from a cool star. What this means is that if you observe the spectrum of a very hot or very cool star with a typical telescope on the surface of Earth, the most common element in that star, hydrogen, will show very weak spectral lines or none at all.

The hydrogen lines in the visible part of the spectrum (called *Balmer lines*) are strongest in stars with intermediate temperatures—not too hot and not too cold. Calculations show that the optimum temperature for producing visible hydrogen lines is about 10,000 K. At this temperature, an appreciable number of hydrogen atoms are excited to the second energy level. They can then absorb additional photons, rise to still-higher levels of excitation, and produce a dark absorption line. Similarly, every other chemical element, in each of its possible stages of ionization, has a characteristic temperature at which it is most effective in producing absorption lines in any particular part of the spectrum.

Classification of Stellar Spectra

Astronomers use the patterns of lines observed in stellar spectra to sort stars into a spectral class. Because a star's temperature determines which absorption lines are present in its spectrum, these spectral classes are a measure of its surface temperature. There are seven standard spectral classes. From hottest to coldest, these seven spectral classes are designated O, B, A, F, G, K, and M. Recently, astronomers have added three additional classes for even cooler objects—L, T, and Y.

At this point, you may be looking at these letters with wonder and asking yourself why astronomers didn't call the spectral types A, B, C, and so on. You will see, as we tell you the history, that it's an instance where tradition won out over common sense.

In the 1880s, Williamina Fleming devised a system to classify stars based on the strength of hydrogen absorption lines. Spectra with the strongest lines were classified as "A" stars, the next strongest "B," and so on down the alphabet to "O" stars, in which the hydrogen lines were very weak. But we saw above that hydrogen lines alone are not a good indicator for classifying stars, since their lines disappear from the visible light spectrum when the stars get too hot or too cold.

In the 1890s, Annie Jump Cannon revised this classification system, focusing on just a few letters from the original system: A, B, F, G, K, M, and O. Instead of starting over, Cannon also rearranged the existing classes—in order of decreasing temperature—into the sequence we have learned: O, B, A, F, G, K, M. As you can read in the feature on Annie Cannon: Classifier of the Stars in this chapter, she classified around 500,000 stars over her lifetime, classifying up to three stars per minute by looking at the stellar spectra.

Link to Learning

For a deep dive into spectral types, explore the interactive project at the Sloan Digital Sky Survey in which you can practice classifying stars yourself.

To help astronomers remember this crazy order of letters, Cannon created a mnemonic, "Oh Be A Fine Girl, Kiss Me." (If you prefer, you can easily substitute "Guy" for "Girl.") Other mnemonics, which we hope will not be relevant for you, include "Oh Brother, Astronomers Frequently Give Killer Midterms" and "Oh Boy, An F Grade Kills Me!" With the new L, T, and Y spectral classes, the mnemonic might be expanded to "Oh Be A Fine Girl (Guy), Kiss Me Like That, Yo!"

Each of these spectral classes, except possibly for the Y class which is still being defined, is further subdivided into 10 subclasses designated by the numbers 0 through 9. A B0 star is the hottest type of B star; a B9 star is the coolest type of B star and is only slightly hotter than an A0 star.

And just one more item of vocabulary: for historical reasons, astronomers call all the elements heavier than helium *metals*, even though most of them do not show metallic properties. (If you are getting annoyed at the peculiar jargon that astronomers use, just bear in mind that every field of human activity tends to develop its own specialized vocabulary. Just try reading a credit card or social media agreement form these days without training in law!)

Let's take a look at some of the details of how the spectra of the stars change with temperature. (It is these details that allowed Annie Cannon to identify the spectral types of stars as quickly as three per minute!) As Figure 17.5 shows, in the hottest O stars (those with temperatures over 28,000 K), only lines of ionized helium and highly ionized atoms of other elements are conspicuous. Hydrogen lines are strongest in A stars with atmospheric temperatures of about 10,000 K. Ionized metals provide

the most conspicuous lines in stars with temperatures from 6000 to 7500 K (spectral type F). In the coolest M stars (below 3500 K), absorption bands of titanium oxide and other molecules are very strong. By the way, the spectral class assigned to the Sun is G2. The sequence of spectral classes is summarized in Table 17.2.

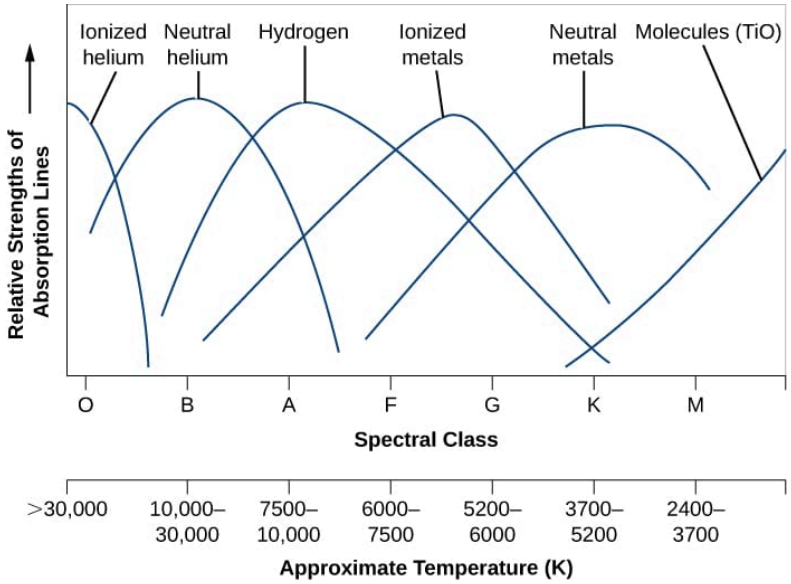


Figure 17.5 : Absorption Lines in Stars of Different Temperatures. This graph shows the strengths of absorption lines of different chemical species (atoms, ions, molecules) as we move from hot (left) to cool (right) stars. The sequence of spectral types is also shown.

Link to Learning

Use the Spectrum Explorer to see how each element's absorption lines change with temperature.

Table 17.2: Spectral Classes for Stars

Spectral Class	Color	Approximate Temperature (K)	Principal Features	Examples
O	Blue	> 30,000	Neutral and ionized helium lines, weak hydrogen lines	10 Lacertae
B	Blue-white	10,000–30,000	Neutral helium lines, strong hydrogen lines	Rigel, Spica
A	White	7500–10,000	Strongest hydrogen lines, weak ionized calcium lines, weak ionized metal (e.g., iron, magnesium) lines	Sirius, Vega
F	Yellow-white	6000–7500	Strong hydrogen lines, strong ionized calcium lines, weak sodium lines, many ionized metal lines	Canopus, Procyon
G	Yellow	5200–6000	Weaker hydrogen lines, strong ionized calcium lines, strong sodium lines, many lines of ionized and neutral metals	Sun, Capella
K	Orange	3700–5200	Very weak hydrogen lines, strong ionized calcium lines, strong sodium lines, many lines of neutral metals	Arcturus, Aldebaran
M	Red	2400–3700	Strong lines of neutral metals and molecular bands of titanium oxide dominate	Betelgeuse, Antares
L	Red	1300–2400	Metal hydride lines, alkali metal lines (e.g., sodium, potassium, rubidium)	Teide 1
T	Magenta	700–1300	Methane lines	Gliese 229B

Spectral Class	Color	Approximate Temperature (K)	Principal Features	Examples
Y	Infrared[1]	< 700	Ammonia lines	WISE 1828+2650

To see how spectral classification works, let's use Figure 17.5. Suppose you have a spectrum in which the hydrogen lines are about half as strong as those seen in an A star. Looking at the lines in our figure, you see that the star could be either a B star or a G star. But if the spectrum also contains helium lines, then it is a B star, whereas if it contains lines of ionized iron and other metals, it must be a G star.

If you look at Figure 17.6, you can see that you, too, could assign a spectral class to a star whose type was not already known. All you have to do is match the pattern of spectral lines to a standard star (like the ones shown in the figure) whose type has already been determined.

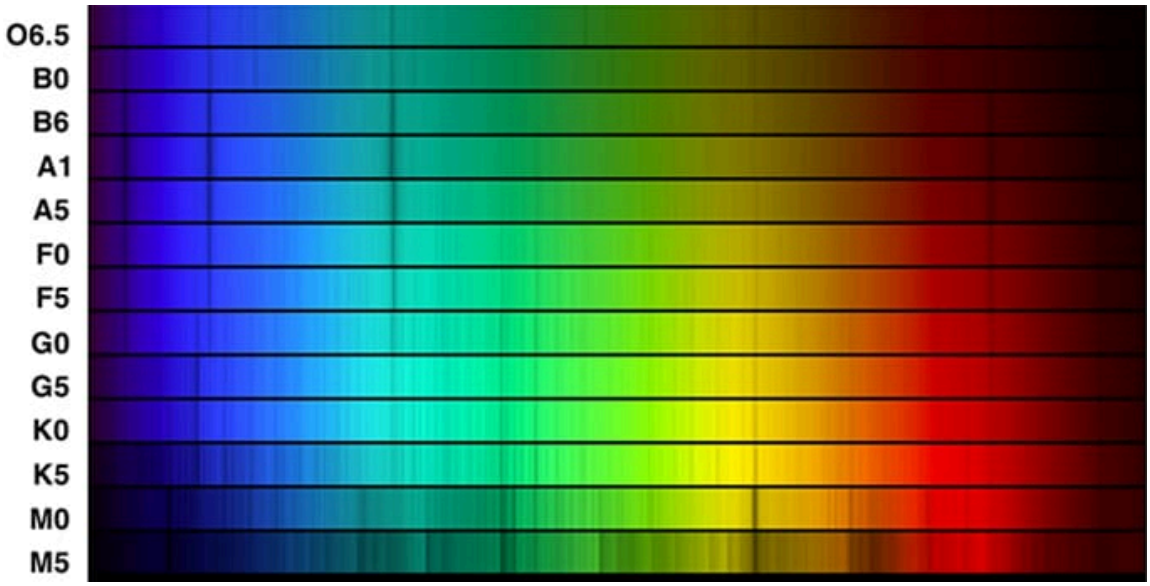


Figure 17.6 : Spectra of Stars with Different Spectral Classes. This image compares the spectra of the different spectral classes. The spectral class assigned to each of these stellar spectra is listed at the left of the picture. The strongest four lines seen at spectral type A1 (one in the red, one in the blue-green, and two in the blue) are Balmer lines of hydrogen. Note how these lines weaken at both higher and lower temperatures, as Figure 17.5 also indicates. The strong pair of closely spaced lines in the yellow in the cool stars is due to neutral sodium (one of the neutral metals in Figure 17.5). (Credit: modification of work by NOAO/AURA/NSF)

Both colors and spectral classes can be used to estimate the temperature of a star. Spectra are harder to measure because the light has to be bright enough to be spread out into all colors of the rainbow, and detectors must be sensitive enough to respond to individual wavelengths. In order to measure colors, the detectors need only respond to the many wavelengths that pass simultaneously through the colored filters that have been chosen—that is, to *all* the blue light or *all* the yellow-green light.

Voyagers in Astronomy: Annie Cannon: Classifier of the Stars

Annie Jump Cannon was born in Delaware in 1863 (Figure 17.7). In 1880, she went to Wellesley College, one of the new breed of US colleges opening up to educate young women. Wellesley, only 5 years old at the time, had the second student physics lab in the country and provided excellent training in basic science. After college, Cannon spent a decade with her parents but was very dissatisfied, longing to do scientific work. After her mother's death in 1893, she returned to Wellesley as a teaching assistant and also to take courses at Radcliffe, the women's college associated with Harvard.


Figure 17.7 : Annie Jump Cannon (1863–1941). Cannon is well-known for her classifications of stellar spectra. (credit: modification of work by Smithsonian Institution)

In the late 1800s, the director of the Harvard Observatory, Edward C. Pickering, needed lots of help with his ambitious program of classifying stellar spectra. The basis for these studies was a monumental collection of nearly a million photographic spectra of stars, obtained from many years of observations made at Harvard College Observatory in Massachusetts as well as at its remote observing stations in South America and South Africa. Pickering quickly discovered that educated young women could be hired as assistants for one-third or one-fourth the salary paid to men, and they would often put up with working conditions and repetitive tasks that men with the same education would not tolerate. These women became known as the Harvard Computers. (We should emphasize that astronomers were not alone in reaching such conclusions about the relatively new idea of upper-class, educated women working outside the home: women were exploited and undervalued in many fields. This is a legacy from which our society is just beginning to emerge.)

Cannon was hired by Pickering as one of the "computers" to help with the classification of spectra. She became so good at it that she could visually examine and determine the spectral types of several hundred stars per hour (dictating her conclusions to an assistant). She made many discoveries while investigating the Harvard photographic plates, including 300 variable stars (stars whose luminosity changes periodically). But her main legacy is a marvelous catalog of spectral types for hundreds of thousands of stars, which served as a foundation for much of twentieth-century astronomy.

In 1911, a visiting committee of astronomers reported that "she is the one person in the world who can do this work quickly and accurately" and urged Harvard to give Cannon an official appointment in keeping with her skill and renown. Not until 1938, however, did Harvard appoint her an astronomer at the university; she was then 75 years old.

Cannon received the first honorary degree Oxford awarded to a woman, and she became the first woman to be elected an officer of the American Astronomical Society, the main professional organization of astronomers in the US. She generously donated the money from one of the major prizes she had won to found a special award for women in astronomy, now known as the Annie Jump Cannon Prize. True to form, she continued classifying stellar spectra almost to the very end of her life in 1941.

Spectral Classes L, T, and Y

The scheme devised by Cannon worked well until 1988, when astronomers began to discover objects even cooler than M9-type stars. We use the word *object* because many of the new discoveries are not true stars. A star is defined as an object that during some part of its lifetime derives 100% of its energy from the same process that makes the Sun shine—the fusion of hydrogen nuclei (protons) into helium. Objects with masses less than about 7.5% of the mass of our Sun (about 0.075 M_{Sun}) do not become hot enough for hydrogen fusion to take place. Even before the first such "failed star" was found, this class of objects, with masses intermediate between stars and planets, was given the name brown dwarfs.

Brown dwarfs are very difficult to observe because they are extremely faint and cool, and they put out most of their light in the infrared part of the spectrum. It was only after the construction of very large telescopes, like the Keck telescopes in Hawaii, and the development of very sensitive infrared detectors, that the search for brown dwarfs succeeded. The first brown dwarf was discovered in 1988, and, as of the summer of 2015, there are more than 2200 known brown dwarfs.

Initially, brown dwarfs were given spectral classes like $M10^{+}$ or "much cooler than M9," but so many are now known that it is possible to begin assigning spectral types. The hottest brown dwarfs are given types L0–L9 (temperatures in the range

2400–1300 K), whereas still cooler (1300–700 K) objects are given types T0–T9 (see Figure 17.8). In class L brown dwarfs, the lines of titanium oxide, which are strong in M stars, have disappeared. This is because the L dwarfs are so cool that atoms and molecules can gather together into dust particles in their atmospheres; the titanium is locked up in the dust grains rather than being available to form molecules of titanium oxide. Lines of steam (hot water vapor) are present, along with lines of carbon monoxide and neutral sodium, potassium, cesium, and rubidium. Methane (CH_4) lines are strong in class-T brown dwarfs, as methane exists in the atmosphere of the giant planets in our own solar system.

In 2009, astronomers discovered ultra-cool brown dwarfs with temperatures of 500–600 K. These objects exhibited absorption lines due to ammonia (NH_3), which are not seen in T dwarfs. A new spectral class, Y, was created for these objects. As of 2021, over 50 brown dwarfs belonging to spectral class Y have been discovered, some with temperatures comparable to that of the human body (about 300 K).[2]

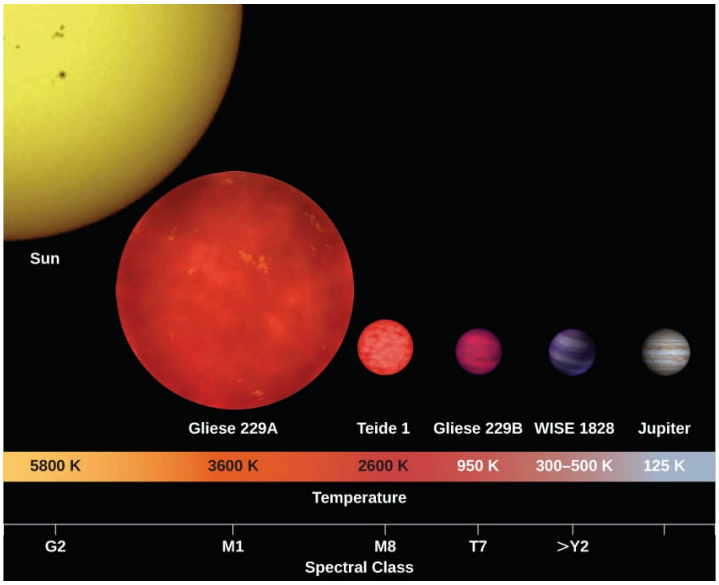


Figure 17.8 : Brown Dwarfs. This illustration shows the sizes and surface temperatures of brown dwarfs Teide 1, Gliese 229B, and WISE1828 in relation to the Sun, a red dwarf star (Gliese 229A), and Jupiter. (credit: modification of work by MPIA/V. Joergens)

Most brown dwarfs start out with atmospheric temperatures and spectra like those of true stars with spectral classes of M6.5 and later, even though the brown dwarfs are not hot and dense enough in their interiors to fuse hydrogen. In fact, the spectra of brown dwarfs and true stars are so similar from spectral types late M through L that it is not possible to distinguish the two types of objects based on spectra alone. An independent measure of mass is required to determine whether a specific object is a brown dwarf or a very low mass star. Since brown dwarfs cool steadily throughout their lifetimes, the spectral type of a given brown dwarf changes with time over a billion years or more from late M through L, T, and Y spectral types.

Low-Mass Brown Dwarfs vs. High-Mass Planets

An interesting property of brown dwarfs is that they are all about the same radius as Jupiter, regardless of their masses. Amazingly, this covers a range of masses from about 13 to 80 times the mass of Jupiter (M_J). This can make distinguishing a low-mass brown dwarf from a high-mass planet very difficult.

So, what is the difference between a low-mass brown dwarf and a high-mass planet? The International Astronomical Union considers the distinctive feature to be *deuterium fusion*. Although brown dwarfs do not sustain regular (proton-proton) hydrogen fusion, they are capable of fusing deuterium (a rare form of hydrogen with one proton and one neutron in its nucleus). The fusion of deuterium can happen at a lower temperature than the fusion of hydrogen. If an object has enough mass to fuse deuterium (about 13 M_J or 0.012 M_{Sun}), it is a brown dwarf. Objects with less than 13 M_J do not fuse deuterium and are usually considered planets.

Footnotes

- 1Absorption by sodium and potassium atoms makes Y dwarfs appear a bit less red than L dwarfs.
- 2The record-holder for the coldest Type Y dwarf is WISE 0855-0714, at around 250 K.

This page titled 5.2.3: The Spectra of Stars (and Brown Dwarfs) is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by OpenStax via source content that was edited to the style and standards of the LibreTexts platform.

5.2.4: Using Spectra to Measure Stellar Radius, Composition, and Motion

Learning Objectives

By the end of this section, you will be able to:

- Understand how astronomers can learn about a star's radius and composition by studying its spectrum
- Explain how astronomers can measure the motion and rotation of a star using the Doppler effect
- Describe the proper motion of a star and how it relates to a star's space velocity

Analyzing the spectrum of a star can teach us all kinds of things in addition to its temperature. We can measure its detailed chemical composition as well as the pressure in its atmosphere. From the pressure, we get clues about its size. We can also measure its motion toward or away from us and estimate its rotation.

Clues to the Size of a Star

As we shall see in The Stars: A Celestial Census, stars come in a wide variety of sizes. At some periods in their lives, stars can expand to enormous dimensions. Stars of such exaggerated size are called giants. Luckily for the astronomer, stellar spectra can be used to distinguish giants from run-of-the-mill stars (such as our Sun).

Suppose you want to determine whether a star is a giant. A giant star has a large, extended photosphere. Because it is so large, a giant star's atoms are spread over a great volume, which means that the density of particles in the star's photosphere is low. As a result, the pressure in a giant star's photosphere is also low. This low pressure affects the spectrum in two ways. First, a star with a lower-pressure photosphere shows narrower spectral lines than a star of the same temperature with a higher-pressure photosphere (Figure 17.9). The difference is large enough that careful study of spectra can tell which of two stars at the same temperature has a higher pressure (and is thus more compressed) and which has a lower pressure (and thus must be extended). This effect is due to collisions between particles in the star's photosphere—more collisions lead to broader spectral lines. Collisions will, of course, be more frequent in a higher-density environment. Think about it like traffic—collisions are much more likely during rush hour, when the density of cars is high.

Second, more atoms are ionized in a giant star than in a star like the Sun with the same temperature. The ionization of atoms in a star's outer layers is caused mainly by photons, and the amount of energy carried by photons is determined by temperature. But how long atoms *stay* ionized depends in part on pressure. Compared with what happens in the Sun (with its relatively dense photosphere), ionized atoms in a giant star's photosphere are less likely to pass close enough to electrons to interact and combine with one or more of them, thereby becoming neutral again. Ionized atoms, as we discussed earlier, have different spectra from atoms that are neutral.

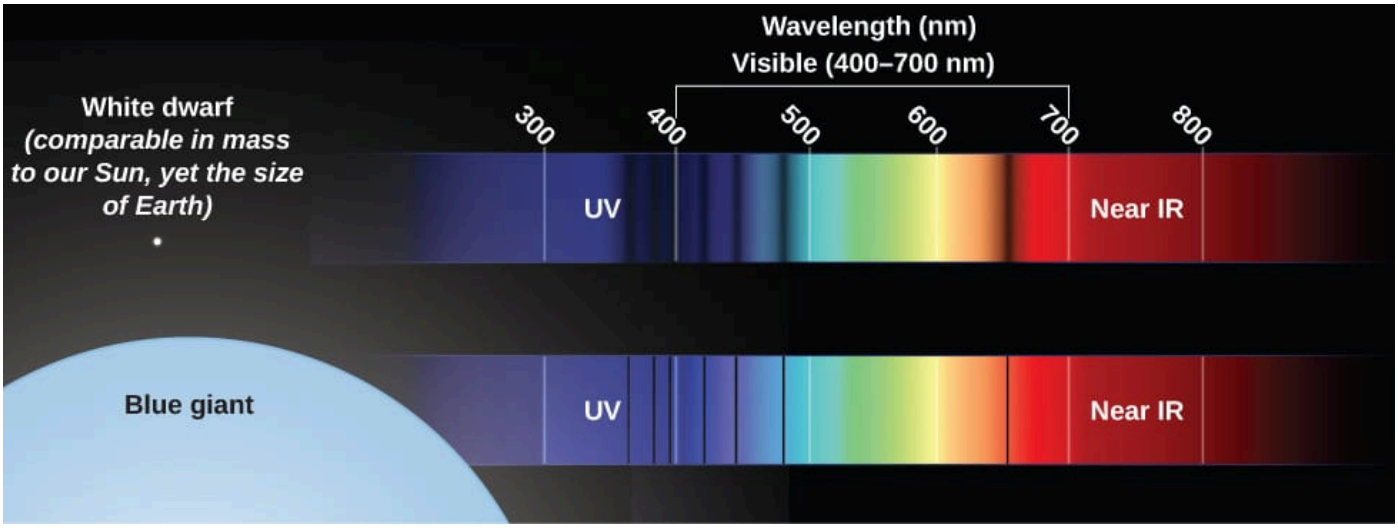


Figure 17.9 : Spectral Lines. This figure illustrates one difference in the spectral lines from stars of the same temperature but different pressures. A giant star with a very-low-pressure photosphere shows very narrow spectral lines (bottom), whereas a smaller star with a higher-pressure photosphere shows much broader spectral lines (top). (credit: modification of work by NASA, ESA, A. Field, and J. Kalirai (STScI))

Abundances of the Elements

Absorption lines of a majority of the known chemical elements have now been identified in the spectra of the Sun and stars. If we see lines of iron in a star's spectrum, for example, then we know immediately that the star must contain iron.

Note that the *absence* of an element's spectral lines does not necessarily mean that the element itself is absent. As we saw, the temperature and pressure in a star's atmosphere will determine what types of atoms are able to produce absorption lines. Only if the physical conditions in a star's photosphere are such that lines of an element *should* (according to calculations) be there can we conclude that the absence of observable spectral lines implies low abundance of the element.

Suppose two stars have identical temperatures and pressures, but the lines of, say, sodium are stronger in one than in the other. Stronger lines mean that there are more atoms in the stellar photosphere absorbing light. Therefore, we know immediately that the star with stronger sodium lines contains more sodium. Complex calculations are required to determine exactly how much more, but those calculations can be done for any element observed in any star with any temperature and pressure.

Of course, astronomy textbooks such as ours always make these things sound a bit easier than they really are. If you look at the stellar spectra such as those in Figure 17.6, you may get some feeling for how hard it is to decode all of the information contained in the thousands of absorption lines. First of all, it has taken many years of careful laboratory work on Earth to determine the precise wavelengths at which hot gases of each element have their spectral lines. Long books and computer databases have been compiled to show the lines of each element that can be seen at each temperature. Second, stellar spectra usually have many lines from a number of elements, and we must be careful to sort them out correctly. Sometimes nature is unhelpful, and lines of different elements have identical wavelengths, thereby adding to the confusion. And third, as we saw in the chapter on Radiation and Spectra, the motion of the star can change the observed wavelength of each of the lines. So, the observed wavelengths may not match laboratory measurements exactly. In practice, analyzing stellar spectra is a demanding, sometimes frustrating task that requires both training and skill.

Studies of stellar spectra have shown that hydrogen makes up about three-quarters of the mass of most stars. Helium is the second-most abundant element, making up almost a quarter of a star's mass. Together, hydrogen and helium make up from 96 to 99% of the mass; in some stars, they amount to more than 99.9%. Among the 4% or less of "heavy elements," oxygen, carbon, neon, iron, nitrogen, silicon, magnesium, and sulfur are among the most abundant. Generally, but not invariably, the elements of lower atomic weight are more abundant than those of higher atomic weight.

Take a careful look at the list of elements in the preceding paragraph. Two of the most abundant are hydrogen and oxygen (which make up water); add carbon and nitrogen and you are starting to write the prescription for the chemistry of an astronomy student. We are made of elements that are common in the universe—just mixed together in a far more sophisticated form (and a much cooler environment) than in a star.

As we mentioned in The Spectra of Stars (and Brown Dwarfs) section, astronomers use the term “metals” to refer to all elements heavier than hydrogen and helium. The fraction of a star’s mass that is composed of these elements is referred to as the star’s *metallicity*. The metallicity of the Sun, for example, is 0.02, since 2% of the Sun’s mass is made of elements heavier than helium.

Appendix K lists how common each element is in the universe (compared to hydrogen); these estimates are based primarily on investigation of the Sun, which is a typical star. Some very rare elements, however, have not been detected in the Sun. Estimates of the amounts of these elements in the universe are based on laboratory measurements of their abundance in primitive meteorites, which are considered representative of unaltered material condensed from the solar nebula (see the Cosmic Samples and the Origin of the Solar System chapter).

Radial Velocity

When we measure the spectrum of a star, we determine the wavelength of each of its lines. If the star is not moving with respect to the Sun, then the wavelength corresponding to each element will be the same as those we measure in a laboratory here on Earth. But if stars are moving toward or away from us, we must consider the *Doppler effect* (see The Doppler Effect section). We should see all the spectral lines of moving stars shifted toward the red end of the spectrum if the star is moving away from us, or toward the blue (violet) end if it is moving toward us (Figure 17.10). The greater the shift, the faster the star is moving. Such motion, along the line of sight between the star and the observer, is called radial velocity and is usually measured in kilometers per second.

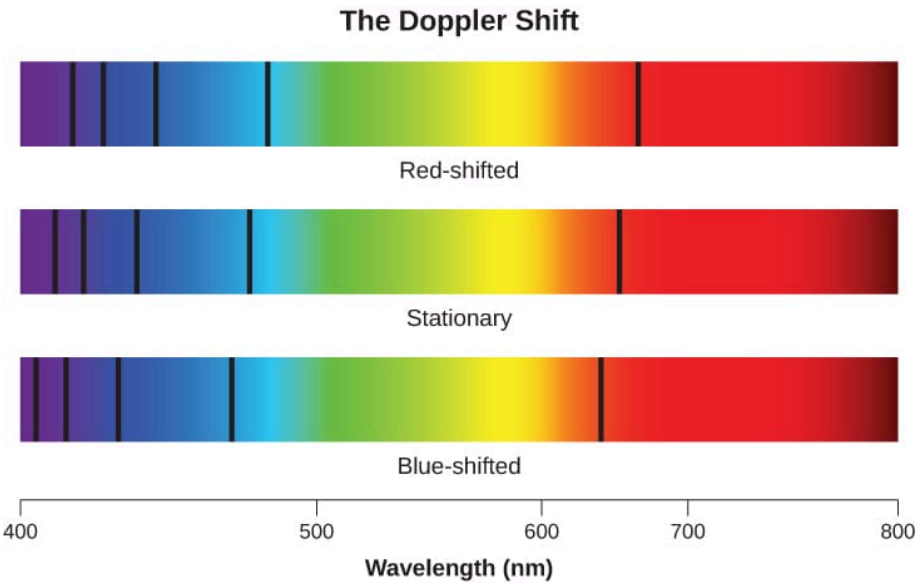


Figure 17.10 : Doppler-Shifted Stars. When the spectral lines of a moving star shift toward the red end of the spectrum, we know that the star is moving away from us. If they shift toward the blue end, the star is moving toward us.

William Huggins, pioneering yet again, in 1868 made the first radial velocity determination of a star. He observed the Doppler shift in one of the hydrogen lines in the spectrum of Sirius and found that this star is moving toward the solar system. (Better measurements now indicate that Sirius is actually moving away, but Huggins is still celebrated as a pioneer in making such measurements.) Today, radial velocity can be measured for any star bright enough for its spectrum to be observed. As we will see in The Stars: A Celestial Census, radial velocity measurements of double stars are crucial in deriving stellar masses.

Link to Learning

In the Spectrum Constructor, you can add absorption lines to the spectrum. Then adjust the velocity; negative indicates the star is moving toward the observer, and positive indicates it is moving away. Explore how the absorption lines are seen to shift to shorter and longer wavelengths.

Proper Motion

There is another type of motion stars can have that cannot be detected with stellar spectra. Unlike radial motion, which is along our line of sight (i.e., toward or away from Earth), this motion, called proper motion, is *transverse*: that is, across our line of sight. We see it as a change in the relative positions of the stars on the celestial sphere (Figure 17.11). These changes are very slow. Even the star with the largest proper motion takes 200 years to change its position in the sky by an amount equal to the width of the full Moon, and the motions of other stars are smaller yet.

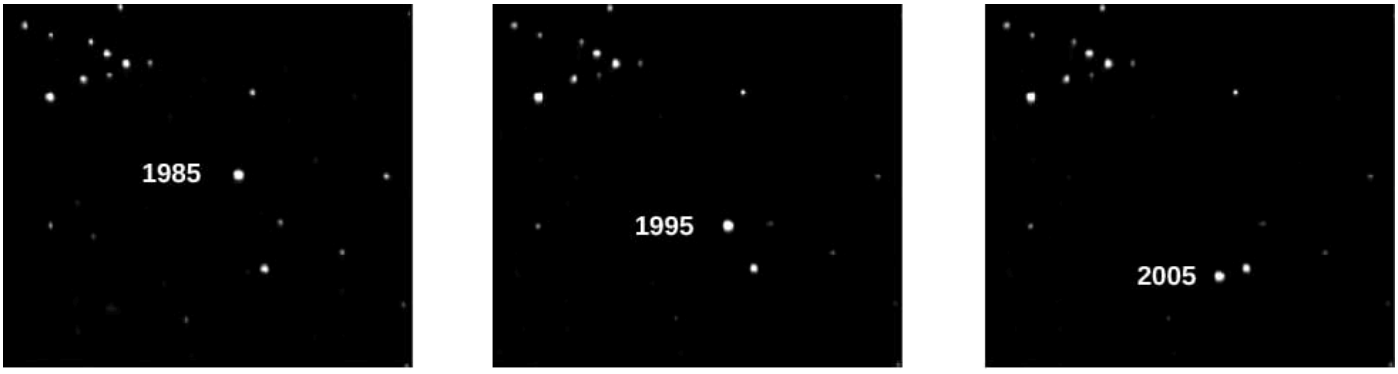


Figure 17.11 : Large Proper Motion. Three photographs of Barnard's star, the star with the largest known proper motion, show how this faint star has moved over a period of 20 years. (modification of work by Steve Quirk)

For this reason, with our naked eyes, we do not notice any change in the positions of the bright stars during the course of a human lifetime. If we could live long enough, however, the changes would become obvious. For example, some 50,000 years from now, terrestrial observers will find the handle of the Big Dipper unmistakably more bent than it is now (Figure 17.12).

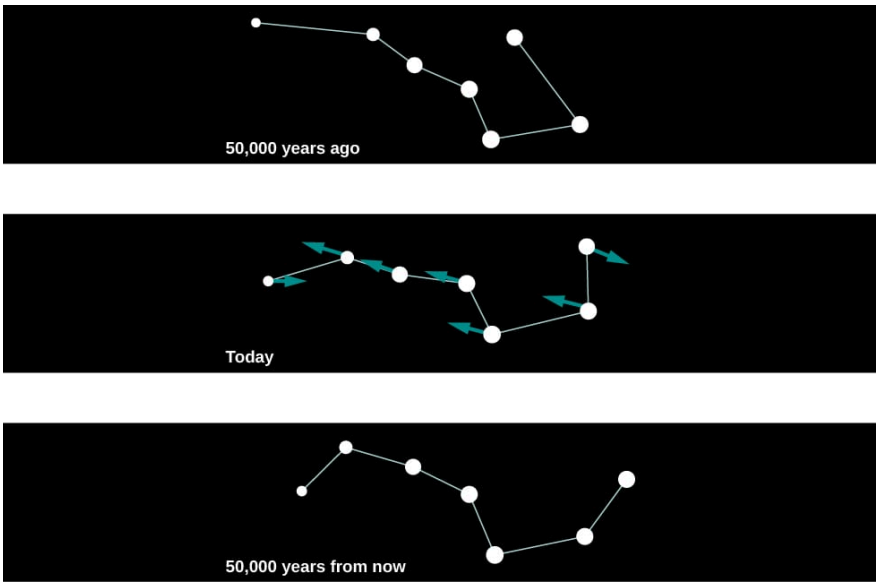


Figure 17.12 : Changes in the Big Dipper. This figure shows changes in the appearance of the Big Dipper due to proper motion of the stars over 100,000 years.

We measure the proper motion of a star in arcseconds (1/3600 of a degree) per year. That is, the measurement of proper motion tells us only by how much of an angle a star has changed its position on the celestial sphere. If two stars at different distances are moving at the same velocity perpendicular to our line of sight, the closer one will show a larger shift in its position on the celestial sphere in a year's time. As an analogy, imagine you are standing at the side of a freeway. Cars will appear to whiz past you. If you then watch the traffic from a vantage point half a mile away, the cars will move much more slowly across your field of vision. In order to convert this angular motion to a velocity, we need to know how far away the star is.

To know the true space velocity of a star—that is, its total speed and the direction in which it is moving through space relative to

the Sun—we must know its radial velocity, proper motion, and distance (Figure 17.13). A star's space velocity can also, over time, cause its distance from the Sun to change significantly. Over several hundred thousand years, these changes can be large enough to affect the apparent brightnesses of nearby stars. Today, Sirius, in the constellation Canis Major (the Big Dog) is the brightest star in the sky, but 100,000 years ago, the star Canopus in the constellation Carina (the Keel) was the brightest one. A little over 200,000 years from now, Sirius will have moved away and faded somewhat, and Vega, the bright blue star in Lyra, will take over its place of honor as the brightest star in Earth's skies.

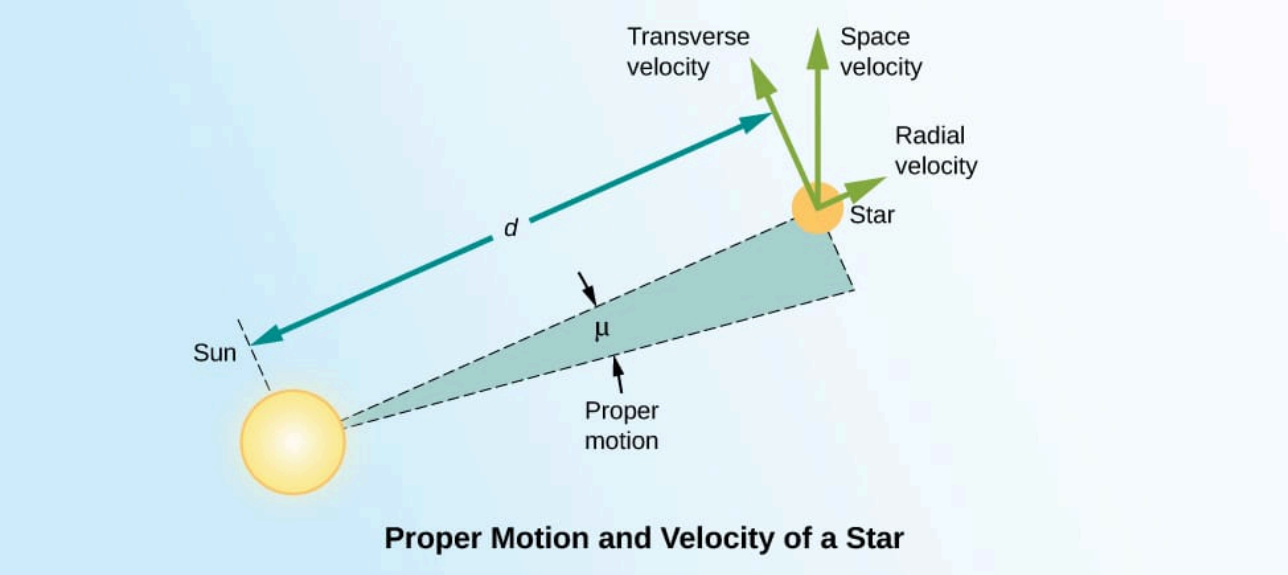


Figure 17.13 : Space Velocity and Proper Motion. This figure shows the true space velocity of a star. The radial velocity is the component of the space velocity projected along the line of sight from the Sun to a star. The transverse velocity is a component of the space velocity projected on the sky. What astronomers measure is proper motion (μ), which is the change in the apparent direction on the sky measured in fractions of a degree. To convert this change in direction to a speed in, say, kilometers per second, it is necessary to also know the distance (*d*) from the Sun to the star.

Rotation

We can also use the Doppler effect to measure how fast a star rotates. If an object is rotating, then one of its sides is approaching us while the other is receding (unless its axis of rotation happens to be pointed exactly toward us). This is clearly the case for the Sun or a planet; we can observe the light from either the approaching or receding edge of these nearby objects and directly measure the Doppler shifts that arise from the rotation.

Stars, however, are so far away that they all appear as unresolved points. The best we can do is to analyze the light from the entire star at once. Due to the Doppler effect, the lines in the light that come from the side of the star rotating toward us are shifted to shorter wavelengths and the lines in the light from the opposite edge of the star are shifted to longer wavelengths. You can think of each spectral line that we observe as the sum or composite of spectral lines originating from different speeds with respect to us. Each point on the star has its own Doppler shift, so the absorption line we see from the whole star is actually much wider than it would be if the star were not rotating. If a star is rotating rapidly, there will be a greater spread of Doppler shifts and all its spectral lines should be quite broad. In fact, astronomers call this effect *line broadening*, and the amount of broadening can tell us the speed at which the star rotates (Figure 17.14).

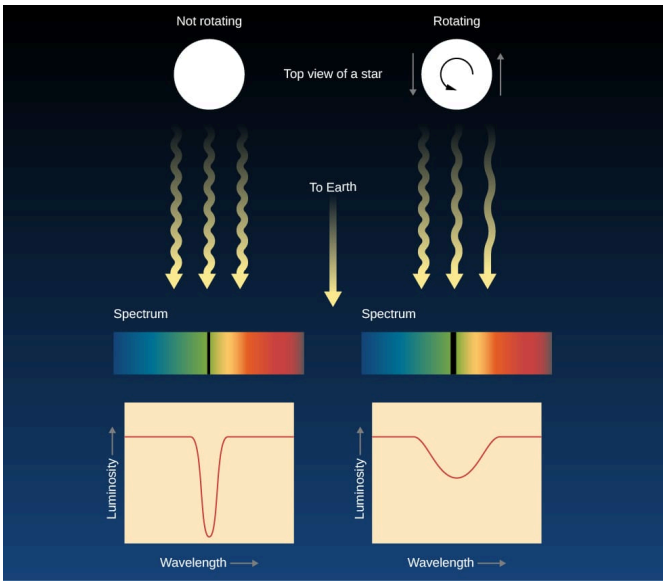


Figure 17.14 : Using a Spectrum to Determine Stellar Rotation. A rotating star will show broader spectral lines than a nonrotating star.

Measurements of the widths of spectral lines show that many stars rotate faster than the Sun, some with periods of less than a day! These rapid rotators spin so fast that their shapes are “flattened” into what we call *oblate spheroids*. An example of this is the star Vega, which rotates once every 12.5 hours. Vega’s rotation flattens its shape so much that its diameter at the equator is 23% wider than its diameter at the poles (Figure 17.15). The Sun, with its rotation period of about a month, rotates rather slowly. Studies have shown that stars decrease their rotational speed as they age. Young stars rotate very quickly, with rotational periods of days or less. Very old stars can have rotation periods of several months.

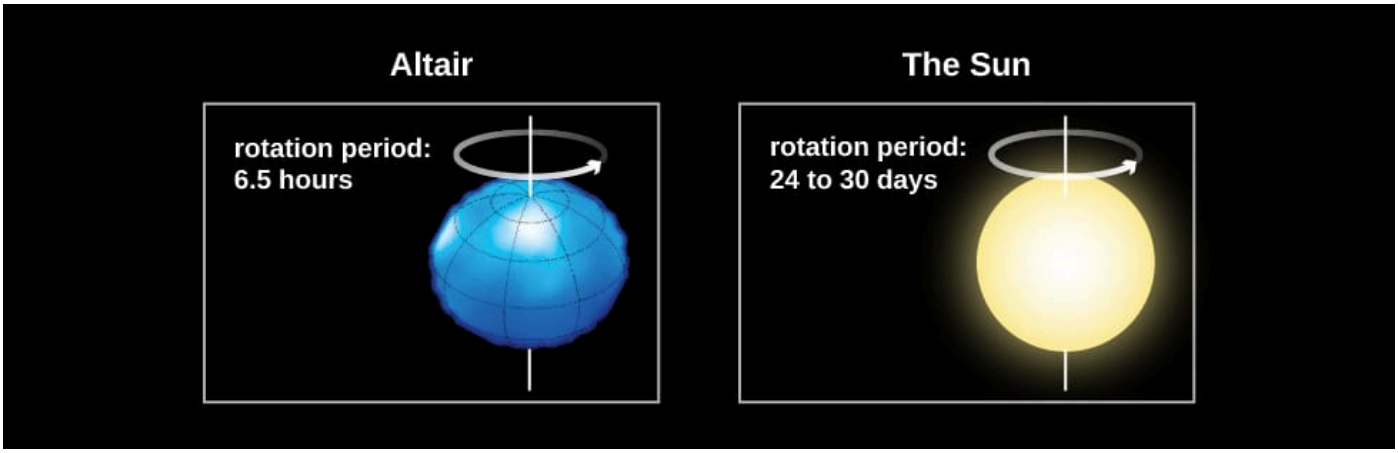


Figure 17.15 : Comparison of Rotating Stars. This illustration compares the more rapidly rotating star Altair to the slower rotating Sun.

As you can see, spectroscopy is an extremely powerful technique that helps us learn all kinds of information about stars that we simply could not gather any other way. We will see in later chapters that these same techniques can also teach us about galaxies, which are the most distant objects that can we observe. Without spectroscopy, we would know next to nothing about the universe beyond the solar system.

Making Connections: Astronomy and Philanthropy

Throughout the history of astronomy, contributions from wealthy patrons of the science have made an enormous difference in building new instruments and carrying out long-term research projects. Edward Pickering’s stellar classification project, which was to stretch over several decades, was made possible by major donations from Anna Draper. She was the widow of Henry Draper, a physician who was one of the most accomplished amateur astronomers of the nineteenth century and the

first person to successfully photograph the spectrum of a star. Anna Draper gave several hundred thousand dollars to Harvard Observatory. As a result, the great spectroscopic survey is still known as the Henry Draper Memorial, and many stars are still referred to by their “HD” numbers in that catalog (such as HD 209458).

In the 1870s, the eccentric piano builder and real estate magnate James Lick (Figure 17.16) decided to leave some of his fortune to build the world’s largest telescope. When, in 1887, the pier to house the telescope was finished, Lick’s body was entombed in it. Atop the foundation rose a 36-inch refractor, which for many years was the main instrument at the Lick Observatory near San Jose.

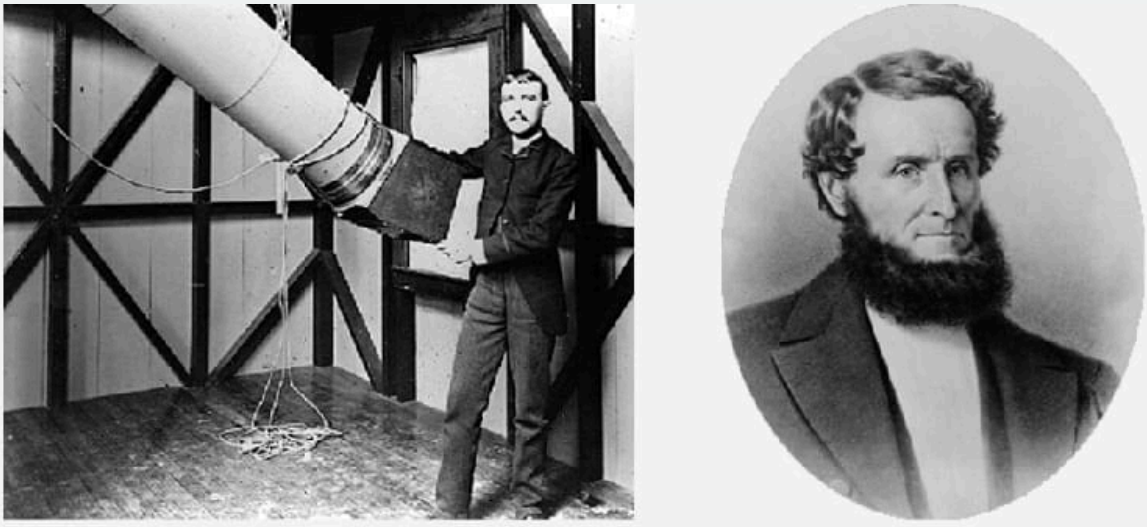


Figure 17.16 : Henry Draper (1837–1882) and James Lick (1796–1876). (a) Draper stands next to a telescope used for photography. After his death, his widow funded further astronomy work in his name. (b) Lick was a philanthropist who provided funds to build a 36-inch refractor not only as a memorial to himself but also to aid in further astronomical research.

The Lick telescope remained the largest in the world until 1897, when George Ellery Hale persuaded railroad millionaire Charles Yerkes to finance the construction of a 40-inch telescope near Chicago. More recently, Howard Keck, whose family made its fortune in the oil industry, gave $70 million from his family foundation to the California Institute of Technology to help build the world’s largest telescope atop the 14,000-foot peak of Maunakea in Hawaii (see the chapter on Astronomical Instruments to learn more about these telescopes). The Keck Foundation was so pleased with what is now called the Keck telescope that they gave $74 million more to build Keck II, another 10-meter reflector on the same volcanic peak.

Now, if any of you become millionaires or billionaires, and astronomy has sparked your interest, do keep an astronomical instrument or project in mind as you plan your estate. But frankly, private philanthropy could not possibly support the full enterprise of scientific research in astronomy. Much of our exploration of the universe is financed by federal agencies such as the National Science Foundation and NASA in the United States, and by similar government agencies in the other countries. In this way, all of us, through a very small share of our tax dollars, are philanthropists for astronomy.

This page titled 5.2.4: Using Spectra to Measure Stellar Radius, Composition, and Motion is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by OpenStax via source content that was edited to the style and standards of the LibreTexts platform.

5.2.5: The H-R Diagram

Learning Objectives

By the end of this section, you will be able to:

- Identify the physical characteristics of stars that are used to create an H–R diagram, and describe how those characteristics vary among groups of stars
- Discuss the physical properties of most stars found at different locations on the H–R diagram, such as radius, and for main sequence stars, mass

In this chapter and Analyzing Starlight, we described some of the characteristics by which we might classify stars and how those are measured. These ideas are summarized in Table 18.2. We have also given an example of a relationship between two of these characteristics in the mass-luminosity relation. When the characteristics of large numbers of stars were measured at the beginning of the twentieth century, astronomers were able to begin a deeper search for patterns and relationships in these data.

Table 18.2: Measuring the Characteristics of Stars

Characteristic	Technique
Surface temperature	1. Determine the color (very rough). 2. Measure the spectrum and get the spectral type.
Chemical composition	Determine which lines are present in the spectrum.
Luminosity	Measure the apparent brightness and compensate for distance.
Radial velocity	Measure the Doppler shift in the spectrum.
Rotation	Measure the width of spectral lines.
Mass	Measure the period and radial velocity curves of spectroscopic binary stars.
Diameter	1. Measure the way a star's light is blocked by the Moon. 2. Measure the light curves and Doppler shifts for eclipsing binary stars.

To help understand what sorts of relationships might be found, let's look briefly at a range of data about human beings. If you want to understand humans by comparing and contrasting their characteristics—without assuming any previous knowledge of these strange creatures—you could try to determine which characteristics lead you in a fruitful direction. For example, you might plot the heights of a large sample of humans against their weights (which is a measure of their mass). Such a plot is shown in Figure 18.12 and it has some interesting features. In the way we have chosen to present our data, height increases upward, whereas weight increases to the left. Notice that humans are not randomly distributed in the graph. Most points fall along a sequence that goes from the upper left to the lower right.

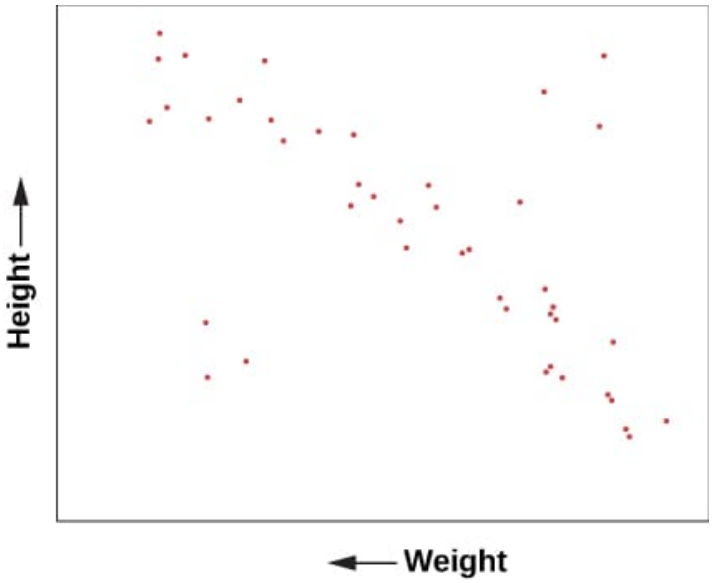


Figure 18.12 : Height versus Weight. The plot of the heights and weights of a representative group of human beings. Most points lie along a “main sequence” representing most people, but there are a few exceptions.

We can conclude from this graph that human height and weight are related. Generally speaking, taller human beings weigh more, whereas shorter ones weigh less. This makes sense if you are familiar with the structure of human beings. Typically, if we have bigger bones, we have more flesh to fill out our larger frame. It’s not mathematically exact—there is a wide range of variation—but it’s not a bad overall rule. And, of course, there are some dramatic exceptions. You occasionally see a short human who is very overweight and would thus be more to the bottom left of our diagram than the average sequence of people. Or you might have a very tall, skinny fashion model with great height but relatively small weight, who would be found near the upper right.

A similar diagram has been found extremely useful for understanding the lives of stars. In 1913, American astronomer Henry Norris Russell plotted the luminosities of stars against their spectral classes (a way of denoting their surface temperatures). This investigation, and a similar independent study in 1911 by Danish astronomer Ejnar Hertzsprung, led to the extremely important discovery that the temperature and luminosity of stars are related (Figure 18.13).

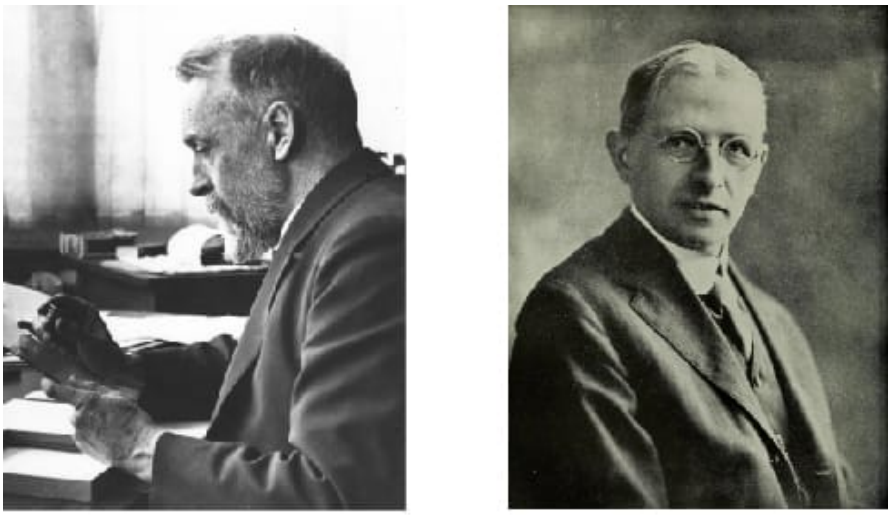


Figure 18.13 : Hertzsprung (1873–1967) and Russell (1877–1957). (a) Ejnar Hertzsprung and (b) Henry Norris Russell independently discovered the relationship between the luminosity and surface temperature of stars that is summarized in what is now called the H–R diagram.

Voyagers in Astronomy

Henry Norris Russell

When Henry Norris Russell graduated from Princeton University, his work had been so brilliant that the faculty decided to create a new level of honors degree beyond "summa cum laude" for him. His students later remembered him as a man whose thinking was three times faster than just about anybody else's. His memory was so phenomenal, he could correctly quote an enormous number of poems and limericks, the entire Bible, tables of mathematical functions, and almost anything he had learned about astronomy. He was nervous, active, competitive, critical, and very articulate; he tended to dominate every meeting he attended. In outward appearance, he was an old-fashioned product of the nineteenth century who wore high-top black shoes and high starched collars, and carried an umbrella every day of his life. His 264 papers were enormously influential in many areas of astronomy.

Born in 1877, the son of a Presbyterian minister, Russell showed early promise. When he was 12, his family sent him to live with an aunt in Princeton so he could attend a top preparatory school. He lived in the same house in that town until his death in 1957 (interrupted only by a brief stay in Europe for graduate work). He was fond of recounting that both his mother and his maternal grandmother had won prizes in mathematics, and that he probably inherited his talents in that field from their side of the family.

Before Russell, American astronomers devoted themselves mainly to surveying the stars and making impressive catalogs of their properties, especially their spectra (as described in Analyzing Starlight. Russell began to see that interpreting the spectra of stars required a much more sophisticated understanding of the physics of the atom, a subject that was being developed by European physicists in the 1910s and 1920s. Russell embarked on a lifelong quest to ascertain the physical conditions inside stars from the clues in their spectra; his work inspired, and was continued by, a generation of astronomers, many trained by Russell and his collaborators.

Russell also made important contributions in the study of binary stars and the measurement of star masses, the origin of the solar system, the atmospheres of planets, and the measurement of distances in astronomy, among other fields. He was an influential teacher and popularizer of astronomy, writing a column on astronomical topics for *Scientific American* magazine for more than 40 years. He and two colleagues wrote a textbook for college astronomy classes that helped train astronomers and astronomy enthusiasts over several decades. That book set the scene for the kind of textbook you are now reading, which not only lays out the facts of astronomy but also explains how they fit together. Russell gave lectures around the country, often emphasizing the importance of understanding modern physics in order to grasp what was happening in astronomy.

Harlow Shapley, director of the Harvard College Observatory, called Russell "the dean of American astronomers." Russell was certainly regarded as the leader of the field for many years and was consulted on many astronomical problems by colleagues from around the world. Today, one of the highest recognitions that an astronomer can receive is an award from the American Astronomical Society called the Russell Prize, set up in his memory.

Features of the H–R Diagram

Following Hertzsprung and Russell, let us plot the temperature (or spectral class) of a selected group of nearby stars against their luminosity and see what we find (Figure 18.14). Such a plot is frequently called the *Hertzsprung–Russell diagram*, abbreviated H–R diagram. It is one of the most important and widely used diagrams in astronomy, with applications that extend far beyond the purposes for which it was originally developed more than a century ago.

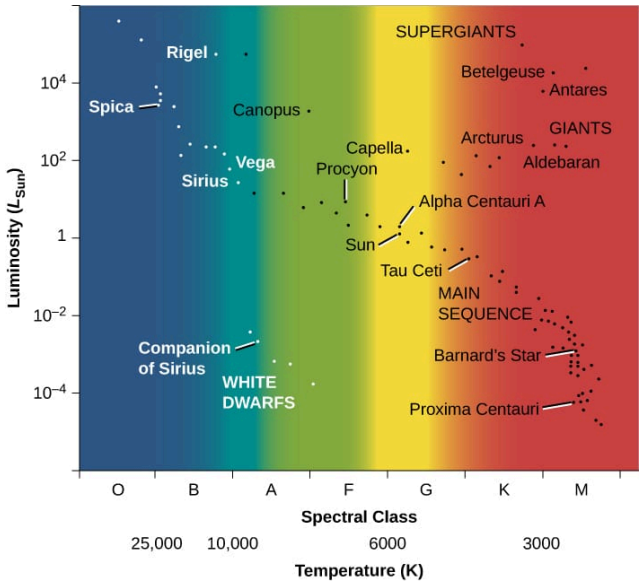


Figure 18.14 : H–R Diagram for a Selected Sample of Stars. In such diagrams, luminosity is plotted along the vertical axis. Along the horizontal axis, we can plot either temperature or spectral type (also sometimes called spectral class). Several of the brightest stars are identified by name. Most stars fall on the main sequence.

It is customary to plot H–R diagrams in such a way that temperature increases toward the left and luminosity toward the top. Notice the similarity to our plot of height and weight for people (Figure 18.12). Stars, like people, are not distributed over the diagram at random, as they would be if they exhibited all combinations of luminosity and temperature. Instead, we see that the stars cluster into certain parts of the H–R diagram. The great majority are aligned along a narrow sequence running from the upper left (hot, highly luminous) to the lower right (cool, less luminous). This band of points is called the main sequence. It represents a relationship between *temperature* and *luminosity* that is followed by most stars. We can summarize this relationship by saying that hotter stars are more luminous than cooler ones.

A number of stars, however, lie above the main sequence on the H–R diagram, in the upper-right region, where stars have low temperature and high luminosity. How can a star be at once cool, meaning each square meter on the star does not put out all that much energy, and yet very luminous? The only way is for the star to be enormous—to have so many square meters on its surface that the *total* energy output is still large. These stars must be *giants* or *supergiants*, the stars of huge diameter we discussed earlier.

There are also some stars in the lower-left corner of the diagram, which have high temperature and low luminosity. If they have high surface temperatures, each square meter on that star puts out a lot of energy. How then can the overall star be dim? It must be that it has a very small total surface area; such stars are known as white dwarfs (white because, at these high temperatures, the colors of the electromagnetic radiation that they emit blend together to make them look bluish-white). We will say more about these puzzling objects in a moment. Figure 18.15 is a schematic H–R diagram for a large sample of stars, drawn to make the different types more apparent.

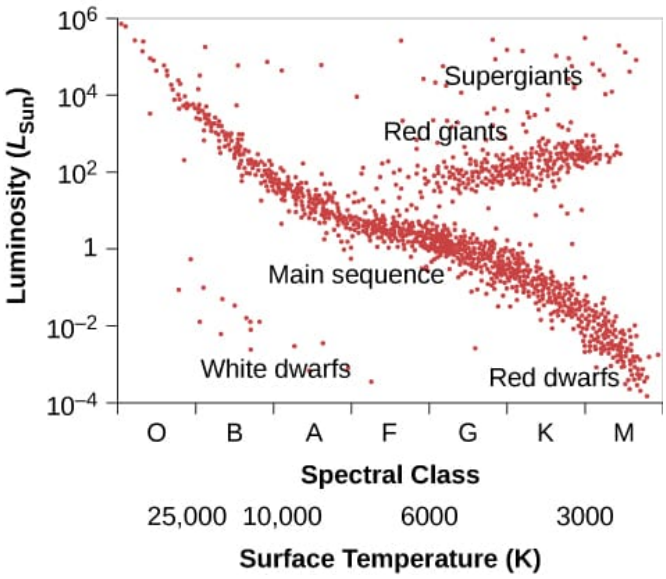


Figure 18.15 : Schematic H–R Diagram for Many Stars. Ninety percent of all stars on such a diagram fall along a narrow band called the main sequence. A minority of stars are found in the upper right; they are both cool (and hence red) and bright, and must be giants. Some stars fall in the lower left of the diagram; they are both hot and dim, and must be white dwarfs.

Now, think back to our discussion of star surveys. It is difficult to plot an H–R diagram that is truly representative of all stars because most stars are so faint that we cannot see those outside our immediate neighborhood. The stars plotted in Figure 18.14 were selected because their distances are known. This sample omits many intrinsically faint stars that are nearby but have not had their distances measured, so it shows fewer faint main-sequence stars than a "fair" diagram would. To be truly representative of the stellar population, an H–R diagram should be plotted for all stars within a certain distance. Unfortunately, our knowledge is reasonably complete only for stars within 10 to 20 light-years of the Sun, among which there are no giants or supergiants. Still, from many surveys (and more can now be done with new, more powerful telescopes), we estimate that about 90% of the true stars overall (excluding brown dwarfs) in our part of space are main-sequence stars, about 10% are white dwarfs, and fewer than 1% are giants or supergiants.

These estimates can be used directly to understand the lives of stars. Permit us another quick analogy with people. Suppose we survey people just like astronomers survey stars, but we want to focus our attention on the location of young people, ages 6 to 18 years. Survey teams fan out and take data about where such youngsters are found at all times during a 24-hour day. Some are found in the local pizza parlor, others are asleep at home, some are at the movies, and many are in school. After surveying a very large number of young people, one of the things that the teams determine is that, averaged over the course of the 24 hours, one-third of all youngsters are found in school.

How can they interpret this result? Does it mean that two-thirds of students are truants and the remaining one-third spend all their time in school? No, we must bear in mind that the survey teams counted youngsters throughout the full 24-hour day. Some survey teams worked at night, when most youngsters were at home asleep, and others worked in the late afternoon, when most youngsters were on their way home from school (and more likely to be enjoying a pizza). If the survey was truly representative, we *can* conclude, however, that if an average of one-third of all youngsters are found in school, then humans ages 6 to 18 years must spend about one-third of *their time* in school.

We can do something similar for stars. We find that, on average, 90% of all stars are located on the main sequence of the H–R diagram. If we can identify some activity or life stage with the main sequence, then it follows that stars must spend 90% of their lives in that activity or life stage.

Understanding the Main Sequence

In The Sun: A Nuclear Powerhouse, we discussed the Sun as a representative star. We saw that what stars such as the Sun "do for a living" is to convert protons into helium deep in their interiors via the process of nuclear fusion, thus producing energy. The fusion of protons to helium is an excellent, long-lasting source of energy for a star because the bulk of every star consists of hydrogen atoms, whose nuclei are protons.

Our computer models of how stars evolve over time show us that a typical star will spend about 90% of its life fusing the abundant hydrogen in its core into helium. This then is a good explanation of why 90% of all stars are found on the main sequence in the H–R diagram. But if all the stars on the main sequence are doing the same thing (fusing hydrogen), why are they distributed along a sequence of points? That is, why do they differ in luminosity and surface temperature (which is what we are plotting on the H–R diagram)?

To help us understand how main-sequence stars differ, we can use one of the most important results from our studies of model stars. Astrophysicists have been able to show that the structure of stars that are in equilibrium and derive all their energy from nuclear fusion is completely and uniquely determined by just two quantities: the *total mass* and the *composition* of the star. This fact provides an interpretation of many features of the H–R diagram.

Imagine a cluster of stars forming from a cloud of interstellar "raw material" whose chemical composition is similar to the Sun's. (We'll describe this process in more detail in The Birth of Stars and Discovery of Planets outside the Solar System, but for now, the details will not concern us.) In such a cloud, all the clumps of gas and dust that become stars begin with the same chemical composition and differ from one another only in mass. Now suppose that we compute a model of each of these stars for the time at which it becomes stable and derives its energy from nuclear reactions, but before it has time to alter its composition appreciably as a result of these reactions.

The models calculated for these stars allow us to determine their luminosities, temperatures, and sizes. If we plot the results from the models—one point for each model star—on the H–R diagram, we get something that looks just like the main sequence we saw for real stars.

And here is what we find when we do this. The model stars with the largest masses are the hottest and most luminous, and they are located at the upper left of the diagram.

The least-massive model stars are the coolest and least luminous, and they are placed at the lower right of the plot. The other model stars all lie along a line running diagonally across the diagram. In other words, *the main sequence turns out to be a sequence of stellar masses*.

This makes sense if you think about it. The most massive stars have the most gravity and can thus compress their centers to the greatest degree. This means they are the hottest inside and the best at generating energy from nuclear reactions deep within. As a result, they shine with the greatest luminosity and have the hottest surface temperatures. The stars with lowest mass, in turn, are the coolest inside and least effective in generating energy. Thus, they are the least luminous and wind up being the coolest on the surface. Our Sun lies somewhere in the middle of these extremes (as you can see in Figure 18.14). The characteristics of representative main-sequence stars (excluding brown dwarfs, which are not true stars) are listed in Table 18.3.

Table 18.3: Characteristics of Main-Sequence Stars

Spectral Type	Mass (Sun = 1)	Luminosity (Sun = 1)	Temperature	Radius (Sun = 1)
O5	40	7×1057×105	40,000 K	18
B0	16	2.7×1052.7×105	28,000 K	7
A0	3.3	55	10,000 K	2.5
F0	1.7	5	7500 K	1.4
G0	1.1	1.4	6000 K	1.1
K0	0.8	0.35	5000 K	0.8
M0	0.4	0.05	3500 K	0.6

Note that we've seen this 90% figure come up before. This is exactly what we found earlier when we examined the mass-luminosity relation (Figure 18.9). We observed that 90% of all stars seem to follow the relationship; these are the 90% of all stars that lie on the main sequence in our H–R diagram. Our models and our observations agree.

What about the other stars on the H–R diagram—the giants and supergiants, and the white dwarfs? As we will see in the next few chapters, these are what main-sequence stars turn into as they age: They are the later stages in a star's life. As a star consumes its nuclear fuel, its source of energy changes, as do its chemical composition and interior structure. These changes cause the star to alter its luminosity and surface temperature so that it no longer lies on the main sequence on our diagram. Because stars

spend much less time in these later stages of their lives, we see fewer stars in those regions of the H–R diagram.

Extremes of Stellar Luminosities, Diameters, and Densities

We can use the H–R diagram to explore the extremes in size, luminosity, and density found among the stars. Such extreme stars are not only interesting to fans of the *Guinness Book of World Records*; they can teach us a lot about how stars work. For example, we saw that the most massive main-sequence stars are the most luminous ones. We know of a few extreme stars that are a million times more luminous than the Sun, with masses that exceed 100 times the Sun's mass. These superluminous stars, which are at the upper left of the H–R diagram, are exceedingly hot, very blue stars of spectral type O. These are the stars that would be the most conspicuous at vast distances in space.

The cool supergiants in the upper corner of the H–R diagram are as much as 10,000 times as luminous as the Sun. In addition, these stars have diameters very much larger than that of the Sun. As discussed above, some supergiants are so large that if the solar system could be centered in one, the star's surface would lie beyond the orbit of Mars (see Figure 18.16). We will have to ask, in coming chapters, what process can make a star swell up to such an enormous size, and how long these "swollen" stars can last in their distended state.

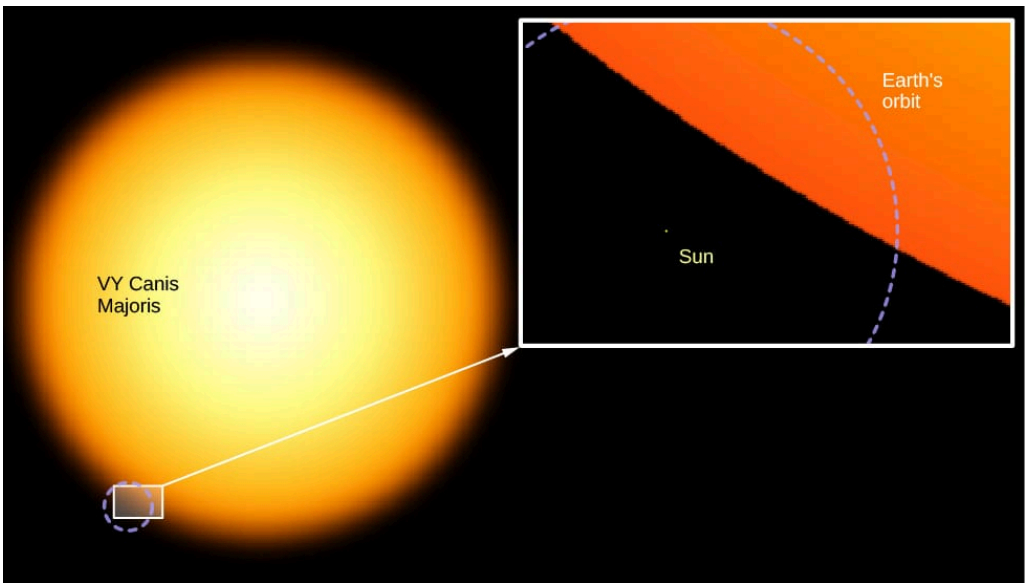


Figure 18.16 : The Sun and a Supergiant. Here you see how small the Sun looks in comparison to one of the largest known stars: VY Canis Majoris, a supergiant.

In contrast, the very common red, cool, low-luminosity stars at the lower end of the main sequence are much smaller and more compact than the Sun. An example of such a red dwarf is Ross 614B, with a surface temperature of 2700 K and only 1/2000 of the Sun's luminosity. We call such a star a dwarf because its diameter is only 1/10 that of the Sun. A star with such a low luminosity also has a low mass (about 1/12 that of the Sun). This combination of mass and diameter means that it is so compressed that the star has an average density about 80 times that of the Sun. Its density must be higher, in fact, than that of any known solid found on the surface of Earth. (Despite this, the star is made of gas throughout because its center is so hot.)

The faint, red, main-sequence stars are not the stars of the most extreme densities, however. The white dwarfs, at the lower-left corner of the H–R diagram, have densities many times greater still.

The White Dwarfs

The first white dwarf star was detected in 1862. Called Sirius B, it forms a binary system with Sirius A, the brightest-appearing star in the sky. It eluded discovery and analysis for a long time because its faint light tends to be lost in the glare of nearby Sirius A (Figure 18.17). (Since Sirius is often called the Dog Star—being the brightest star in the constellation of Canis Major, the big dog—Sirius B is sometimes nicknamed the Pup.)

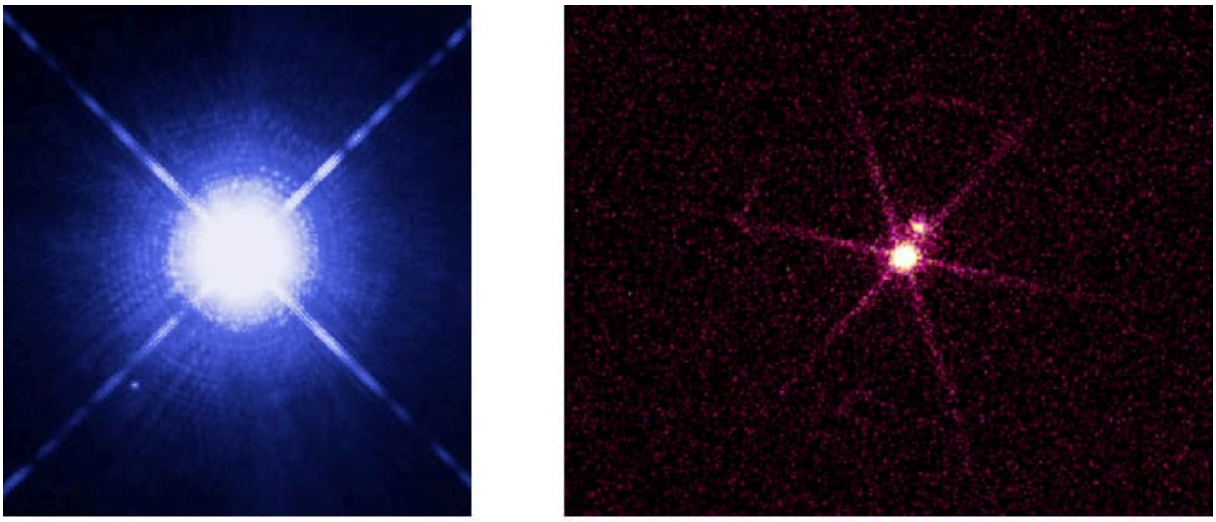


Figure 18.17 : Two Views of Sirius and Its White Dwarf Companion. (a) The (visible light) image, taken with the Hubble Space Telescope, shows bright Sirius A, and, below it and off to its left, faint Sirius B. (b) This image of the Sirius star system was taken with the Chandra X-Ray Telescope. Now, the bright object is the white dwarf companion, Sirius B. Sirius A is the faint object above it; what we are seeing from Sirius is probably not actually X-ray radiation but rather ultraviolet light that has leaked into the detector. Note that the ultraviolet intensities of these two objects are completely reversed from the situation in visible light because Sirius B is hotter and emits more higher-frequency radiation. (credit a: modification of work by NASA, H.E. Bond and E. Nelan (Space Telescope Science Institute), M. Barstow and M. Burleigh (University of Leicester) and J.B. Holberg (University of Arizona); credit b: modification of work by NASA/SAO/CXC)

We have now found thousands of white dwarfs. Table 18.1 shows that about 7% of the true stars (spectral types O–M) in our local neighborhood are white dwarfs. A good example of a typical white dwarf is the nearby star 40 Eridani B. Its surface temperature is a relatively hot 12,000 K, but its luminosity is only 1/275 L_{Sun}. Calculations show that its radius is only 1.4% of the Sun's, or about the same as that of Earth, and its volume is 2.5×10^{-6} that of the Sun. Its mass, however, is 0.57 times the Sun's mass, just a little more than half. To fit such a substantial mass into so tiny a volume, the star's density must be about 210,000 times the density of the Sun, or more than 300,000 g/cm^3. A teaspoonful of this material would have a mass of some 1.6 tons! At such enormous densities, matter cannot exist in its usual state; we will examine the particular behavior of this type of matter in The Death of Stars. For now, we just note that white dwarfs are dying stars, reaching the end of their productive lives and ready for their stories to be over.

The British astrophysicist (and science popularizer) Arthur Eddington (1882–1944) described the first known white dwarf this way:

> The message of the companion of Sirius, when decoded, ran: "I am composed of material three thousand times denser than anything you've ever come across. A ton of my material would be a little nugget you could put in a matchbox." What reply could one make to something like that? Well, the reply most of us made in 1914 was, "Shut up; don't talk nonsense."

Today, however, astronomers not only accept that stars as dense as white dwarfs exist but (as we will see) have found even denser and stranger objects in their quest to understand the evolution of different types of stars.

Link to Learning

Use this interactive H-R diagram to explore the sizes of stars based on their placement. Overlay the closest stars and the brightest stars, and consider where they lie on the figure. How many stars are among both the closest and brightest stars?

This page titled 5.2.5: The H-R Diagram is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by OpenStax via source content that was edited to the style and standards of the LibreTexts platform.

5.3: Stellar Evolution

This section follows the life cycle of stars from their formation in clouds of gas and dust through the final stages of their evolution. It examines how a star's position on the H-R diagram changes over time and emphasizes the importance of stellar mass in determining its evolutionary path. The section traces the development of low- and high-mass stars from the main sequence through red giant stages and ultimately to very different stellar endpoints.

This page titled 5.3: Stellar Evolution is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Robert Wagner.

5.3.1: Star Formation

Learning Objectives

By the end of this section, you will be able to:

- Identify the sometimes-violent processes by which parts of a molecular cloud collapse to produce stars
- Recognize some of the structures seen in images of molecular clouds like the one in Orion
- Explain how the environment of a molecular cloud enables the formation of stars
- Describe how advancing waves of star formation cause a molecular cloud to evolve

As we begin our exploration of how stars are formed, let's review some basics about stars discussed in earlier chapters:

- Stable (main-sequence) stars such as our Sun maintain equilibrium by producing energy through nuclear fusion in their cores. The ability to generate energy by fusion defines a star.
- Each second in the Sun, approximately 600 million tons of hydrogen undergo fusion into helium, with about 4 million tons turning into energy in the process. This rate of hydrogen use means that eventually the Sun (and all other stars) will run out of central fuel.
- Stars come with many different masses, ranging from 1/12 solar masses (M_{Sun}) to roughly 100–200 M_{Sun}. There are far more low-mass than high-mass stars.
- The most massive main-sequence stars (spectral type O) are also the most luminous and have the highest surface temperature. The lowest-mass stars on the main sequence (spectral type M or L) are the least luminous and the coolest.
- A galaxy of stars such as the Milky Way contains enormous amounts of gas and dust—enough to make billions of stars like the Sun.

If we want to find stars still in the process of formation, we must look in places that have plenty of the raw material from which stars are assembled. Since stars are made of gas, we focus our attention (and our telescopes) on the dense and cold clouds of gas and dust that dot the Milky Way (see Figure 21.1 and Figure 21.2).



(a) (b)

Figure 21.2 : Pillars of Dust and Dense Globules in M16. (a) This Hubble Space Telescope image of the central regions of M16 (also known as the Eagle Nebula) shows huge columns of cool gas, (including molecular hydrogen, H2) and dust. These columns are of higher density than the surrounding regions and have resisted evaporation by the ultraviolet radiation from a cluster of hot stars just beyond the upper-right corner of this image. The tallest pillar is about 1 light-year long, and the M16 region is about 7000 light-years away from us. (b) This close-up view of one of the pillars shows some very dense globules, many of which harbor embryonic stars. Astronomers coined the term *evaporating gas globules* (EGGs) for these structures, in part so that they could say we found EGGs

inside the Eagle Nebula. It is possible that because these EGGs are exposed to the relentless action of the radiation from nearby hot stars, some may not yet have collected enough material to form a star. (credit a : modification of work by NASA, ESA, and the Hubble Heritage Team (STScI/AURA); credit b: modification of work by NASA, ESA, STScI, J. Hester and P. Scowen (Arizona State University))

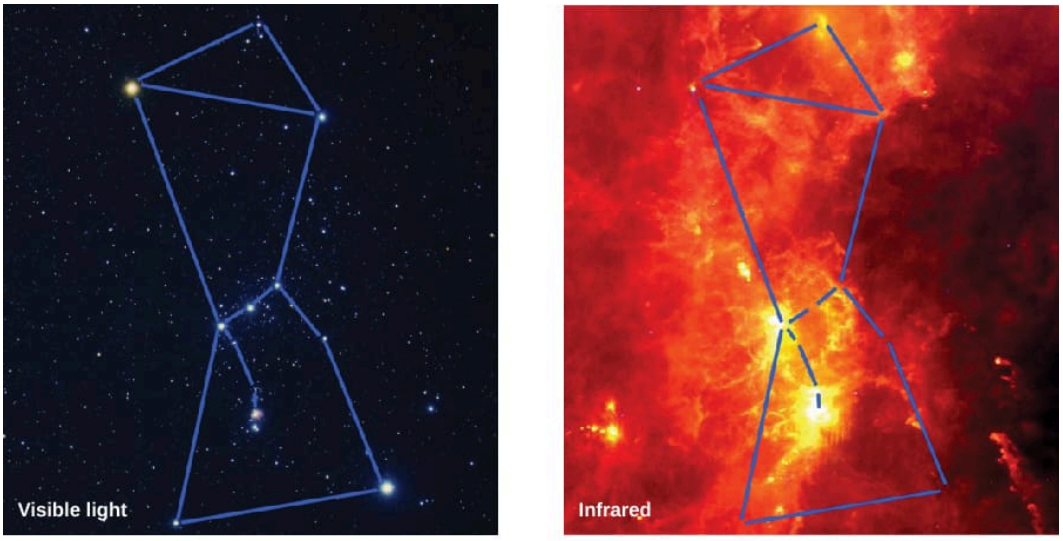
Molecular Clouds: Stellar Nurseries

As we saw in Between the Stars: Gas and Dust in Space, the most massive reservoirs of interstellar matter—and some of the most massive objects in the Milky Way Galaxy—are the giant molecular clouds. These clouds have cold interiors with characteristic temperatures of only 10–20 K; most of their gas atoms are bound into molecules. These clouds turn out to be the birthplaces of most stars in our Galaxy.

The masses of molecular clouds range from a thousand times the mass of the Sun to about 3 million solar masses. Molecular clouds have a complex filamentary structure, similar to cirrus clouds in Earth's atmosphere, but much less dense. The molecular cloud filaments can be up to 1000 light-years long. Within the clouds are cold, dense regions with typical masses of 50 to 500 times the mass of the Sun; we give these regions the highly technical name *clumps*. Within these clumps, there are even denser, smaller regions called cores. The cores are the embryos of stars. The conditions in these cores—low temperature and high density—are just what is required to make stars. Remember that the essence of the life story of any star is the ongoing competition between two forces: *gravity* and *pressure*. The force of gravity, pulling inward, tries to make a star collapse. Internal pressure produced by the motions of the gas atoms, pushing outward, tries to force the star to expand. When a star is first forming, low temperature (and hence, low pressure) and high density (hence, greater gravitational attraction) both work to give gravity the advantage. In order to form a star—that is, a dense, hot ball of matter capable of starting nuclear reactions deep within—we need a typical core of interstellar atoms and molecules to shrink in radius and increase in density by a factor of nearly 10^{20}. It is the force of gravity that produces this drastic collapse.

The Orion Molecular Cloud

Let's discuss what happens in regions of star formation by considering a nearby site where stars are forming right now. One of the best-studied stellar nurseries is in the constellation of Orion, The Hunter, about 1500 light-years away (Figure 21.3). The pattern of the hunter is easy to recognize by the conspicuous "belt" of three stars that mark his waist. The Orion molecular cloud is much larger than the star pattern and is truly an impressive structure. In its long dimension, it stretches over a distance of about 100 light-years. The total quantity of molecular gas is about 200,000 times the mass of the Sun. Most of the cloud does not glow with visible light but betrays its presence by the radiation that the dusty gas gives off at infrared and radio wavelengths.



(a) (b)

Figure 21.3 : Orion in Visible and Infrared. (a) The Orion star group was named after the legendary hunter in Greek mythology. Three stars close together in a link mark Orion's belt. The ancients imagined a sword hanging from the belt; the object at the end of the blue line in this sword is the Orion Nebula. (b) This wide-angle, infrared view of the same area was taken with the Infrared Astronomical Satellite. Heated

dust clouds dominate in this false-color image, and many of the stars that stood out on part (a) are now invisible. An exception is the cool, red-giant star Betelgeuse, which can be seen as a yellowish point at the left vertex of the blue triangle (at Orion's left armpit). The large, yellow ring to the right of Betelgeuse is the remnant of an exploded star. The infrared image lets us see how large and full of cooler material the Orion molecular cloud really is. On the visible-light image at left, you see only two colorful regions of interstellar matter—the two, bright yellow splotches at the left end of and below Orion's belt. The lower one is the Orion Nebula and the higher one is the region of the Horsehead Nebula. (credit: modification of work by NASA, visible light: Akira Fujii; infrared: Infrared Astronomical Satellite)

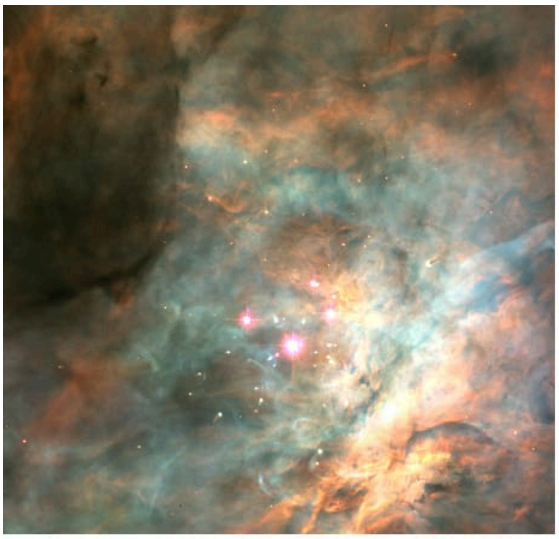
The stars in Orion's belt are typically about 5 million years old, whereas the stars near the middle of the "sword" hanging from Orion's belt are only 300,000 to 1 million years old. The region about halfway down the sword where star formation is still taking place is called the Orion Nebula. About 2200 young stars are found in this region, which is only slightly larger than a dozen light-years in diameter. The Orion Nebula also contains a tight cluster of stars called the Trapezium (Figure 21.5). The brightest Trapezium stars can be seen easily with a small telescope.



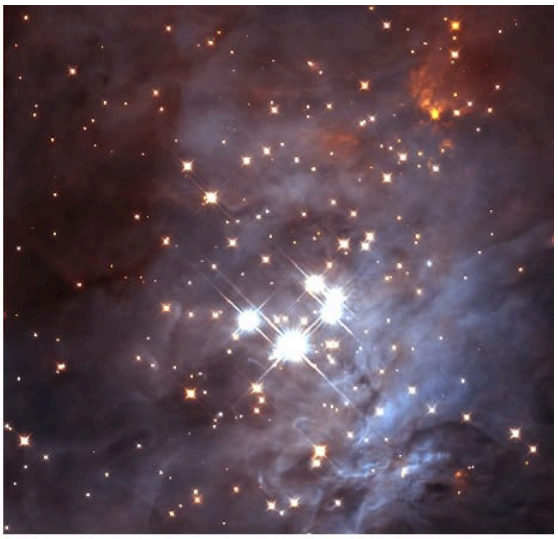
(a) (b)

Figure 21.4 : Orion Nebula. (a) The Orion Nebula is shown in visible light. (b) With near-infrared radiation, we can see more detail within the dusty nebula since infrared can penetrate dust more easily than can visible light. (credit a: modification of work by Filip Lolić; credit b: modification of work by NASA/JPL-Caltech/T. Megeath (University of Toledo, Ohio))

Compare this with our own solar neighborhood, where the typical spacing between stars is about 3 light-years. Only a small number of stars in the Orion cluster can be seen with visible light, but infrared images—which penetrate the dust better—detect the more than 2000 stars that are part of the group (Figure 21.5).



(a)



(b)

Figure 21.5 : Central Region of the Orion Nebula. The Orion Nebula harbors some of the youngest stars in the solar neighborhood. At the heart of the nebula is the Trapezium cluster, which includes four very bright stars that provide much of the energy that causes the nebula to glow so brightly. In these images, we see a section of the nebula in (a) visible light and (b) infrared. The four bright stars in the center of the visible-light image are the Trapezium stars. Notice that most of the stars seen in the infrared are completely hidden by dust in the visible-light image. (credit a: modification of work by NASA, C.R. O'Dell and S.K. Wong (Rice University); credit b: modification of work by NASA; K.L. Luhman (Harvard-Smithsonian Center for Astrophysics); and G. Schneider, E. Young, G. Rieke, A. Cotera, H. Chen, M. Rieke, R. Thompson (Steward Observatory, University of Arizona))

Studies of Orion and other star-forming regions show that star formation is not a very efficient process. In the region of the Orion Nebula, about 1% of the material in the cloud has been turned into stars. That is why we still see a substantial amount of gas and dust near the Trapezium stars. The leftover material is eventually heated, either by the radiation and winds from the hot stars that form or by explosions of the most massive stars. (We will see in later chapters that the most massive stars go through their lives very quickly and end by exploding.)

Link to Learning

Take a journey through the Orion Nebula to view a nice narrated video tour of this region.

Whether gently or explosively, the material in the neighborhood of the new stars is blown away into interstellar space. Older groups or clusters of stars can now be easily observed in visible light because they are no longer shrouded in dust and gas (Figure 21.6).

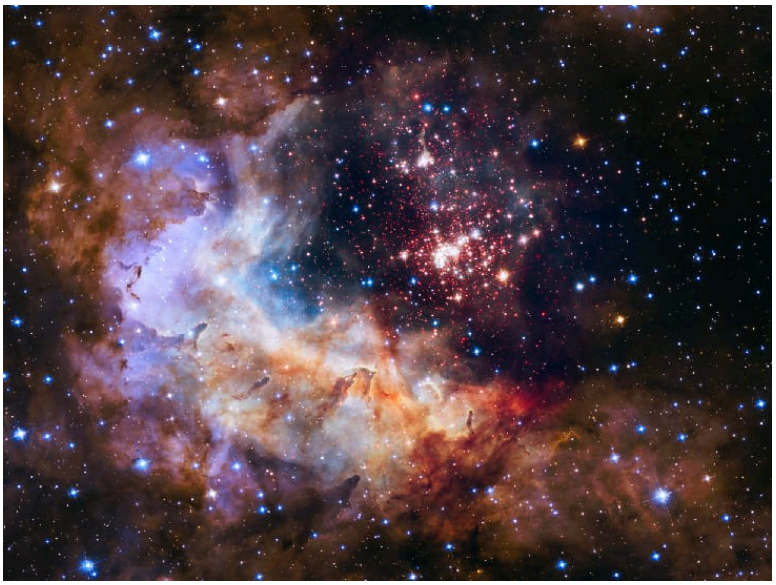


Figure 21.6 : Westerlund 2. This young cluster of stars known as Westerlund 2 formed within the Carina star-forming region about 2 million years ago. Stellar winds and pressure produced by the radiation from the hot stars within the cluster are blowing and sculpting the surrounding gas and dust. The nebula still contains many globules of dust. Stars are continuing to form within the denser globules and pillars of the nebula. This Hubble Space Telescope image includes near-infrared exposures of the star cluster and visible-light observations of the surrounding nebula. Colors in the nebula are dominated by the red glow of hydrogen gas, and blue-green emissions from glowing oxygen. (credit: NASA, ESA, the Hubble Heritage Team (STScI/AURA), A. Nota (ESA/STScI), and the Westerlund 2 Science Team)

Although we do not know what initially caused stars to begin forming in Orion, there is good evidence that the first generation of stars triggered the formation of additional stars, which in turn led to the formation of still more stars (Figure 21.7).

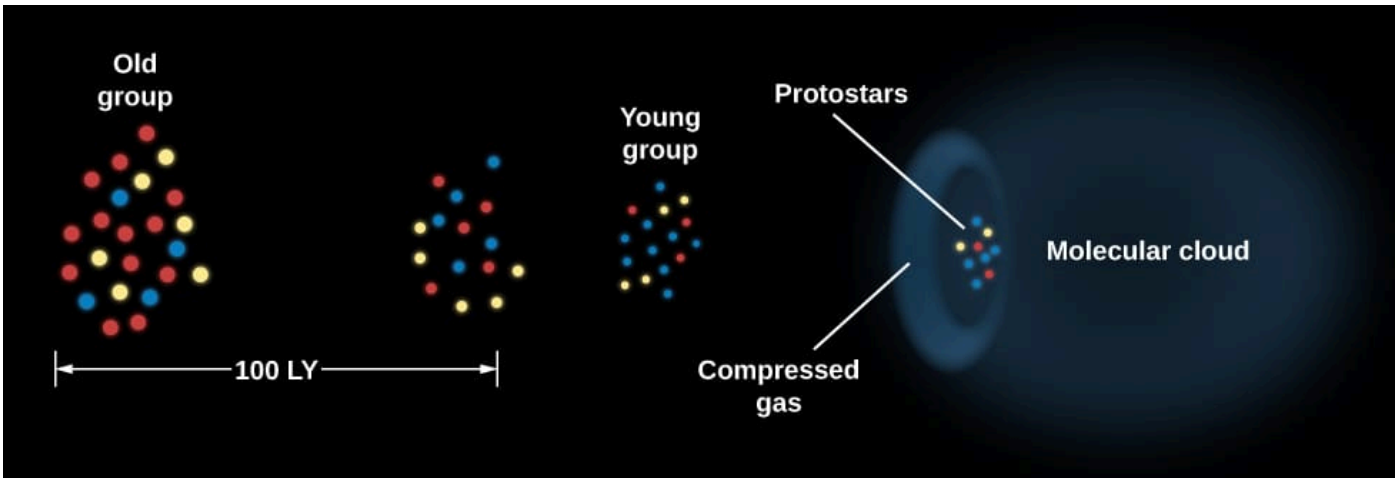


Figure 21.7 : Propagating Star Formation. Star formation can move progressively through a molecular cloud. The oldest group of stars lies to the left of the diagram and has expanded because of the motions of individual stars. Eventually, the stars in the group will disperse and no longer be recognizable as a cluster. The youngest group of stars lies to the right, next to the molecular cloud. This group of stars is only 1 to 2 million years old. The pressure of the hot, ionized gas surrounding these stars compresses the material in the nearby edge of the molecular cloud and initiates the gravitational collapse that will lead to the formation of more stars.

The basic idea of triggered star formation is this: when a massive star is formed, it emits a large amount of ultraviolet radiation and ejects high-speed gas in the form of a stellar wind. This injection of energy heats the gas around the stars and causes it to expand. When massive stars exhaust their supply of fuel, they explode, and the energy of the explosion also heats the gas. The hot gases pile into the surrounding cold molecular cloud, compressing the material in it and increasing its density. If this increase in density is large enough, gravity will overcome pressure, and stars will begin to form in the compressed gas. Such a chain reaction—where the brightest and hottest stars of one area become the cause of star formation “next door”—seems to have occurred not only in Orion but also in many other molecular clouds.

There are many molecular clouds that form only (or mainly) low-mass stars. Because low-mass stars do not have strong winds and do not die by exploding, triggered star formation cannot occur in these clouds. There are also stars that form in relative isolation in small cores. Therefore, not all star formation is originally triggered by the death of massive stars. However, there are likely to be other possible triggers, such as spiral density waves and other processes we do not yet understand.

The Birth of a Star

Although regions such as Orion give us clues about how star formation begins, the subsequent stages are still shrouded in mystery (and a lot of dust). There is an enormous difference between the density of a molecular cloud core and the density of the youngest stars that can be detected. Direct observations of this collapse to higher density are nearly impossible for two reasons. First, the dust-shrouded interiors of molecular clouds where stellar births take place cannot be observed with visible light. Second, the timescale for the initial collapse—thousands of years—is very short, astronomically speaking. Since each star spends such a tiny fraction of its life in this stage, relatively few stars are going through the collapse process at any given time. Nevertheless, through a combination of theoretical calculations and the limited observations available, astronomers have pieced together a picture of what the earliest stages of stellar evolution are likely to be.

The first step in the process of creating stars is the formation of dense cores within a clump of gas and dust (Figure 21.8(a)). It is generally thought that all the material for the star comes from the core, the larger structure surrounding the forming star. Eventually, the gravitational force of the infalling gas becomes strong enough to overwhelm the pressure exerted by the cold material that forms the dense cores. The material then undergoes a rapid collapse, and the density of the core increases greatly as a result. During the time a dense core is contracting to become a true star, but before the fusion of protons to produce helium begins, we call the object a protostar.

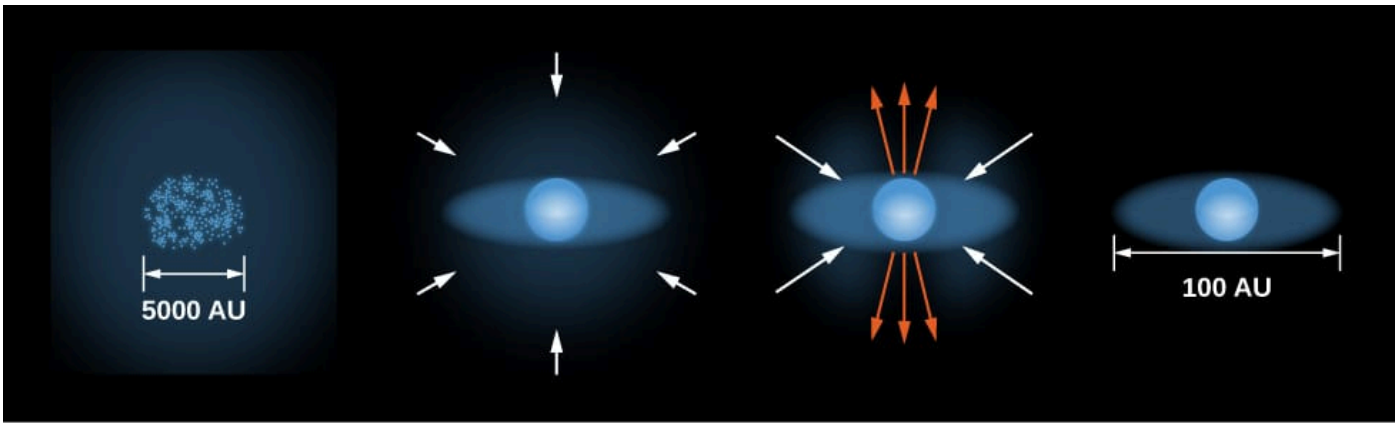


Figure 21.8 : Formation of a Star. (a) Dense cores form within a molecular cloud. (b) A protostar with a surrounding disk of material forms at the center of a dense core, accumulating additional material from the molecular cloud through gravitational attraction. (c) A stellar wind breaks out but is confined by the disk to flow out along the two poles of the star. (d) Eventually, this wind sweeps away the cloud material and halts the accumulation of additional material, and a newly formed star, surrounded by a disk, becomes observable. These sketches are not drawn to the same scale. The diameter of a typical envelope that is supplying gas to the newly forming star is about 5000 AU. The typical diameter of the disk is about 100 AU or slightly larger than the diameter of the orbit of Pluto.

The natural turbulence inside a clump tends to give any portion of it some initial spinning motion (even if it is very slow). As a result, each collapsing core is expected to spin. According to the law of conservation of angular momentum (discussed in the chapter on Orbits and Gravity), a rotating body spins more rapidly as it decreases in size. In other words, if the object can turn its material around a smaller circle, it can move that material more quickly—like a figure skater spinning more rapidly as she brings her arms in tight to her body. This is exactly what happens when a core contracts to form a protostar: as it shrinks, its rate of spin increases.

But all directions on a spinning sphere are not created equal. As the protostar rotates, it is much easier for material to fall right onto the poles (which spin most slowly) than onto the equator (where material moves around most rapidly). Therefore, gas and dust falling in toward the protostar's equator are "held back" by the rotation and form a whirling extended disk around the equator (part b in Figure 21.8). You may have observed this same "equator effect" on the amusement park ride in which you stand with

your back to a cylinder that is spun faster and faster. As you spin really fast, you are pushed against the wall so strongly that you cannot possibly fall toward the center of the cylinder. Gas can, however, fall onto the protostar easily from directions away from the star's equator.

The protostar and disk at this stage are embedded in an envelope of dust and gas from which material is still falling onto the protostar. This dusty envelope blocks visible light, but infrared radiation can get through. As a result, in this phase of its evolution, the protostar itself is emitting infrared radiation and so is observable only in the infrared region of the spectrum. Once almost all of the available material has been accreted and the central protostar has reached nearly its final mass, it is given a special name: it is called a *T Tauri star*, named after one of the best studied and brightest members of this class of stars, which was discovered in the constellation of Taurus. (Astronomers have a tendency to name types of stars after the first example they discover or come to understand. It's not an elegant system, but it works.) Only stars with masses less than or similar to the mass of the Sun become T Tauri stars. Massive stars do not go through this stage, although they do appear to follow the formation scenario illustrated in Figure 21.8.

Winds and Jets

Recent observations suggest that T Tauri stars may actually be stars in a middle stage between protostars and hydrogen-fusing stars such as the Sun. High-resolution infrared images have revealed jets of material as well as *stellar winds* coming from some T Tauri stars, proof of interaction with their environment. A stellar wind consists mainly of protons (hydrogen nuclei) and electrons streaming away from the star at speeds of a few hundred kilometers per second (several hundred thousand miles per hour). When the wind first starts up, the disk of material around the star's equator blocks the wind in its direction. Where the wind particles *can* escape most effectively is in the direction of the star's poles.

Astronomers have actually seen evidence of these beams of particles shooting out in opposite directions from the polar regions of newly formed stars. In many cases, these beams point back to the location of a protostar that is still so completely shrouded in dust that we cannot yet see it (Figure 21.9).

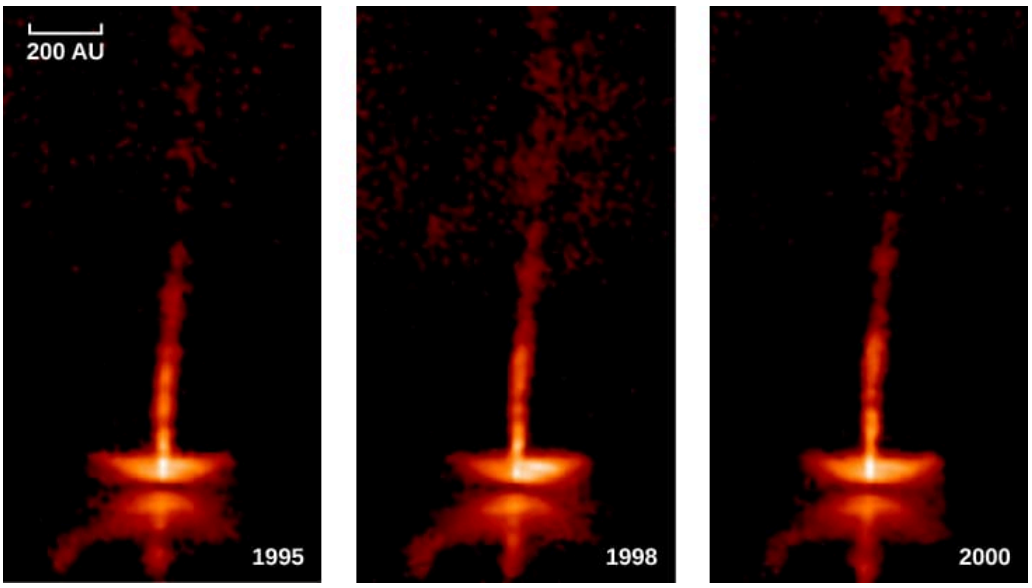


Figure 21.9 : Gas Jets Flowing away from a Protostar. Here we see the neighborhood of a protostar, known to us as HH 34 because it is a Herbig-Haro object. The star is about 450 light-years away and only about 1 million years old. Light from the star itself is blocked by a disk, which is larger than 60 billion kilometers in diameter and is seen almost edge-on. Jets are seen emerging perpendicular to the disk. The material in these jets is flowing outward at speeds up to 580,000 kilometers per hour. The series of three images shows changes during a period of 5 years. Every few months, a compact clump of gas is ejected, and its motion outward can be followed. The changes in the brightness of the disk may be due to motions of clouds within the disk that alternately block some of the light and then let it through. This image corresponds to the stage in the life of a protostar shown in part (c) of Figure 21.8. (credit: modification of work by Hubble Space Telescope, NASA, ESA)

On occasion, the jets of high-speed particles streaming away from the protostar collide with a somewhat-denser lump of gas nearby, excite its atoms, and cause them to emit light. These glowing regions, each of which is known as a Herbig-Haro (HH) object after the two astronomers who first identified them, allow us to trace the progress of the jet to a distance of a light-year or

more from the star that produced it. Figure 21.10 shows two spectacular images of HH objects.

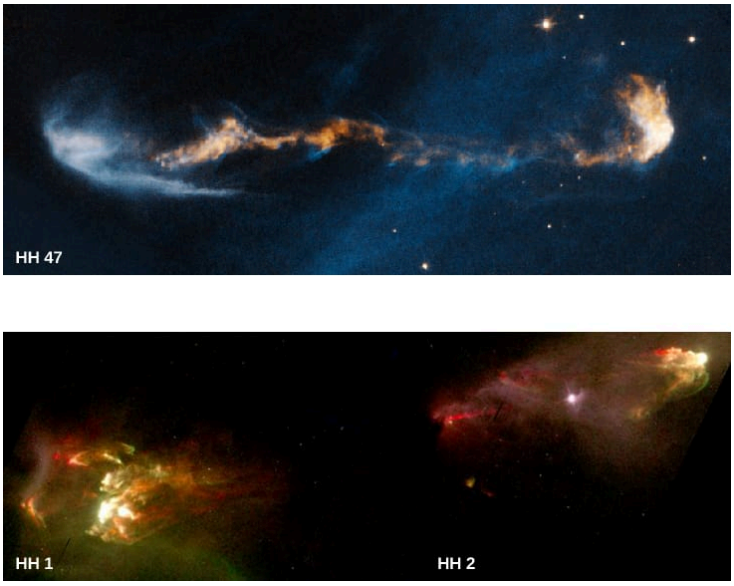


Figure 21.10 : Outflows from Protostars. These images were taken with the Hubble Space Telescope and show jets flowing outward from newly formed stars. In the HH47 image, a protostar 1500 light-years away (invisible inside a dust disk at the left edge of the image) produces a very complicated jet. The star may actually be wobbling, perhaps because it has a companion. Light from the star illuminates the white region at the left because light can emerge perpendicular to the disk (just as the jet does). At right, the jet is plowing into existing clumps of interstellar gas, producing a shock wave that resembles an arrowhead. The HH1/2 image shows a double-beam jet emanating from a protostar (hidden in a dust disk in the center) in the constellation of Orion. Tip to tip, these jets are more than 1 light-year long. The bright regions (first identified by Herbig and Haro) are places where the jet is a slamming into a clump of interstellar gas and causing it to glow. (credit “HH 47”: modification of work by NASA, ESA, and P. Hartigan (Rice University); credit “HH 1 and HH 2: modification of work by J. Hester, WFPC2 Team, NASA)

The wind from a forming star will ultimately sweep away the material that remains in the obscuring envelope of dust and gas, leaving behind the naked disk and protostar, which can then be seen with visible light. We should note that at this point, the protostar itself is still contracting slowly and has not yet reached the main-sequence stage on the H–R diagram (a concept introduced in the chapter The Stars: A Celestial Census). The disk can be detected directly when observed at infrared wavelengths or when it is seen silhouetted against a bright background (Figure 21.11).

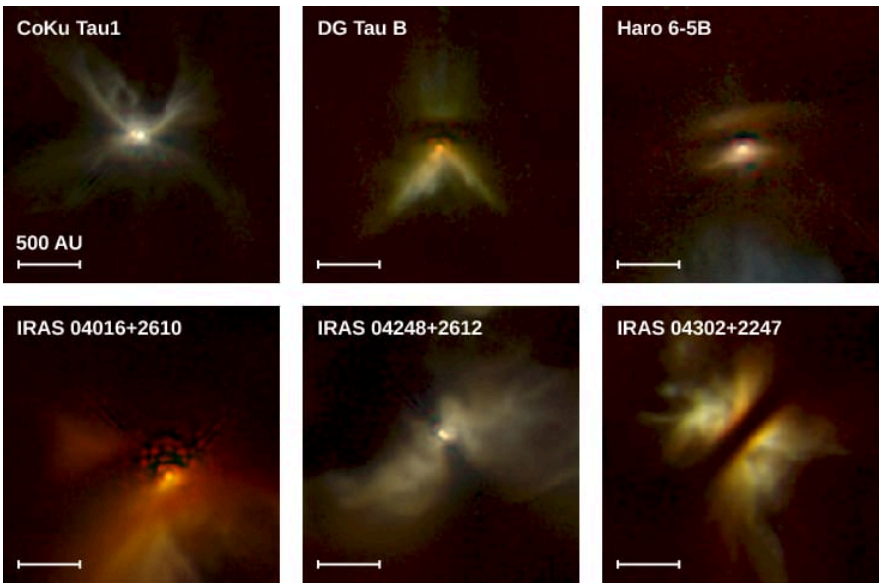


Figure 21.11 : Disks around Protostars. These Hubble Space Telescope infrared images show disks around young stars in the constellation of Taurus, in a region about 450 light-years away. In some cases, we can see the central star (or stars—some are binaries). In other cases, the dark, horizontal

bands indicate regions where the dust disk is so thick that even infrared radiation from the star embedded within it cannot make its way through. The brightly glowing regions are starlight reflected from the upper and lower surfaces of the disk, which are less dense than the central, dark regions. (Credit: modification of work by D. Padgett (IPAC/Caltech), W. Brandner (IPAC), K. Stapelfeldt (JPL) and NASA)

This description of a protostar surrounded by a rotating disk of gas and dust sounds very much like what happened in our solar system when the Sun and planets formed. Indeed, one of the most important discoveries from the study of star formation in the last decade of the twentieth century was that disks are an inevitable byproduct of the process of creating stars. The next questions that astronomers set out to answer was: will the disks around protostars also form planets? And if so, how often? We will return to these questions later in this chapter.

To keep things simple, we have described the formation of single stars. Many stars, however, are members of binary or triple systems, where several stars are born together. In this case, the stars form in nearly the same way. Widely separated binaries may each have their own disk; close binaries may share a single disk.

This page titled 5.3.1: Star Formation is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by OpenStax via source content that was edited to the style and standards of the LibreTexts platform.

5.3.2: The H-R Diagram and the Study of Stellar Evolution

Learning Objectives

By the end of this section, you will be able to:

- Determine the age of a protostar using an H–R diagram and the protostar's luminosity and temperature
- Explain the interplay between gravity and pressure, and how the contracting protostar changes its position in the H–R diagram as a result

One of the best ways to summarize all of these details about how a star or protostar changes with time is to use a Hertzsprung-Russell (H–R) diagram. Recall from The Stars: A Celestial Census that, when looking at an H–R diagram, the temperature (the horizontal axis) is plotted increasing toward the left. As a star goes through the stages of its life, its luminosity and temperature change. Thus, its position on the H–R diagram, in which luminosity is plotted against temperature, also changes. As a star ages, we must replot it in different places on the diagram. Therefore, astronomers often speak of a star *moving* on the H–R diagram, or of its evolution tracing out a path on the diagram. In this context, "tracing out a path" has nothing to do with the star's motion through space; this is just a shorthand way of saying that its temperature and luminosity change as it evolves.

Link to Learning

Watch an animation of the stars in the Omega Centauri cluster as they rearrange according to luminosity and temperature, forming a Hertzsprung-Russell (H–R) diagram.

To estimate just how much the luminosity and temperature of a star change as it ages, we must resort to calculations. Theorists compute a series of models for a star, with each successive model representing a later point in time. Stars may change for a variety of reasons. Protostars, for example, change in size because they are contracting, and their temperature and luminosity change as they do so. After nuclear fusion begins in the star's core (see Stars from Adolescence to Old Age), main-sequence stars change because they are using up their nuclear fuel.

Given a model that represents a star at one stage of its evolution, we can calculate what it will be like at a slightly later time. At each step, the model predicts the luminosity and size of the star, and from these values, we can figure out its surface temperature. A series of points on an H–R diagram, calculated in this way, allows us to follow the life changes of a star and hence is called its *evolutionary track*.

Evolutionary Tracks

Let's now use these ideas to follow the evolution of protostars that are on their way to becoming main-sequence stars. The evolutionary tracks of newly forming stars with a range of stellar masses are shown in Figure 21.12. These young stellar objects are not yet producing energy by nuclear reactions, but they derive energy from gravitational contraction—through the sort of process proposed for the Sun by Helmhotz and Kelvin in this last century (see the chapter on The Sun: A Nuclear Powerhouse).

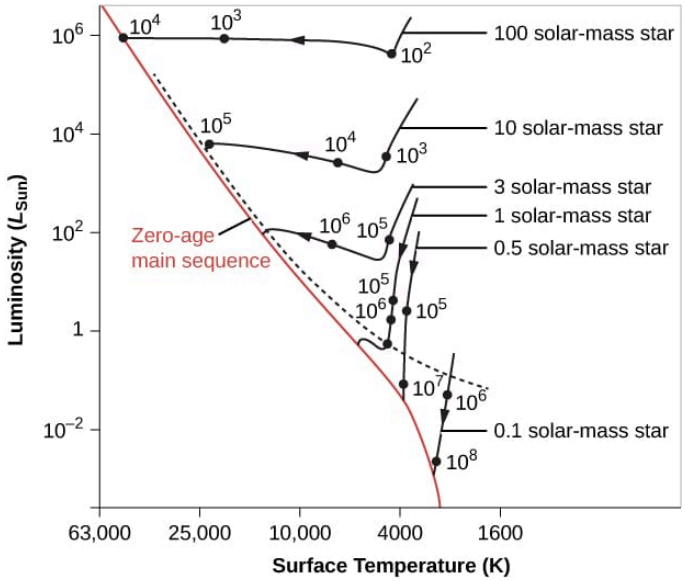


Figure 21.12 : Evolutionary Tracks for Contracting Protostars. Tracks are plotted on the H–R diagram to show how stars of different masses change during the early parts of their lives. The number next to each dark point on a track is the rough number of years it takes an embryo star to reach that stage (the numbers are the result of computer models and are therefore not well known). Note that the surface temperature (K) on the horizontal axis increases toward the left. You can see that the more mass a star has, the shorter time it takes to go through each stage. Stars above the dashed line are typically still surrounded by infalling material and are hidden by it.

Initially, a protostar remains fairly cool with a very large radius and a very low density. It is transparent to infrared radiation, and the heat generated by gravitational contraction can be radiated away freely into space. Because heat builds up slowly inside the protostar, the gas pressure remains low, and the outer layers fall almost unhindered toward the center. Thus, the protostar undergoes very rapid collapse, a stage that corresponds to the roughly vertical lines at the right of Figure 21.12. As the star shrinks, its surface area gets smaller, and so its total luminosity decreases. The rapid contraction stops only when the protostar becomes dense and opaque enough to trap the heat released by gravitational contraction.

When the star begins to retain its heat, the contraction becomes much slower, and changes inside the contracting star keep the luminosity of stars like our Sun roughly constant. The surface temperatures start to build up, and the star “moves” to the left in the H–R diagram. Stars first become visible only after the stellar wind described earlier clears away the surrounding dust and gas. This can happen during the rapid-contraction phase for low-mass stars, but high-mass stars remain shrouded in dust until they end their early phase of gravitational contraction (see the dashed line in Figure 21.12).

To help you keep track of the various stages that stars go through in their lives, it can be useful to compare the development of a star to that of a human being. (Clearly, you will not find an exact correspondence, but thinking through the stages in human terms may help you remember some of the ideas we are trying to emphasize.) Protostars might be compared to human embryos—as yet unable to sustain themselves but drawing resources from their environment as they grow. Just as the birth of a child is the moment it is called upon to produce its own energy (through eating and breathing), so astronomers say that a star is born when it is able to sustain itself through nuclear reactions (by making its own energy.)

When the star’s central temperature becomes high enough (about 12 million K) to fuse hydrogen into helium, we say that the star has reached the main sequence (a concept introduced in The Stars: A Celestial Census). It is now a full-fledged star, more or less in equilibrium, and its rate of change slows dramatically. Only the gradual depletion of hydrogen as it is transformed into helium in the core slowly changes the star’s properties.

The mass of a star determines exactly where it falls on the main sequence. As Figure 21.12 shows, massive stars on the main sequence have high temperatures and high luminosities. Low-mass stars have low temperatures and low luminosities.

Objects of extremely low mass never achieve high-enough central temperatures to ignite nuclear reactions. The lower end of the main sequence stops where stars have a mass just barely great enough to sustain nuclear reactions at a sufficient rate to stop gravitational contraction. This critical mass is calculated to be about 0.075 times the mass of the Sun. As we discussed in the chapter on Analyzing Starlight, objects below this critical mass are called either brown dwarfs or planets. At the other extreme, the upper end of the main sequence terminates at the point where the energy radiated by the newly forming massive star becomes so great that it halts the accretion of additional matter. The upper limit of stellar mass is between 100 and 200 solar masses.

Evolutionary Timescales

How long it takes a star to form depends on its mass. The numbers that label the points on each track in Figure 21.12 are the times, in years, required for the embryo stars to reach the stages we have been discussing. Stars of masses much higher than the Sun's reach the main sequence in a few thousand to a million years. The Sun required millions of years before it was born. Tens of millions of years are required for stars of lower mass to evolve to the lower main sequence. (We will see that this turns out to be a general principle: massive stars go through *all* stages of evolution faster than low-mass stars do.)

We will take up the subsequent stages in the life of a star in Stars from Adolescence to Old Age, examining what happens after stars arrive in the main sequence and begin a "prolonged adolescence" and "adulthood" of fusing hydrogen to form helium. But now we want to examine the connection between the formation of stars and planets.

This page titled 5.3.2: The H-R Diagram and the Study of Stellar Evolution is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by OpenStax via source content that was edited to the style and standards of the LibreTexts platform.

5.3.3: Evolution from the Main Sequence to Red Giants

Learning Objectives

By the end of this section, you will be able to:

- Explain the zero-age main sequence
- Describe what happens to main-sequence stars of various masses as they exhaust their hydrogen supply

One of the best ways to get a "snapshot" of a group of stars is by plotting their properties on an H–R diagram. We have already used the H–R diagram to follow the evolution of protostars up to the time they reach the main sequence. Now we'll see what happens next.

Once a star has reached the main-sequence stage of its life, it derives its energy almost entirely from the conversion of hydrogen to helium via the process of nuclear fusion in its core (see The Sun: A Nuclear Powerhouse). Since hydrogen is the most abundant element in stars, this process can maintain the star's equilibrium for a long time. Thus, all stars remain on the main sequence for most of their lives. Some astronomers like to call the main-sequence phase the star's "prolonged adolescence" or "adulthood" (continuing our analogy to the stages in a human life).

The left-hand edge of the main-sequence band in the H–R diagram is called the zero-age main sequence (see Figure 18.15). We use the term *zero-age* to mark the time when a star stops contracting, settles onto the main sequence, and begins to fuse hydrogen in its core. The zero-age main sequence is a continuous line in the H–R diagram that shows where stars of different masses but similar chemical composition can be found when they begin to fuse hydrogen.

Since only 0.7% of the hydrogen used in fusion reactions is converted into energy, fusion does not change the *total* mass of the star appreciably during this long period. It does, however, change the chemical composition in its central regions where nuclear reactions occur: hydrogen is gradually depleted, and helium accumulates. This change of composition changes the luminosity, temperature, size, and interior structure of the star. When a star's luminosity and temperature begin to change, the point that represents the star on the H–R diagram moves away from the zero-age main sequence.

Calculations show that the temperature and density in the inner region slowly increase as helium accumulates in the center of a star. As the temperature gets hotter, each proton acquires more energy of motion on average; this means it is more likely to interact with other protons, and as a result, the rate of fusion also increases. For the proton-proton cycle described in The Sun: A Nuclear Powerhouse, the rate of fusion goes up roughly as the temperature to the fourth power.

If the rate of fusion goes up, the rate at which energy is being generated also increases, and the luminosity of the star gradually rises. Initially, however, these changes are small, and stars remain within the main-sequence band on the H–R diagram for most of their lifetimes.

Example 22.1: Star Temperature and Rate of Fusion

If a star's temperature were to double, by what factor would its rate of fusion increase?

Solution

Since the rate of fusion (like temperature) goes up to the fourth power, it would increase by a factor of 2^4, or 16 times.

Exercise 5.3.3.1

If the rate of fusion of a star increased 256 times, by what factor would the temperature increase?

Answer available at the end of this section.

Lifetimes on the Main Sequence

How many years a star remains in the main-sequence band depends on its mass. You might think that a more massive star, having more fuel, would last longer, but it's not that simple. The lifetime of a star in a particular stage of evolution depends on how much nuclear fuel it has and on *how quickly* it uses up that fuel. (In the same way, how long people can keep spending money depends not only on how much money they have but also on how quickly they spend it. This is why many lottery winners who go on spending sprees quickly wind up poor again.) In the case of stars, more massive ones use up their fuel much more quickly than stars of low mass.

The reason massive stars are such spendthrifts is that, as we saw above, the rate of fusion depends *very* strongly on the star's core temperature. And what determines how hot a star's central regions get? It is the *mass* of the star—the weight of the overlying layers determines how high the pressure in the core must be: higher mass requires higher pressure to balance it. Higher pressure, in turn, is produced by higher temperature. The higher the temperature in the central regions, the faster the star races through its storehouse of central hydrogen. Although massive stars have more fuel, they burn it so prodigiously that their lifetimes are much shorter than those of their low-mass counterparts. You can also understand now why the most massive main-sequence stars are also the most luminous. Like new rock stars with their first platinum album, they spend their resources at an astounding rate.

The main-sequence lifetimes of stars of different masses are listed in Table 22.1. This table shows that the most massive stars spend only a few million years on the main sequence. A star of 1 solar mass remains there for roughly 10 billion years, while a star of about 0.4 solar mass has a main-sequence lifetime of some 200 billion years, which is longer than the current age of the universe. (Bear in mind, however, that every star spends *most* of its total lifetime on the main sequence. Stars devote an average of 90% of their lives to peacefully fusing hydrogen into helium.)

Table 22.1: Lifetimes of Main-Sequence Stars

Spectral Type	Surface Temperature (K)	Mass (Mass of Sun = 1)	Lifetime on Main Sequence (years)
O5	54,000	40	1 million
B0	29,200	16	10 million
A0	9600	3.3	500 million
F0	7350	1.7	2.7 billion
G0	6050	1.1	9 billion
K0	5240	0.8	14 billion
M0	3750	0.4	200 billion

These results are not merely of academic interest. Human beings developed on a planet around a G-type star. This means that the Sun's stable main-sequence lifetime is so long that it afforded life on Earth plenty of time to evolve. When searching for intelligent life like our own on planets around other stars, it would be a pretty big waste of time to search around O- or B-type stars. These stars remain stable for such a short time that the development of creatures complicated enough to take astronomy courses is very unlikely.

From Main-Sequence Star to Red Giant

Eventually, all the hydrogen in a star's core, where it is hot enough for fusion reactions, is used up. The core then contains only helium, "contaminated" by whatever small percentage of heavier elements the star had to begin with. The helium in the core can be thought of as the accumulated "ash" from the nuclear "burning" of hydrogen during the main-sequence stage.

Energy can no longer be generated by hydrogen fusion in the stellar core because the hydrogen is all gone and, as we will see, the fusion of helium requires much higher temperatures. Since the central temperature is not yet high enough to fuse helium, there is no nuclear energy source to supply heat to the central region of the star. The long period of stability now ends, gravity again takes over, and the core begins to contract. Once more, the star's energy is partially supplied by gravitational energy, in the way described by Kelvin and Helmholtz (see Sources of Sunshine: Thermal and Gravitational Energy). As the star's core shrinks, the energy of the inward-falling material is converted to heat.

The heat generated in this way, like all heat, flows outward to where it is a bit cooler. In the process, the heat raises the temperature of a layer of hydrogen that spent the whole long main-sequence time just outside the core. Like an understudy waiting in the wings of a hit Broadway show for a chance at fame and glory, this hydrogen was almost (but not quite) hot enough to undergo fusion and take part in the main action that sustains the star. Now, the additional heat produced by the shrinking core puts this hydrogen "over the limit," and a shell of hydrogen nuclei just outside the core becomes hot enough for hydrogen fusion to begin.

New energy produced by fusion of this hydrogen now pours outward from this shell and begins to heat up layers of the star farther out, causing them to expand. Meanwhile, the helium core continues to contract, producing more heat right around it. This leads to more fusion in the shell of fresh hydrogen outside the core (Figure 22.2). The additional fusion produces still more energy, which also flows out into the upper layer of the star.

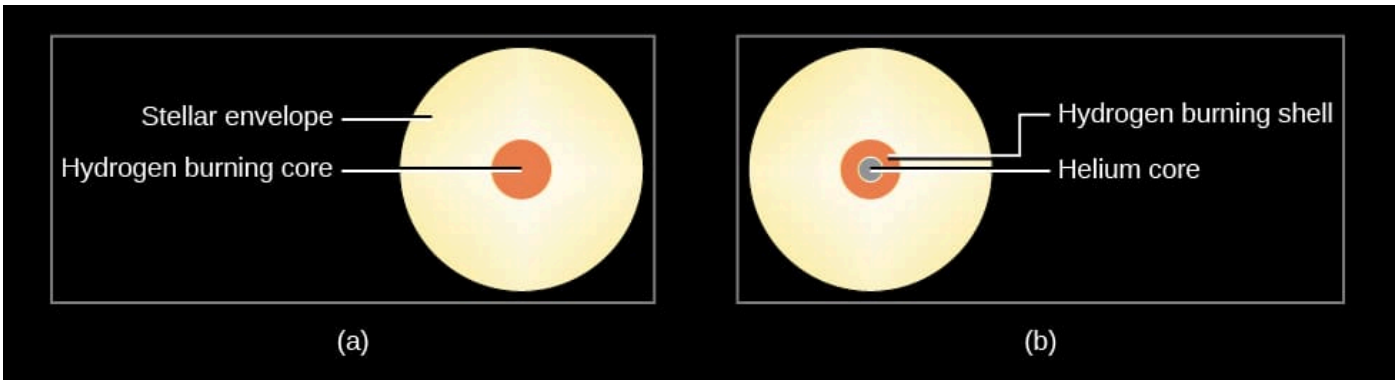


Figure 22.2 : Star Layers during and after the Main Sequence. (a) During the main sequence, a star has a core where fusion takes place and a much larger envelope that is too cold for fusion. (b) When the hydrogen in the core is exhausted (made of helium, not hydrogen), the core is compressed by gravity and heats up. The additional heat starts hydrogen fusion in a layer just outside the core. Note that these parts of the Sun are not drawn to scale.

Most stars actually generate more energy each second when they are fusing hydrogen in the shell surrounding the helium core than they did when hydrogen fusion was confined to the central part of the star; thus, they increase in luminosity. With all the new energy pouring outward, the outer layers of the star begin to expand, and the star eventually grows and grows until it reaches enormous proportions (Figure 22.3).

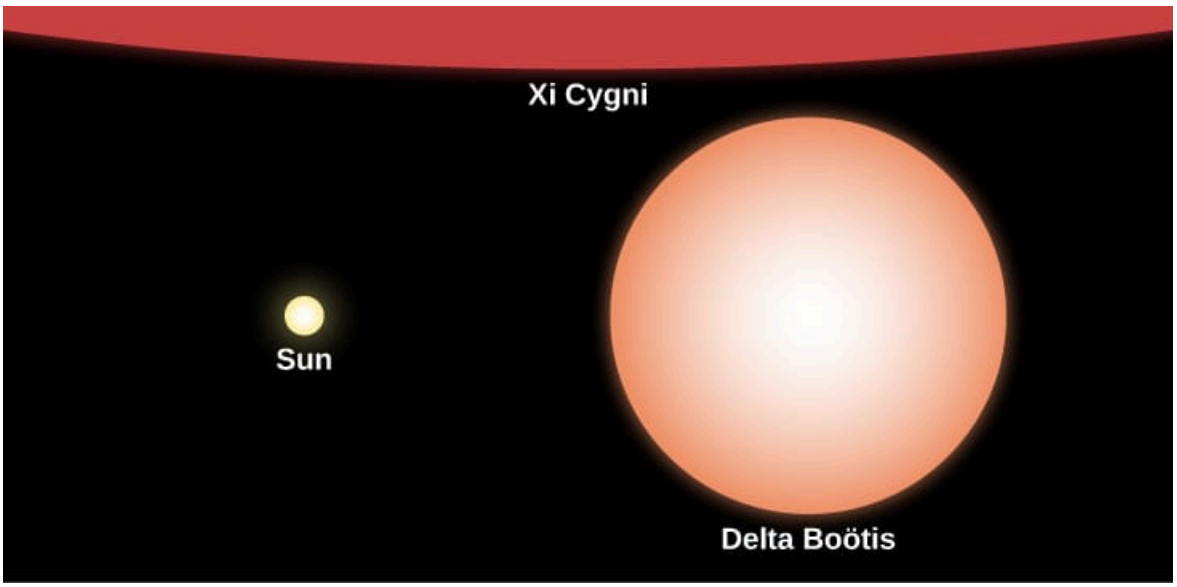


Figure 22.3 : Relative Sizes of Stars. This image compares the size of the Sun to that of Delta Boötis, a giant star, and Xi Cygni, a supergiant. Note that Xi Cygni is so large in comparison to the other two stars that only a small portion of it is visible at the top of the frame.

When you take the lid off a pot of boiling water, the steam can expand and it cools down. In the same way, the expansion of a star's outer layers causes the temperature at the surface to decrease. As it cools, the star's overall color becomes redder. (We saw

in Radiation and Spectra that a red color corresponds to cooler temperature.)

So the star becomes simultaneously more luminous and cooler. On the H–R diagram, the star therefore leaves the main-sequence band and moves upward (brighter) and to the right (cooler surface temperature). Over time, massive stars become red supergiants, and lower-mass stars like the Sun become red giants. (We first discussed such giant stars in The Stars: A Celestial Census; here we see how such "swollen" stars originate.) You might also say that these stars have "split personalities": their cores are contracting while their outer layers are expanding. (Note that red giant stars do not actually look deep red; their colors are more like orange or orange-red.)

Just how different are these red giants and supergiants from a main-sequence star? Table 22.2 compares the Sun with the red supergiant Betelgeuse, which is visible above Orion's belt as the bright red star that marks the hunter's armpit. Relative to the Sun, this supergiant has a much larger radius, a much lower average density, a cooler surface, and a much hotter core.

Table 22.2: Comparing a Supergiant with the Sun

Property	Sun	Betelgeuse
Mass (2×1033 g)(2×1033 g)	1	16
Radius (km)	700,000	500,000,000
Surface temperature (K)	5,800	3,600
Core temperature (K)	15,000,000	160,000,000
Luminosity (4×1026 W)(4×1026 W)	1	46,000
Average density (g/cm^3)	1.4	1.3×10-71.3×10-7
Age (millions of years)	4,500	10

Red giants can become so large that if we were to replace the Sun with one of them, its outer atmosphere would extend to the orbit of Mars or even beyond (Figure 22.4). This is the next stage in the life of a star as it moves (to continue our analogy to human lives) from its long period of "youth" and "adulthood" to "old age." By considering the relative ages of the Sun and Betelgeuse, we can also see that the idea that "bigger stars die faster" is indeed true here. Betelgeuse is a mere 10 million years old, which is relatively young compared with our Sun's 4.5 billion years, but it is already nearing its death throes as a red supergiant.

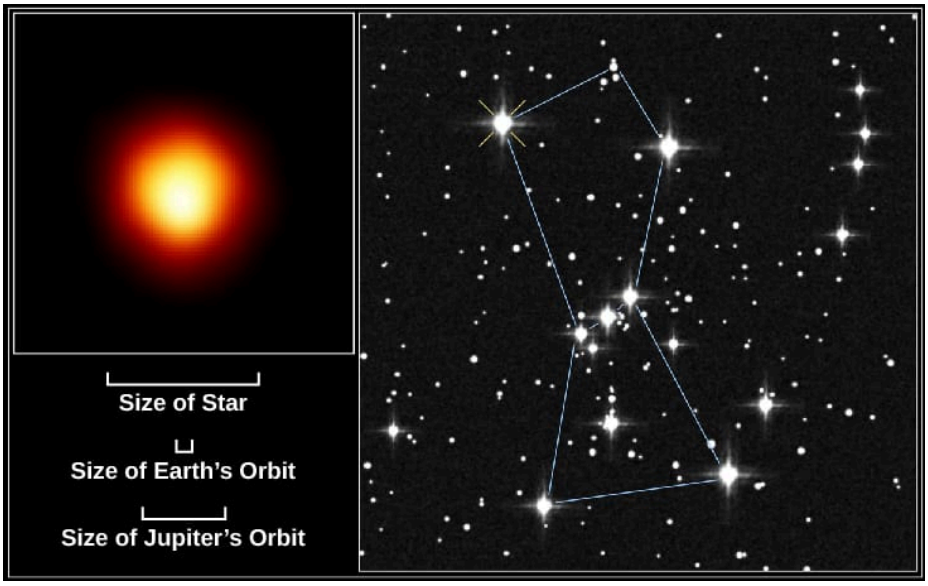


Figure 22.4 : Betelgeuse. Betelgeuse is in the constellation Orion, the hunter; in the right image, it is marked with a yellow "X" near the top left. In the left image, we see it in ultraviolet with the Hubble Space Telescope, in the first direct image ever made of the surface of another star. As shown by the scale at the bottom, Betelgeuse has an extended atmosphere so large that, if it were at the center of our solar system, it would stretch past the orbit of Jupiter. (credit: Modification of work by Andrea Dupree (Harvard-Smithsonian CfA), Ronald Gilliland (STScI), NASA and ESA)

Models for Evolution to the Giant Stage

As we discussed earlier, astronomers can construct computer models of stars with different masses and compositions to see how stars change throughout their lives. Figure 22.5, which is based on theoretical calculations by University of Illinois astronomer Icko Iben, shows an H–R diagram with several tracks of evolution from the main sequence to the giant stage. Tracks are shown for stars with different masses (from 0.5 to 15 times the mass of our Sun) and with chemical compositions similar to that of the Sun. The red line is the initial or zero-age main sequence. The numbers along the tracks indicate the time, in years, required for each star to reach those points in their evolution after leaving the main sequence. Once again, you can see that the more massive a star is, the more quickly it goes through each stage in its life.

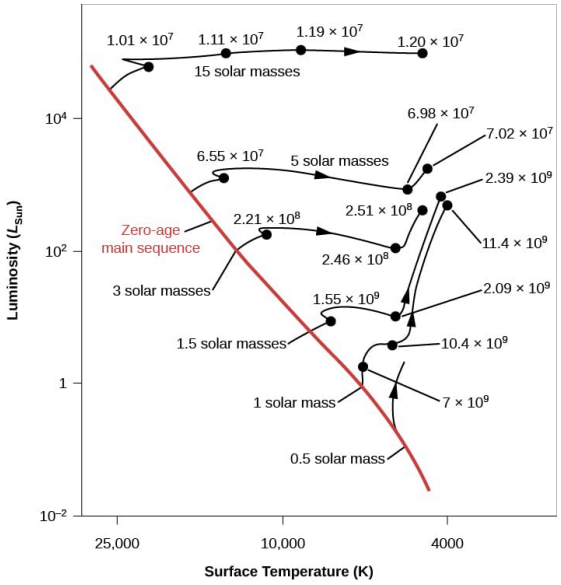


Figure 22.5 : Evolutionary Tracks of Stars of Different Masses. The solid black lines show the predicted evolution from the main sequence through the red giant or supergiant stage on the H–R diagram. Each track is labeled with the mass of the star it is describing. The numbers show how many years each star takes to become a giant. The red line is the zero-age main sequence. While theorists debate the exact number of years shown here, our main point should be clear. The more massive the star, the shorter time it takes for each stage in its life.

Note that the most massive star in this diagram has a mass similar to that of Betelgeuse, and so its evolutionary track shows approximately the history of Betelgeuse. The track for a 1-solar-mass star shows that the Sun is still in the main-sequence phase of evolution, since it is only about 4.5 billion years old. It will be billions of years before the Sun begins its own “climb” away from the main sequence—the expansion of its outer layers that will make it a red giant.

Link to Learning

Use the Star in a Box simulation to explore the evolution of stars of different masses. Select the star’s mass, and then hit play to see how its luminosity, temperature, and size change during its lifetime.

Answers

Exercise 5.3.3.1

The temperature would increase by a factor of $256^{0.25}$ (that is, the 4th root of 256), or 4 times.

This page titled 5.3.3: Evolution from the Main Sequence to Red Giants is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by OpenStax via source content that was edited to the style and standards of the LibreTexts platform.

5.3.4: The Death of Low-Mass Stars

Learning Objectives

By the end of this section, you will be able to:

- Describe the physical characteristics of degenerate matter and explain how the mass and radius of degenerate stars are related
- Plot the future evolution of a white dwarf and show how its observable features will change over time
- Distinguish which stars will become white dwarfs

Let's begin with those stars whose final mass just before death is less than about 1.4 times the mass of the Sun (M_{Sun}). (We will explain why this mass is the crucial dividing line in a moment.) Note that most stars in the universe fall into this category. The number of stars decreases as mass increases; really massive stars are rare (see The Stars: A Celestial Census). This is similar to the music business where only a few musicians ever become superstars. Furthermore, many stars with an initial mass much greater than 1.4 M_{Sun} will be reduced to that level by the time they die. For example, we now know that stars that start out with masses of at least 8.0 M_{Sun} (and possibly as much as 10 M_{Sun}) manage to lose enough mass during their lives to fit into this category (an accomplishment anyone who has ever attempted to lose weight would surely envy).

A Star in Crisis

In the last chapter, we left the life story of a star with a mass like the Sun's just after it had climbed up to the red-giant region of the H–R diagram for a second time and had shed some of its outer layers to form a planetary nebula. Recall that during this time, the *core* of the star was undergoing an "energy crisis." Earlier in its life, during a brief stable period, helium in the core had gotten hot enough to fuse into carbon (and oxygen). But after this helium was exhausted, the star's core had once more found itself without a source of pressure to balance gravity and so had begun to contract.

This collapse is the final event in the life of the core. Because the star's mass is relatively low, it cannot push its core temperature high enough to begin another round of fusion (in the same way larger-mass stars can). The core continues to shrink until it reaches a density equal to nearly a million times the density of water! That is 200,000 times greater than the average density of Earth. At this extreme density, a new and different way for matter to behave kicks in and helps the star achieve a final state of equilibrium. In the process, what remains of the star becomes one of the strange *white dwarfs* that we met in The Stars: A Celestial Census.

Degenerate Stars

Because white dwarfs are far denser than any substance on Earth, the matter inside them behaves in a very unusual way—unlike anything we know from everyday experience. At this high density, gravity is incredibly strong and tries to shrink the star still further, but all the *electrons* resist being pushed closer together and set up a powerful pressure inside the core. This pressure is the result of the fundamental rules that govern the behavior of electrons (the quantum physics you were introduced to in The Sun: A Nuclear Powerhouse). According to these rules (known to physicists as the Pauli exclusion principle), which have been verified in studies of atoms in the laboratory, no two electrons can be in the same place at the same time doing the same thing. We specify the *place* of an electron by its position in space, and we specify what it is doing by its motion and the way it is spinning.

The temperature in the interior of a star is always so high that the atoms are stripped of virtually all their electrons. For most of a star's life, the density of matter is also relatively low, and the electrons in the star are moving rapidly. This means that no two of them will be in the same place moving in exactly the same way at the same time. But this all changes when a star exhausts its store of nuclear energy and begins its final collapse.

As the star's core contracts, electrons are squeezed closer and closer together. Eventually, a star like the Sun becomes so dense that further contraction would in fact require two or more electrons to violate the rule against occupying the same place and moving in the same way. Such a dense gas is said to be degenerate (a term coined by physicists and not related to the electron's moral character). The electrons in a degenerate gas resist further crowding with tremendous pressure. (It's as if the electrons

said, “You can press inward all you want, but there is simply no room for any other electrons to squeeze in here without violating the rules of our existence.”)

The degenerate electrons do not require an input of heat to maintain the pressure they exert, and so a star with this kind of structure, if nothing disturbs it, can last essentially forever. (Note that the repulsive force between degenerate electrons is different from, and much stronger than, the normal electrical repulsion between charges that have the same sign.)

The electrons in a degenerate gas do move about, as do particles in any gas, but not with a lot of freedom. A particular electron cannot change position or momentum until another electron in an adjacent stage gets out of the way. The situation is much like that in the parking lot after a big football game. Vehicles are closely packed, and a given car cannot move until the one in front of it moves, leaving an empty space to be filled.

Of course, the dying star also has atomic nuclei in it, not just electrons, but it turns out that the nuclei must be squeezed to much higher densities before their quantum nature becomes apparent. As a result, in white dwarfs, the nuclei do not exhibit degeneracy pressure. Hence, in the white dwarf stage of stellar evolution, it is the degeneracy pressure of the electrons, and not of the nuclei, that halts the collapse of the core.

White Dwarfs

White dwarfs, then, are stable, compact objects with electron-degenerate cores that cannot contract any further. Calculations showing that white dwarfs are the likely end state of low-mass stars were first carried out by the Indian-American astrophysicist Subrahmanyan Chandrasekhar. He was able to show how much a star will shrink before the degenerate electrons halt its further contraction and hence what its final diameter will be (Figure 23.2).

When Chandrasekhar made his calculation about white dwarfs, he found something very surprising: the radius of a white dwarf shrinks as the mass in the star increases (the larger the mass, the more tightly packed the electrons can become, resulting in a smaller radius). According to the best theoretical models, a white dwarf with a mass of about 1.4 M_{Sun} or larger would have a radius of zero. What the calculations are telling us is that even the force of degenerate electrons cannot stop the collapse of a star with more mass than this. The maximum mass that a star can end its life with and still become a white dwarf—1.4 M_{Sun}—is called the Chandrasekhar limit. Stars with end-of-life masses that exceed this limit have a different kind of end in store—one that we will explore in the next section.

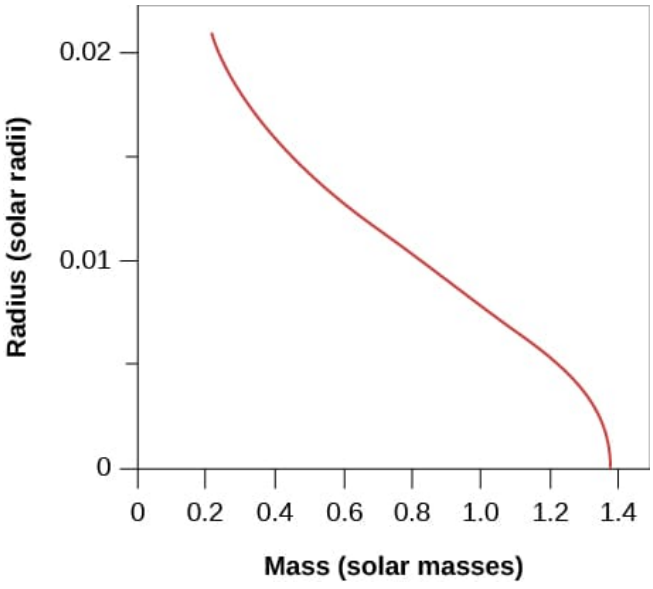


Figure 23.2 : Relating Masses and Radii of White Dwarfs. Models of white-dwarf structure predict that as the mass of the star increases (toward the right), its radius gets smaller and smaller.

Voyagers in Astronomy: Subrahmanyan Chandrasekhar

Born in 1910 in Lahore, India, Subrahmanyan Chandrasekhar (known as Chandra to his friends and colleagues) grew up in a home that encouraged scholarship and an interest in science (Figure 23.3). His uncle, C. V. Raman, was a physicist who won the 1930 Nobel Prize. A precocious student, Chandra tried to read as much as he could about the latest ideas in physics and astronomy, although obtaining technical books was not easy in India at the time. He finished college at age 19 and won a scholarship to study in England. It was during the long boat voyage to get to graduate school that he first began doing calculations about the structure of white dwarf stars.

Chandra developed his ideas during and after his studies as a graduate student, showing—as we have discussed—that white dwarfs with masses greater than 1.4 times the mass of the Sun cannot exist and that the theory predicts the existence of other kinds of stellar corpses. He wrote later that he felt very shy and lonely during this period, isolated from students, afraid to assert himself, and sometimes waiting for hours to speak with some of the famous professors he had read about in India. His calculations soon brought him into conflict with certain distinguished astronomers, including Sir Arthur Eddington, who publicly ridiculed Chandra's ideas. At a number of meetings of astronomers, such leaders in the field as Henry Norris Russell refused to give Chandra the opportunity to defend his ideas, while allowing his more senior critics lots of time to criticize them.

Yet Chandra persevered, writing books and articles elucidating his theories, which turned out not only to be correct, but to lay the foundation for much of our modern understanding of the death of stars. In 1983, he received the Nobel Prize in physics for this early work.

In 1937, Chandra came to the United States and joined the faculty at the University of Chicago, where he remained for the rest of his life. There he devoted himself to research and teaching, making major contributions to many fields of astronomy, from our understanding of the motions of stars through the Galaxy to the behavior of the bizarre objects called black holes (see Black Holes and Curved Spacetime). In 1999, NASA named its sophisticated orbiting X-ray telescope (designed in part to explore such stellar corpses) the Chandra X-ray Observatory.



Figure 23.3 : S. Chandrasekhar (1910–1995). Chandra's research provided the basis for much of what we now know about stellar corpses. (credit: modification of work by American Institute of Physics)

Chandra spent a great deal of time with his graduate students, supervising the research of more than 50 PhDs during his life. He took his teaching responsibilities very seriously: during the 1940s, while based at the Yerkes Observatory, he willingly drove the more than 100-mile trip to the university each week to teach a class of only a few students.

Chandra also had a deep devotion to music, art, and philosophy, writing articles and books about the relationship between the humanities and science. He once wrote that "one can learn science the way one enjoys music or art. . . . Heisenberg had a marvelous phrase 'shuddering before the beautiful'. . . that is the kind of feeling I have."

Link to Learning

Using the Hubble Space Telescope, astronomers were able to detect images of faint white dwarf stars and other “stellar corpses” in the M4 star cluster, located about 7200 light-years away.

The Ultimate Fate of White Dwarfs

If the birth of a main-sequence star is defined by the onset of fusion reactions, then we must consider the end of all fusion reactions to be the time of a star’s death. As the core is stabilized by degeneracy pressure, a last shudder of fusion passes through the outside of the star, consuming the little hydrogen still remaining. Now the star is a true white dwarf: nuclear fusion in its interior has ceased. Figure 23.4 shows the path of a star like the Sun on the H–R diagram during its final stages.

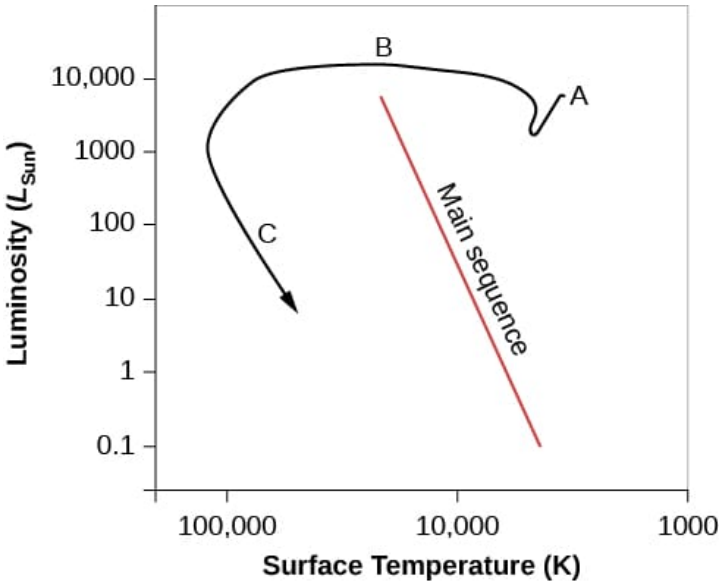
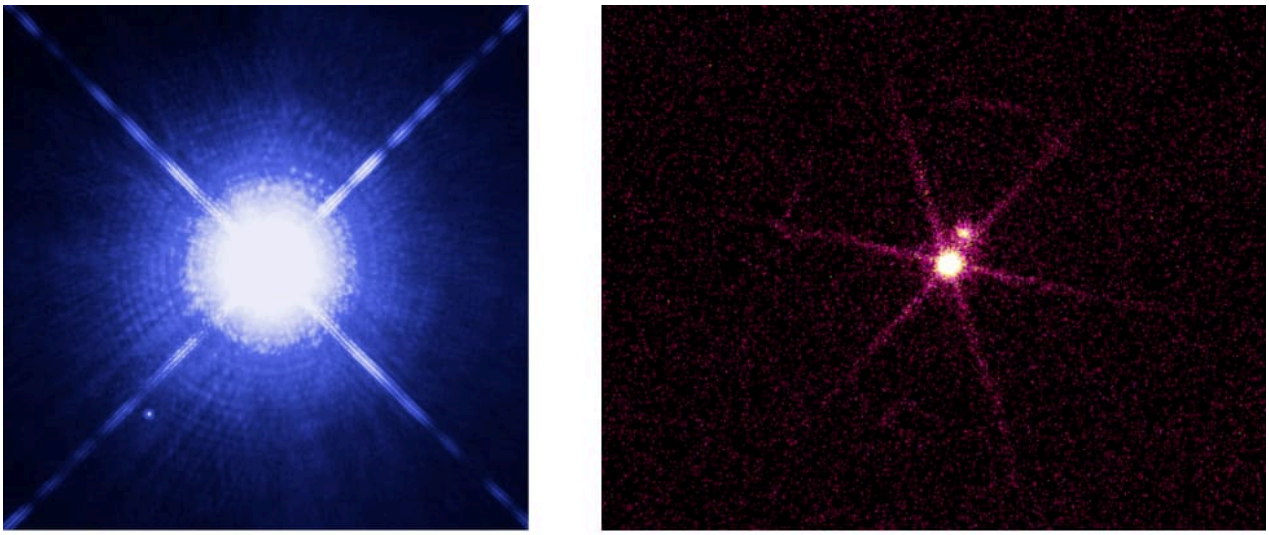


Figure 23.4 : Evolutionary Track for a Star Like the Sun. This diagram shows the changes in luminosity and surface temperature for a star with a mass like the Sun’s as it nears the end of its life. After the star becomes a giant again (point A on the diagram), it will lose more and more mass as its core begins to collapse. The mass loss will expose the hot inner core, which will appear at the center of a planetary nebula. In this stage, the star moves across the diagram to the left as it becomes hotter and hotter during its collapse (point B). At first, the luminosity remains nearly constant, but as the star begins to cool off, it becomes less and less bright (point C). It is now a white dwarf and will continue to cool slowly for billions of years until all of its remaining store of energy is radiated away. (This assumes the Sun will lose between 46–50% of its mass during the giant stages, based upon various theoretical models).

Since a stable white dwarf can no longer contract or produce energy through fusion, its only energy source is the heat represented by the motions of the atomic nuclei in its interior. The light it emits comes from this internal stored heat, which is substantial. Gradually, however, the white dwarf radiates away all its heat into space. After many billions of years, the nuclei will be moving much more slowly, and the white dwarf will no longer shine (Figure 23.5). It will then be a black dwarf—a cold stellar corpse with the mass of a star and the size of a planet. It will be composed mostly of carbon, oxygen, and neon, the products of the most advanced fusion reactions of which the star was capable.



(a) (b)

Figure 23.5 : Visible Light and X-Ray Images of the Sirius Star System. (a) This image taken by the Hubble Space Telescope shows Sirius A (the large bright star), and its companion star, the white dwarf known as Sirius B (the tiny, faint star at the lower left). Sirius A and B are 8.6 light-years from Earth and are our fifth-closest star system. Note that the image has intentionally been overexposed to allow us to see Sirius B. (b) The same system is shown in X-ray taken with the Chandra Space Telescope. Note that Sirius A is fainter in X-rays than the hot white dwarf that is Sirius B. (credit a: modification of work by NASA, ESA, H. Bond, M. Barstow(University of Leicester); credit b: modification of work by NASA/SAO/CXC)

We have one final surprise as we leave our low-mass star in the stellar graveyard. Calculations show that as a degenerate star cools, the atoms inside it in essence "solidify" into a giant, highly compact lattice (organized rows of atoms, just like in a crystal). When carbon is compressed and crystallized in this way, it becomes a giant *diamond-like* star. A white dwarf star might make the most impressive engagement present you could ever see, although any attempt to mine the diamond-like material inside would crush an ardent lover instantly!

Link to Learning

Learn about a recent "diamond star" find, a cold, white dwarf star detected in 2014, which is considered the coldest and dimmest found to date, at the website of the National Radio Astronomy Observatory.

Evidence That Stars Can Shed a Lot of Mass as They Evolve

Whether or not a star will become a white dwarf depends on how much mass is lost in the red-giant and earlier phases of evolution. All stars that have masses below the Chandrasekhar limit when they run out of fuel will become white dwarfs, no matter what mass they were born with. But which stars shed enough mass to reach this limit?

One strategy for answering this question is to look in young, open clusters (which were discussed in Star Clusters). The basic idea is to search for young clusters that contain one or more white dwarf stars. Remember that more massive stars go through all stages of their evolution more rapidly than less massive ones. Suppose we find a cluster that has a white dwarf member and also contains stars on the main sequence that have 6 times the mass of the Sun. This means that only those stars with masses greater than 6 M_{Sun} have had time to exhaust their supply of nuclear energy and complete their evolution to the white dwarf stage. The star that turned into the white dwarf must therefore have had a main-sequence mass of more than 6 M_{Sun}, since stars with lower masses have not yet had time to use up their stores of nuclear energy. The star that became the white dwarf must, therefore, have gotten rid of at least 4.6 M_{Sun} so that its mass at the time nuclear energy generation ceased could be less than 1.4 M_{Sun}.

Astronomers continue to search for suitable clusters to make this test, and the evidence so far suggests that stars with masses up to about 8 M_{Sun} can shed enough mass to end their lives as white dwarfs. Stars like the Sun will probably lose about 45% of their initial mass and become white dwarfs with masses less than 1.4 M_{Sun}.

This page titled 5.3.4: The Death of Low-Mass Stars is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by OpenStax via source content that was edited to the style and standards of the LibreTexts platform.

5.3.5: Evolution of Massive Stars-An Explosive Finish

Learning Objectives

By the end of this section, you will be able to:

- Describe the interior of a massive star before a supernova
- Explain the steps of a core collapse and explosion
- List the hazards associated with nearby supernovae

Thanks to mass loss, then, stars with starting masses of at least 8 M_{Sun} (and perhaps even more) probably end their lives as white dwarfs. But we know stars can have masses as large as 150 (or more) M_{Sun}. They have a different kind of death in store for them. As we will see, these stars die with a bang.

Nuclear Fusion of Heavy Elements

After the helium in its core is exhausted (see The Evolution of More Massive Stars), the evolution of a massive star takes a significantly different course from that of lower-mass stars. In a massive star, the weight of the outer layers is sufficient to force the carbon core to contract until it becomes hot enough to fuse carbon into oxygen, neon, and magnesium. This cycle of contraction, heating, and the ignition of another nuclear fuel repeats several more times. After each of the possible nuclear fuels is exhausted, the core contracts again until it reaches a new temperature high enough to fuse still-heavier nuclei. The products of carbon fusion can be further converted into silicon, sulfur, calcium, and argon. And these elements, when heated to a still-higher temperature, can combine to produce iron. Massive stars go through these stages very, very quickly. In really massive stars, some fusion stages toward the very end can take only months or even days! This is a far cry from the millions of years they spend in the main-sequence stage.

At this stage of its evolution, a massive star resembles an onion with an iron core. As we get farther from the center, we find shells of decreasing temperature in which nuclear reactions involve nuclei of progressively lower mass—silicon and sulfur, oxygen, neon, carbon, helium, and finally, hydrogen (Figure 23.6).

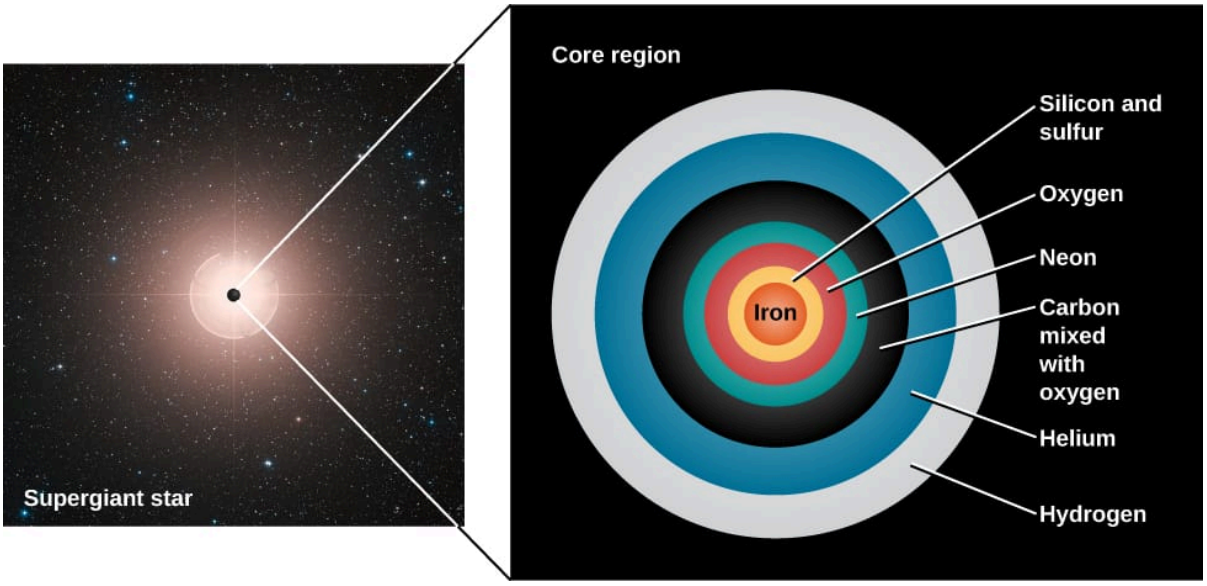


Figure 23.6 : Structure of an Old Massive Star. Just before its final gravitational collapse, the core of a massive star resembles an onion. The iron core is surrounded by layers of silicon and sulfur, oxygen, neon, carbon mixed with some oxygen, helium, and finally hydrogen. Outside the core, the composition is mainly hydrogen and helium. (Note that this diagram is not precisely to scale but is just meant to convey the general idea of what such a star would be like.) (credit: modification of work by ESO, Digitized Sky Survey)

But there is a limit to how long this process of building up elements by fusion can go on. The fusion of silicon into iron turns out to be the last step in the sequence of nonexplosive element production. Up to this point, each fusion reaction has *produced* energy because the nucleus of each fusion product has been a bit more stable than the nuclei that formed it. As discussed in The Sun: A Nuclear Powerhouse, light nuclei give up some of their binding energy in the process of fusing into more tightly bound, heavier nuclei. It is this released energy that maintains the outward pressure in the core so that the star does not collapse. But of all the nuclei known, iron is the most tightly bound and thus the most stable.

You might think of the situation like this: all smaller nuclei want to "grow up" to be like iron, and they are willing to pay (*produce* energy) to move toward that goal. But iron is a mature nucleus with good self-esteem, perfectly content being iron; it requires payment (must *absorb* energy) to change its stable nuclear structure. This is the exact opposite of what has happened in each nuclear reaction so far: instead of *providing* energy to balance the inward pull of gravity, any nuclear reactions involving iron would *remove* some energy from the core of the star.

Unable to generate energy, the star now faces catastrophe.

Collapse into a Ball of Neutrons

When nuclear reactions stop, the core of a massive star is supported by degenerate electrons, just as a white dwarf is. For stars that begin their evolution with masses of at least 10 M_{Sun}, this core is likely made mainly of iron. (For stars with initial masses in the range 8 to 10 M_{Sun}, the core is likely made of oxygen, neon, and magnesium, because the star never gets hot enough to form elements as heavy as iron. The exact composition of the cores of stars in this mass range is very difficult to determine because of the complex physical characteristics in the cores, particularly at the very high densities and temperatures involved.) We will focus on the more massive iron cores in our discussion.

While no energy is being generated within the white dwarf core of the star, fusion still occurs in the shells that surround the core. As the shells finish their fusion reactions and stop producing energy, the ashes of the last reaction fall onto the white dwarf core, increasing its mass. As Figure 23.2 shows, a higher mass means a smaller core. The core can contract because even a degenerate gas is still mostly empty space. Electrons and atomic nuclei are, after all, extremely small. The electrons and nuclei in a stellar core may be crowded compared to the air in your room, but there is still lots of space between them.

The electrons at first resist being crowded closer together, and so the core shrinks only a small amount. Ultimately, however, the iron core reaches a mass so large that even degenerate electrons can no longer support it. When the density reaches 4×10^{11} g/cm^3 (400 billion times the density of water), some electrons are actually squeezed into the atomic nuclei, where they combine with protons to form neutrons and neutrinos. This transformation is not something that is familiar from everyday life, but becomes very important as such a massive star core collapses.

Some of the electrons are now gone, so the core can no longer resist the crushing mass of the star's overlying layers. The core begins to shrink rapidly. More and more electrons are now pushed into the atomic nuclei, which ultimately become so saturated with neutrons that they cannot hold onto them.

At this point, the neutrons are squeezed out of the nuclei and can exert a new force. As is true for electrons, it turns out that the neutrons strongly resist being in the same place and moving in the same way. The force that can be exerted by such *degenerate neutrons* is much greater than that produced by degenerate electrons, so unless the core is too massive, they can ultimately stop the collapse.

This means the collapsing core can reach a stable state as a crushed ball made mainly of neutrons, which astronomers call a neutron star. We don't have an exact number (a "Chandrasekhar limit") for the maximum mass of a neutron star, but calculations tell us that the upper mass limit of a body made of neutrons might only be about 3 M_{Sun}. So if the mass of the core were greater than this, then even neutron degeneracy would not be able to stop the core from collapsing further. The dying star must end up as something even more extremely compressed, which until recently was believed to be only one possible type of object—the state of ultimate compaction known as a black hole (which is the subject of our next chapter). This is because no force was believed to exist that could stop a collapse beyond the neutron star stage.

Collapse and Explosion

When the collapse of a high-mass star's core is stopped by degenerate neutrons, the core is saved from further destruction, but

it turns out that the rest of the star is literally blown apart. Here's how it happens.

The collapse that takes place when electrons are absorbed into the nuclei is very rapid. In less than a second, a core with a mass of about 1 M_{Sun}, which originally was approximately the size of Earth, collapses to a diameter of less than 20 kilometers. The speed with which material falls inward reaches one-fourth the speed of light. The collapse halts only when the density of the core exceeds the density of an atomic nucleus (which is the densest form of matter we know). A typical neutron star is so compressed that to duplicate its density, we would have to squeeze all the people in the world into a single sugar cube! This would give us one sugar cube's worth (one cubic centimeter's worth) of a neutron star.

The neutron degenerate core strongly resists further compression, abruptly halting the collapse. The shock of the sudden jolt initiates a shock wave that starts to propagate outward. However, this shock alone is not enough to create a star explosion. The energy produced by the outflowing matter is quickly absorbed by atomic nuclei in the dense, overlying layers of gas, where it breaks up the nuclei into individual neutrons and protons.

Our understanding of nuclear processes indicates (as we mentioned above) that each time an electron and a proton in the star's core merge to make a neutron, the merger releases a *neutrino*. These ghostly subatomic particles, introduced in The Sun: A Nuclear Powerhouse, carry away some of the nuclear energy. It is their presence that launches the final disastrous explosion of the star. The total energy contained in the neutrinos is huge. In the initial second of the star's explosion, the power carried by the neutrinos (10^{46} watts) is greater than the power put out by all the stars in over a billion galaxies.

While neutrinos ordinarily do not interact very much with ordinary matter (we earlier accused them of being downright antisocial), matter near the center of a collapsing star is so dense that the neutrinos do interact with it to some degree. They deposit some of this energy in the layers of the star just outside the core. This huge, sudden input of energy reverses the infall of these layers and drives them explosively outward. Most of the mass of the star (apart from that which went into the neutron star in the core) is then ejected outward into space. As we saw earlier, such an explosion requires a star of at least 8 M_{Sun}, and the neutron star can have a mass of at most 3 M_{Sun}. Consequently, at least five times the mass of our Sun is ejected into space in each such explosive event!

The resulting explosion is called a supernova (Figure 23.7). When these explosions happen close by, they can be among the most spectacular celestial events, as we will discuss in the next section. (Actually, there are at least two different types of supernova explosions: the kind we have been describing, which is the collapse of a massive star, is called, for historical reasons, a type II supernova. We will describe how the types differ later in this chapter).



Figure 23.7 : Five Supernova Explosions in Other Galaxies. The arrows in the top row of images point to the supernovae. The bottom row shows the host galaxies before or after the stars exploded. Each of these supernovae exploded between 3.5 and 10 billion years ago. Note that the supernovae when they first explode can be as bright as an entire galaxy. (credit: modification of work by NASA, ESA, and A. Riess (STScI))

Table 23.1 summarizes the discussion so far about what happens to stars and substellar objects of different initial masses at the

ends of their lives. Like so much of our scientific understanding, this list represents a progress report: it is the best we can do with our present models and observations. The mass limits corresponding to various outcomes may change somewhat as models are improved. There is much we do not yet understand about the details of what happens when stars die.

Table 23.1: The Ultimate Fate of Stars and Substellar Objects with Different Masses

Initial Mass (Mass of Sun = 1)[1]	Final State at the End of Its Life
< 0.01	Planet
0.01 to 0.08	Brown dwarf
0.08 to 0.25	White dwarf made mostly of helium
0.25 to 8	White dwarf made mostly of carbon and oxygen
8 to 10	White dwarf made of oxygen, neon, and magnesium
10 to 40	Supernova explosion that leaves a neutron star
> 40	Supernova explosion that leaves a black hole

The Supernova Giveth and the Supernova Taketh Away

After the supernova explosion, the life of a massive star comes to an end. But the death of each massive star is an important event in the history of its galaxy. The elements built up by fusion during the star's life are now "recycled" into space by the explosion, making them available to enrich the gas and dust that form new stars and planets. Because these heavy elements ejected by supernovae are critical for the formation of planets and the origin of life, it's fair to say that without mass loss from supernovae and planetary nebulae, neither the authors nor the readers of this book would exist.[2]

But the supernova explosion has one more creative contribution to make, one we alluded to in Stars from Adolescence to Old Age when we asked where the atoms in your jewelry came from. The supernova explosion produces a flood of energetic neutrons that barrel through the expanding material. These neutrons can be absorbed by iron and other nuclei where they can turn into protons. Thus, they can build up elements that are more massive than iron, possibly including such terrestrial favorites as gold, silver and uranium. Supernovae (and, as we will shortly see, the explosive mergers of neutron stars) are the only candidates we have for places where such heavier atoms can be made. Next time you wear some gold jewelry (or give some to your sweetheart), bear in mind that those gold atoms were forged long ago in these kinds of celestial explosions!

When supernovae explode, these elements (as well as the ones the star made during more stable times) are ejected into the existing gas between the stars and mixed with it. Thus, supernovae play a crucial role in enriching their galaxy with heavier elements, allowing, among other things, the chemical elements that make up earthlike planets and the building blocks of life to become more common as time goes on (Figure 23.8).

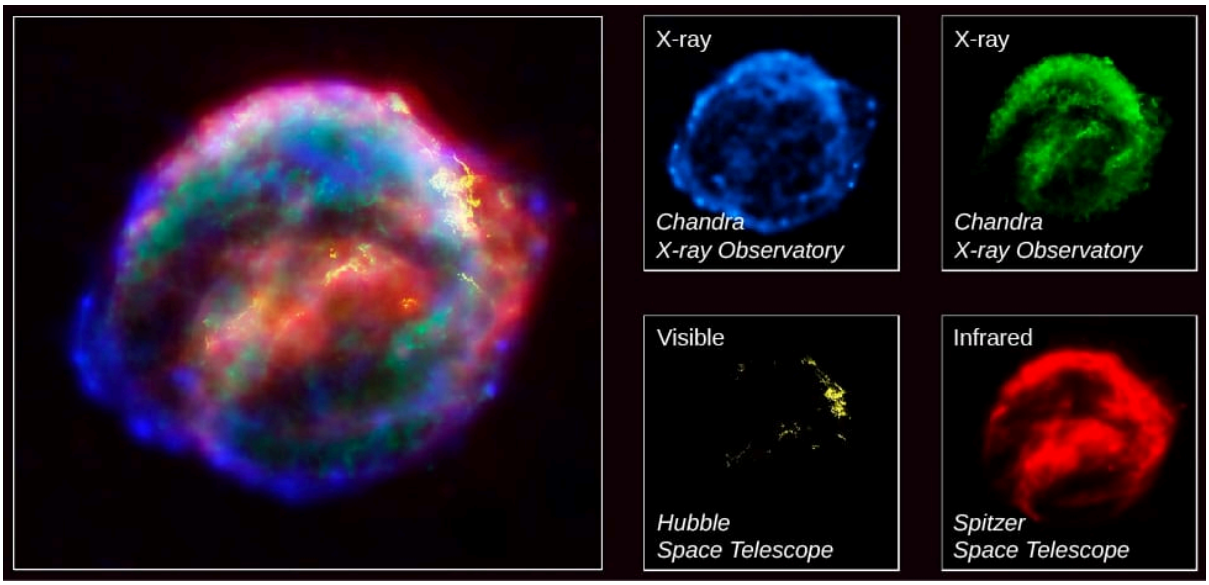


Figure 23.8 : Kepler Supernova Remant. This image shows the expanding remains of a supernova explosion, which was first seen about 400 years ago by sky watchers, including the famous astronomer Johannes Kepler. The bubble-shaped shroud of gas and dust is now 14 light-years wide and is expanding at 2,000 kilometers per second (4 million miles per hour). The remnant emits energy at wavelengths from X-rays (shown in blue and green) to visible light (yellow) and into the infrared (red). The expanding shell is rich in iron, which was produced in the star that exploded. The main image combines the individual single-color images seen at the bottom into one multi-wavelength picture. (credit: modification of work by NASA, ESA, R. Sankrit and W. Blair (Johns Hopkins University))

Supernovae are also thought to be the source of many of the high-energy *cosmic ray* particles discussed in Cosmic Rays. Trapped by the magnetic field of the Galaxy, the particles from exploded stars continue to circulate around the vast spiral of the Milky Way. Scientists speculate that high-speed cosmic rays hitting the genetic material of Earth organisms over billions of years may have contributed to the steady *mutations*—subtle changes in the genetic code—that drive the evolution of life on our planet. In all the ways we have mentioned, supernovae have played a part in the development of new generations of stars, planets, and life.

But supernovae also have a dark side. Suppose a life form has the misfortune to develop around a star that happens to lie near a massive star destined to become a supernova. Such life forms may find themselves snuffed out when the harsh radiation and high-energy particles from the neighboring star's explosion reach their world. If, as some astronomers speculate, life can develop on many planets around long-lived (lower-mass) stars, then the suitability of that life's *own star* and planet may not be all that matters for its long-term evolution and survival. Life may well have formed around a number of pleasantly stable stars only to be wiped out because a massive nearby star suddenly went supernova. Just as children born in a war zone may find themselves the unjust victims of their violent neighborhood, life too close to a star that goes supernova may fall prey to having been born in the wrong place at the wrong time.

What is a safe distance to be from a supernova explosion? A lot depends on the violence of the particular explosion, what type of supernova it is (see The Evolution of Binary Star Systems), and what level of destruction we are willing to accept. Calculations suggest that a supernova less than 50 light-years away from us would certainly end all life on Earth, and that even one 100 light-years away would have drastic consequences for the radiation levels here. One minor extinction of sea creatures about 2 million years ago on Earth may actually have been caused by a supernova at a distance of about 120 light-years.

The good news is that there are at present no massive stars that promise to become supernovae within 50 light-years of the Sun. (This is in part because the kinds of massive stars that become supernovae are overall quite rare.) The massive star closest to us, Spica (in the constellation of Virgo), is about 260 light-years away, probably a safe distance, even if it were to explode as a supernova in the near future.

Example 23.1: Extreme Gravity

In this section, you were introduced to some very dense objects. How would those objects' gravity affect you? Recall that the force of gravity, F, between two bodies is calculated as

F=GM1M2R2F=GM1M2R2

where G is the gravitational constant, 6.67×10-11 Nm2/kg2,M1 and M26.67×10-11 Nm2/kg2,M1 and M2 are the masses of the two bodies, and R is their separation. Also, from Newton's second law,

F=M×aF=M×a

where a is the acceleration of a body with mass M.

So let's consider the situation of a mass—say, you—standing on a body, such as Earth or a white dwarf (where we assume you will be wearing a heat-proof space suit). You are M_1 and the body you are standing on is M_2. The distance between you and the center of gravity of the body on which you stand is its radius, R. The force exerted on you is

F=M1×a=GM1M2/R2F=M1×a=GM1M2/R2

Solving for a, the acceleration of gravity on that world, we get

g=(G×M)R2g=(G×M)R2

Note that we have replaced the general symbol for acceleration, a, with the symbol scientists use for the acceleration of gravity, g.

Say that a particular white dwarf has the mass of the Sun (2×1030 kg)(2×1030 kg) but the radius of Earth (6.4×106 m)(6.4×106 m). What is the acceleration of gravity at the surface of the white dwarf?

Solution

The acceleration of gravity at the surface of the white dwarf is

g(white dwarf)=(G×MSun)REarth2=(6.67×10−11m2/kg s2×2×1030kg)(6.4×106m)2=3.26×106m/s2g(white dwarf)=(G×MSun)REarth2=(6.67×10−11m2/kg s2×2×1030kg)(6.4×106m)2=3.26×106m/s2

Compare this to g on the surface of Earth, which is 9.8 m/s^2.

Exercise 5.3.5.1

What is the acceleration of gravity at the surface if the white dwarf has the twice the mass of the Sun and is only half the radius of Earth?

Answer available at the end of this section.

Footnotes

- 1Stars in the mass ranges 0.25–8 and 8–10 may later produce a type of supernova different from the one we have discussed so far. These are discussed in The Evolution of Binary Star Systems.
- 2To tell the full story, there is another way to make and recycle these heavier elements which we will discuss in The Milky Way Galaxy. That is the merger of two neutron stars, which also causes a huge explosion.

Answers

Exercise 5.3.5.1

g(white dwarf)=(G×2MSun)(0.5REarth)2=(6.67×10−11m2/kg s2×4×1030kg)(3.2×106)2=2.61×107m/s2g(white dwarf)=(G×2MSun)(0.5REarth)2=(6.67×10−11m2/kg s2×4×1030kg)(3.2×106)2=2.61×107m/s2

This page titled 5.3.5: Evolution of Massive Stars-An Explosive Finish is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by OpenStax via source content that was edited to the style and standards of the LibreTexts platform.

6: GALAXIES AND THE UNIVERSE

6.1: The Architecture of the Galaxy

6.2: Types of Galaxies

6.3: The Expanding Universe

6.4: The Beginning of the Universe

6.5: The Cosmic Microwave Background

Chapter 6: Galaxies and the Universe — Introduction

Stars do not exist in isolation. They are members of enormous systems called galaxies, and galaxies themselves are part of the larger structure of the universe. Our Sun is just one of the many stars in the Milky Way Galaxy, so understanding our own galaxy provides a starting point for exploring the universe on its largest scales.

In this chapter, we will examine the structure of the Milky Way and compare it with other types of galaxies. We will then look beyond individual galaxies to one of the most important discoveries of modern astronomy: **the universe is expanding**. By observing the motions of distant galaxies, astronomers can investigate how the universe has changed over time and trace its history back toward a much hotter and denser beginning.

Finally, we will examine the **cosmic microwave background**, radiation left over from the early universe. Together, the expansion of the universe and the cosmic microwave background provide important observational evidence that allows scientists to reconstruct the history of the cosmos.

This page titled 6: Galaxies and the Universe is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Robert Wagner.

6.1: The Architecture of the Galaxy

Learning Objectives

By the end of this section, you will be able to:

- Explain why William and Caroline Herschel concluded that the Milky Way has a flattened structure centered on the Sun and solar system
- Describe the challenges of determining the Galaxy's structure from our vantage point within it
- Identify the main components of the Galaxy

The Milky Way Galaxy surrounds us, and you might think it is easy to study because it is so close. However, the very fact that we are embedded within it presents a difficult challenge. Suppose you were given the task of mapping New York City. You could do a much better job from a helicopter flying over the city than you could if you were standing in Times Square. Similarly, it would be easier to map our Galaxy if we could only get a little way outside it, but instead we are trapped inside and way out in its suburbs—far from the galactic equivalent of Times Square.

Herschel Measures the Galaxy

In 1785, William Herschel (Figure 25.2) made the first important discovery about the architecture of the Milky Way Galaxy. Using a large reflecting telescope that he had built, William and his sister Caroline counted stars in different directions of the sky. They found that most of the stars they could see lay in a flattened structure encircling the sky, and that the numbers of stars were about the same in any direction around this structure. Herschel therefore concluded that the stellar system to which the Sun belongs has the shape of a disk or wheel (he might have called it a Frisbee except Frisbees hadn't been invented yet), and that the Sun must be near the hub of the wheel (Figure 25.3).



Figure 25.2 : William Herschel (1738–1822) and Caroline Herschel (1750–1848). William Herschel was a German musician who emigrated to England and took up astronomy in his spare time. He discovered the planet Uranus, built several large telescopes, and made measurements of the Sun's place in the Galaxy, the Sun's motion through space, and the comparative brightnesses of stars. This painting shows William and his sister Caroline polishing a telescope lens. (credit: modification of work by the Wellcome Library)

To understand why Herschel reached this conclusion, imagine that you are a member of a band standing in formation during halftime at a football game. If you count the band members you see in different directions and get about the same number each time, you can conclude that the band has arranged itself in a circular pattern with you at the center. Since you see no band members above you or underground, you know that the circle made by the band is much flatter than it is wide.

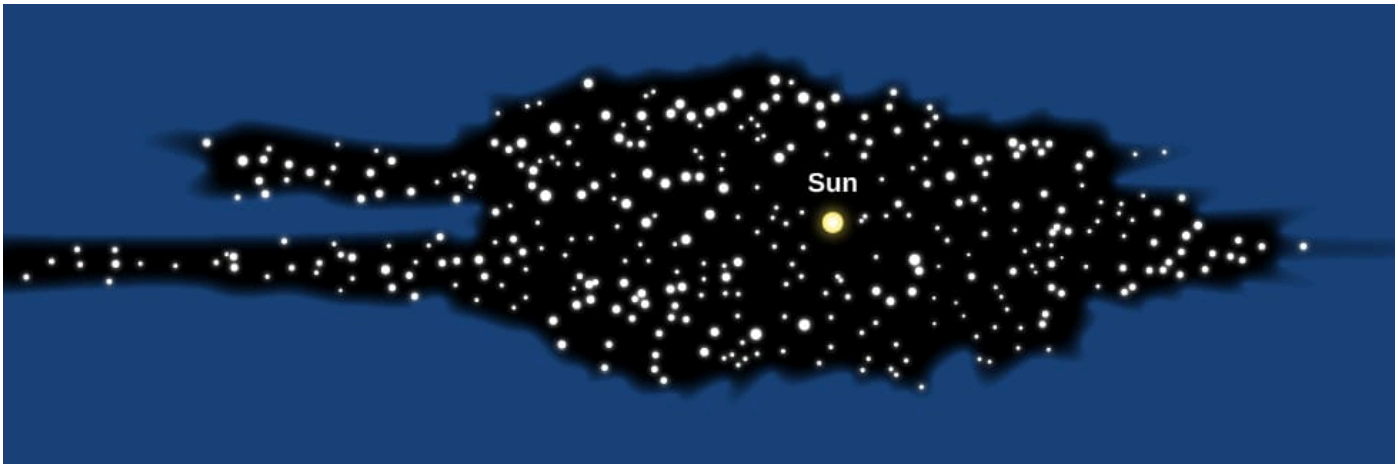


Figure 25.3 : Herschel's Diagram of the Milky Way. Herschel constructed this cross section of the Galaxy by counting stars in various directions.

We now know that Herschel was right about the shape of our system, but wrong about where the Sun lies within the disk. As we saw in Between the Stars: Gas and Dust in Space, we live in a dusty Galaxy. Because interstellar dust absorbs the light from stars, Herschel could see only those stars within about 6000 light-years of the Sun. Today we know that this is a very small section of the entire 100,000-light-year-diameter disk of stars that makes up the Galaxy.

Voyagers in Astronomy: Harlow Shapley: Mapmaker to the Stars

Until the early 1900s, astronomers generally accepted Herschel's conclusion that the Sun is near the center of the Galaxy. The discovery of the Galaxy's true size and our actual location came about largely through the efforts of Harlow Shapley. In 1917, he was studying RR Lyrae variable stars in globular clusters. By comparing the known intrinsic luminosity of these stars to how bright they appeared, Shapley could calculate how far away they are. (Recall that it is distance that makes the stars look dimmer than they would be "up close," and that the brightness fades as the distance squared.) Knowing the distance to any star in a cluster then tells us the distance to the cluster itself.

Globular clusters can be found in regions that are free of interstellar dust and so can be seen at very large distances. When Shapley used the distances and directions of 93 globular clusters to map out their positions in space, he found that the clusters are distributed in a spherical volume, which has its center not at the Sun but at a distant point along the Milky Way in the direction of Sagittarius. Shapley then made the bold assumption, verified by many other observations since then, that the point on which the system of globular clusters is centered is also the center of the entire Galaxy (Figure 25.4).

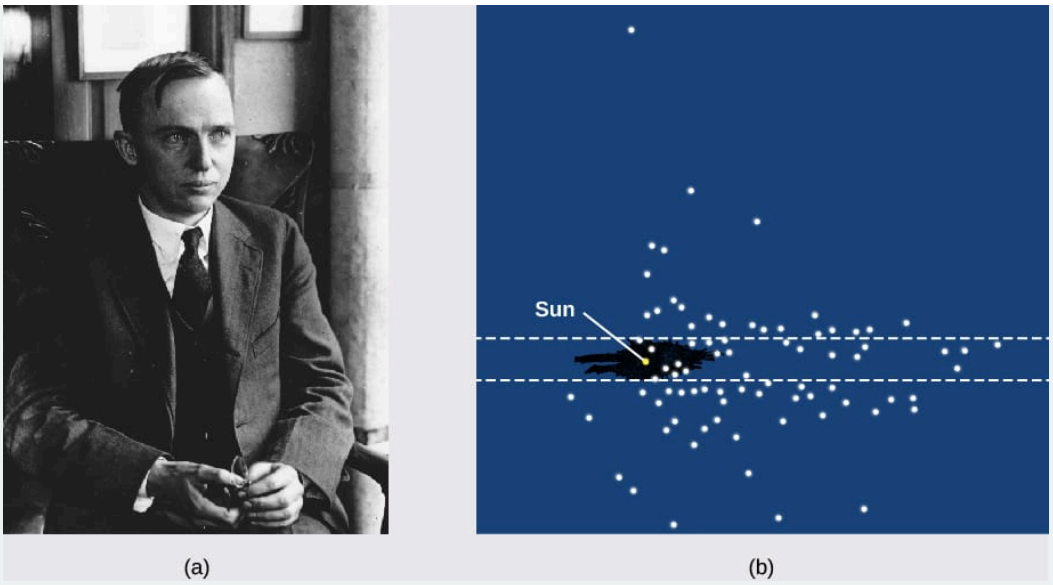


Figure 25.4 : Harlow Shapley and His Diagram of the Milky Way. (a) Shapley poses for a formal portrait. (b) His diagram shows the location of globular clusters, with the position of the Sun also marked. The black area shows Herschel's old diagram, centered on the Sun, approximately to scale.

Shapley's work showed once and for all that our star has no special place in the Galaxy. We are in a nondescript region of the Milky Way, only one of 200 to 400 billion stars that circle the distant center of our Galaxy.

Born in 1885 on a farm in Missouri, Harlow Shapley at first dropped out of school with the equivalent of only a fifth-grade education. He studied at home and at age 16 got a job as a newspaper reporter covering crime stories. Frustrated by the lack of opportunities for someone who had not finished high school, Shapley went back and completed a six-year high-school program in only two years, graduating as class valedictorian.

In 1907, at age 22, he went to the University of Missouri, intent on studying journalism, but found that the school of journalism would not open for a year. Leafing through the college catalog (or so he told the story later), he chanced to see "Astronomy" among the subjects beginning with "A." Recalling his boyhood interest in the stars, he decided to study astronomy for the next year (and the rest, as the saying goes, is history).

Upon graduation Shapley received a fellowship for graduate study at Princeton and began to work with the brilliant Henry Norris Russell (see the Henry Norris Russell feature box). For his PhD thesis, Shapley made major contributions to the methods of analyzing the behavior of eclipsing binary stars. He was also able to show that cepheid variable stars are not binary systems, as some people thought at the time, but individual stars that pulsate with striking regularity.

Impressed with Shapley's work, George Ellery Hale offered him a position at the Mount Wilson Observatory, where the young man took advantage of the clear mountain air and the 60-inch reflector to do his pioneering study of variable stars in globular clusters.

Shapley subsequently accepted the directorship of the Harvard College Observatory, and over the next 30 years, he and his collaborators made contributions to many fields of astronomy, including the study of neighboring galaxies, the discovery of dwarf galaxies, a survey of the distribution of galaxies in the universe, and much more. He wrote a series of nontechnical books and articles and became known as one of the most effective popularizers of astronomy. Shapley enjoyed giving lectures around the country, including at many smaller colleges where students and faculty rarely got to interact with scientists of his caliber.

During World War II, Shapley helped rescue many scientists and their families from Eastern Europe; later, he helped found UNESCO, the United Nations Educational, Scientific, and Cultural Organization. He wrote a pamphlet called *Science from Shipboard* for men and women in the armed services who had to spend many weeks on board transport ships to Europe. And during the difficult period of the 1950s, when congressional committees began their "witch hunts" for communist

sympathizers (including such liberal leaders as Shapley), he spoke out forcefully and fearlessly in defense of the freedom of thought and expression. A man of many interests, he was fascinated by the behavior of ants, and wrote scientific papers about them as well as about galaxies.

By the time he died in 1972, Shapley was acknowledged as one of the pivotal figures of modern astronomy, a "twentieth-century Copernicus" who mapped the Milky Way and showed us our place in the Galaxy.

Link to Learning

To find more information about Shapley's life and work, see the entry for him on the Bruce Medalists website. (This site features the winners of the Bruce Medal of the Astronomical Society of the Pacific, one of the highest honors in astronomy; the list is a who's who of some of the greatest astronomers of the last twelve decades.)

Disks and Haloes

With modern instruments, astronomers can now penetrate the "smog" of the Milky Way by studying radio and infrared emissions from distant parts of the Galaxy. Measurements at these wavelengths (as well as observations of other galaxies like ours) have given us a good idea of what the Milky Way would look like if we *could* observe it from a distance.

Figure 25.5 sketches what we would see if we could view the Galaxy face-on and edge-on. The brightest part of the Galaxy consists of a thin, circular, rotating disk of stars distributed across a region about 100,000 light-years in diameter and about 2000 light-years thick. (Given how thin the disk is, perhaps a CD is a more appropriate analogy than a wheel.) The very youngest stars, and the dust and gas from which stars form, are found typically within 100 light-years of the plane of the Milky Way Galaxy. The mass of the interstellar matter is about 15% of the mass of the stars in this disk.

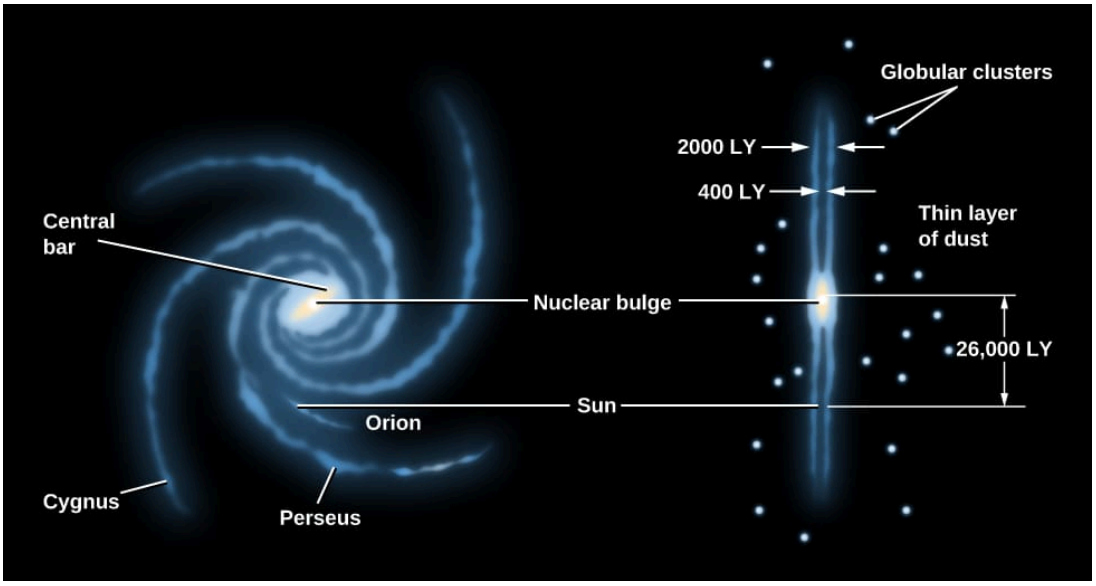


Figure 25.5 : Schematic Representation of the Galaxy. The left image shows the face-on view of the spiral disk; the right image shows the view looking edge-on along the disk. The major spiral arms are labeled. The Sun is located on the inside edge of the short Orion spur.

As the diagram in Figure 25.5 shows, the stars, gas, and dust are not spread evenly throughout the disk but are concentrated into a central bar and a series of spiral arms. Recent infrared observations have confirmed that the central bar is composed mostly of old yellow-red stars. The two main spiral arms appear to connect with the ends of the bar. They are highlighted by the blue light from young hot stars. We know many other spiral galaxies that also have bar-shaped concentrations of stars in their central regions; for that reason they are called *barred spirals*. Figure 25.6 shows two other galaxies—one without a bar and one with a strong bar—to give you a basis for comparison to our own. We will describe our spiral structure in more detail shortly. The Sun is located about halfway between the center of the Galaxy and the edge of the disk and only about 70 light-years above its central plane.



(a)



(b)

Figure 25.6 : Unbarred and Barred Spiral Galaxies. (a) This image shows the unbarred spiral galaxy M74. It contains a small central bulge of mostly old yellow-red stars, along with spiral arms that are highlighted with the blue light from young hot stars. (b) This image shows the strongly barred spiral galaxy NGC 1365. The bulge and the fainter bar both appear yellowish because the brightest stars in them are mostly old yellow and red giants. Two main spiral arms project from the ends of the bar. As in M74, these spiral arms are populated with blue stars and red patches of glowing gas—hallmarks of recent star formation. The Milky Way Galaxy is thought to have a barred spiral structure that is intermediate between these two examples. (credit a: modification of work by ESO/PESSTO/ S. Smartt; credit b: modification of work by ESO)

Our thin disk of young stars, gas, and dust is embedded in a thicker but more diffuse disk of older stars; this thicker disk extends about 1000 light-years above and 1000 light-years below the midplane of the thin disk and contains only about 5% as much mass as the thin disk. The stars thin out with distance from the galactic plane and don't have a sharp edge. Approximately 2/3 of the stars in the thick disk are within 1000 light-years of midplane.

Close in to the galactic center (within about 10,000 light-years), the stars are no longer confined to the disk but form a central bulge (or nuclear bulge). When we observe with visible light, we can glimpse the stars in the bulge only in those rare directions where there happens to be relatively little interstellar dust. The first picture that actually succeeded in showing the bulge as a whole was taken at infrared wavelengths (Figure 25.7).

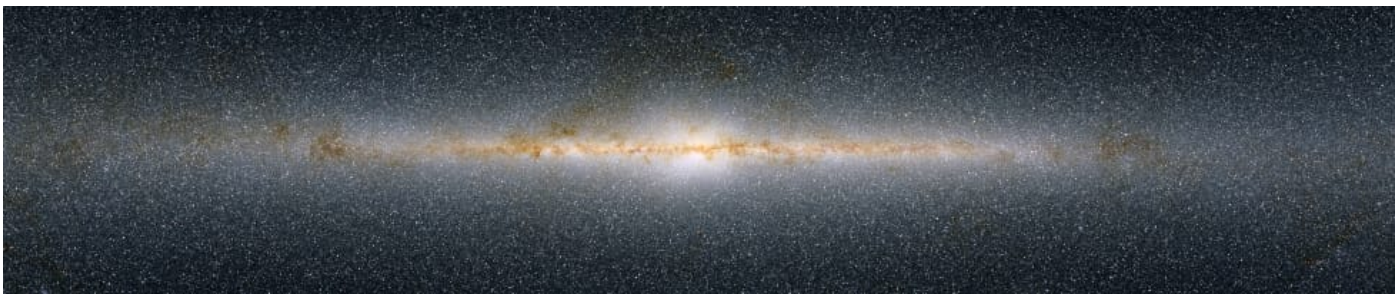


Figure 25.7 : Inner Part of the Milky Way Galaxy. This beautiful infrared map, showing half a billion stars, was obtained as part of the Two Micron All Sky Survey (2MASS). Because interstellar dust does not absorb infrared as strongly as visible light, this view reveals the previously hidden bulge of old stars that surrounds the center of our Galaxy, along with the Galaxy's thin disk component. (credit: modification of work by 2MASS/J. Carpenter, T. H. Jarrett, and R. Hurt)

The fact that much of the bulge is obscured by dust makes its shape difficult to determine. For a long time, astronomers assumed it was spherical. However, infrared images and other data indicate that the bulge is about two times longer than it is wide, and shaped rather like a peanut. The relationship between this elongated inner bulge and the larger bar of stars remains uncertain. At the very center of the nuclear bulge is a tremendous concentration of matter, which we will discuss later in this chapter.

In our Galaxy, the thin and thick disks and the nuclear bulge are embedded in a spherical halo of very old, faint stars that extends to a distance of at least 150,000 light-years from the galactic center. Most of the globular clusters are also found in this halo.

The mass in the Milky Way extends even farther out, well beyond the boundary of the luminous stars to a distance of at least 200,000 light-years from the center of the Galaxy. This invisible mass has been give the name *dark matter* because it emits no light and cannot be seen with any telescope. Its composition is unknown, and it can be detected only because of its gravitational effects on the motions of luminous matter that we can see. We know that this extensive dark matter halo exists because of its effects on the orbits of distant star clusters and other dwarf galaxies that are associated with the Galaxy. This mysterious halo will be a subject of the section on The Mass of the Galaxy, and the properties of dark matter will be discussed more in the chapter on The Big Bang.

Some vital statistics of the thin and thick disks and the stellar halo are given in Table 25.1, with an illustration in Figure 25.8. Note particularly how the ages of stars correlate with where they are found. As we shall see, this information holds important clues to how the Milky Way Galaxy formed.

Table 25.1: Characteristics of the Milky Way Galaxy

Property	Thin Disk	Thick Disk	Stellar Halo (Excludes Dark Matter)
Stellar mass	4×1010 4×1010 M_{Sun}	A few percent of the thin disk mass	10^{10} M_{Sun}
Luminosity	3×10103×1010 L_{Sun}	A few percent of the thin disk luminosity	8×1088×108 L_{Sun}
Typical age of stars	1 million to 10 billion years	11 billion years	13 billion years
Heavier-element abundance	High	Intermediate	Very low
Rotation	High	Intermediate	Very low

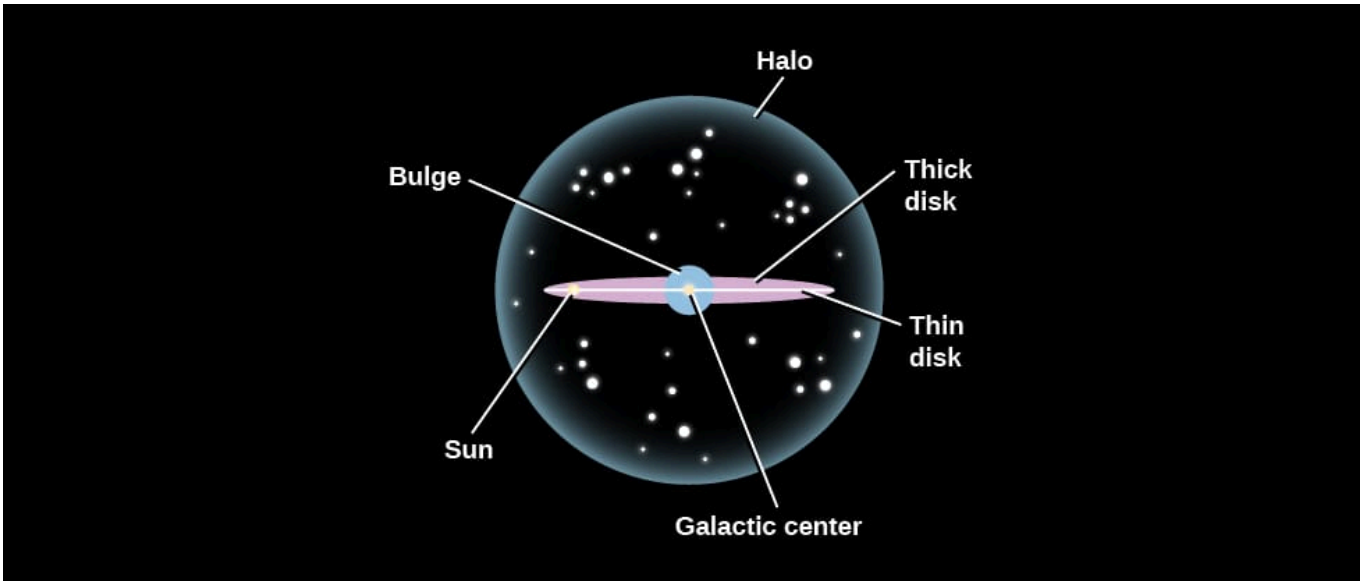


Figure 25.8 : Major Parts of the Milky Way Galaxy. This schematic shows the major components of our Galaxy.

Establishing this overall picture of the Galaxy from our dust-shrouded viewpoint inside the thin disk has been one of the great achievements of modern astronomy (and one that took decades of effort by astronomers working with a wide range of telescopes). One thing that helped enormously was the discovery that our Galaxy is not unique in its characteristics. There are many other flat, spiral-shaped islands of stars, gas, and dust in the universe. For example, the Milky Way somewhat resembles the Andromeda galaxy, which, at a distance of about 2.3 million light-years, is our nearest neighboring giant spiral galaxy. Just as you can get a much better picture of yourself if someone else takes the photo from a distance away, pictures and other diagnostic observations of nearby galaxies that resemble ours have been vital to our understanding of the properties of the Milky Way.

Link to Learning

Start this video, then click-and-drag to look in all directions from Earth in the Gaia space telescope's visualization of the stars of our galaxy.

The Milky Way viewer allows you to move to different locations in the Milky Way (using the Move and Zoom buttons) and view different stellar populations.

Making Connections: The Milky Way Galaxy in Myth and Legend

To most of us living in the twenty-first century, the Milky Way Galaxy is an elusive sight. We must make an effort to leave our well-lit homes and streets and venture beyond our cities and suburbs into less populated environments. Once the light pollution subsides to negligible levels, the Milky Way can be readily spotted arching over the sky on clear, moonless nights. The Milky Way is especially bright in late summer and early fall in the Northern Hemisphere. Some of the best places to view the Milky Way are in our national and state parks, where residential and industrial developments have been kept to a minimum. Some of these parks host special sky-gazing events that are definitely worth checking out—especially during the two weeks surrounding the new moon, when the faint stars and Milky Way don't have to compete with the Moon's brilliance.

Go back a few centuries, and these starlit sights would have been the norm rather than the exception. Before the advent of electric or even gas lighting, people relied on short-lived fires to illuminate their homes and byways. Consequently, their night skies were typically much darker. Confronted by myriad stellar patterns and the Milky Way's gauzy band of diffuse light, people of all cultures developed myths to make sense of it all.

Some of the oldest myths relating to the Milky Way are maintained by the aboriginal Australians through their rock painting and storytelling. These legacies are thought to go back tens of thousands of years, to when the aboriginal people were being "dreamed" along with the rest of the cosmos. The Milky Way played a central role as an arbiter of the Creation. Taking the form of a great serpent, it joined with the Earth serpent to dream and thus create all the creatures on Earth.

The ancient Greeks viewed the Milky Way as a spray of milk that spilled from the breast of the goddess Hera. In this legend, Zeus had secretly placed his infant son Heracles at Hera's breast while she was asleep in order to give his half-human son immortal powers. When Hera awoke and found Heracles suckling, she pushed him away, causing her milk to spray forth into the cosmos (Figure 25.9).

The dynastic Chinese regarded the Milky Way as a "silver river" that was made to separate two star-crossed lovers. To the east of the Milky Way, Zhi Nu, the weaving maiden, was identified with the bright star Vega in the constellation of Lyra the Harp. To the west of the Milky Way, her lover Niu Lang, the cowherd, was associated with the star Altair in the constellation of Aquila the Eagle. They had been exiled on opposite sides of the Milky Way by Zhi Nu's mother, the Queen of Heaven, after she heard of their secret marriage and the birth of their two children. However, once a year, they are permitted to reunite. On the seventh day of the seventh lunar month (which typically occurs in our month of August), they would meet on a bridge over the Milky Way that thousands of magpies had made (Figure 25.9). This romantic time continues to be celebrated today as Qi Xi, meaning "Double Seventh," with couples reenacting the cosmic reunion of Zhi Nu and Niu Lang.



(a)



(b)

Figure 25.9 : The Milky Way in Myth. (a) *Origin of the Milky Way* by Jacopo Tintoretto (circa 1575) illustrates the Greek myth that explains the formation of the Milky Way. (b) *The Moon of the Milky Way* by Japanese painter Tsukioka Yoshitoshi depicts the Chinese legend of Zhi Nu and Niu Lang.

To the Quechua Indians of Andean Peru, the Milky Way was seen as the celestial abode for all sorts of cosmic creatures. Arrayed along the Milky Way are myriad dark patches that they identified with partridges, llamas, a toad, a snake, a fox, and other animals. The Quechua's orientation toward the dark regions rather than the glowing band of starlight appears to be unique among all the myth makers. Likely, their access to the richly structured southern Milky Way had something to do with it.

Among Finns, Estonians, and related northern European cultures, the Milky Way is regarded as the "pathway of birds" across the night sky. Having noted that birds seasonally migrate along a north-south route, they identified this byway with the Milky Way. Recent scientific studies have shown that this myth is rooted in fact: the birds of this region use the Milky Way as a guide for their annual migrations.

Today, we regard the Milky Way as our galactic abode, where the foment of star birth and star death plays out on a grand stage, and where sundry planets have been found to be orbiting all sorts of stars. Although our perspective on the Milky Way is based on scientific investigations, we share with our forebears an affinity for telling stories of origin and transformation. In these regards, the Milky Way continues to fascinate and inspire us.

This page titled 6.1: The Architecture of the Galaxy is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by OpenStax via source content that was edited to the style and standards of the LibreTexts platform.

6.2: Types of Galaxies

Learning Objectives

By the end of this section, you will be able to:

- Describe the properties and features of elliptical, spiral, and irregular galaxies
- Explain what may cause a galaxy's appearance to change over time

Having established the existence of other galaxies, Hubble and others began to observe them more closely—noting their shapes, their contents, and as many other properties as they could measure. This was a daunting task in the 1920s when obtaining a single photograph or spectrum of a galaxy could take a full night of tireless observing. Today, larger telescopes and electronic detectors have made this task less difficult, although observing the most distant galaxies (those that show us the universe in its earliest phases) still requires enormous effort.

The first step in trying to understand a new type of object is often simply to describe it. Remember, the first step in understanding stellar spectra was simply to sort them according to appearance (see Analyzing Starlight). As it turns out, the biggest and most luminous galaxies come in one of two basic shapes: either they are flatter and have spiral arms, like our own Galaxy, or they appear to be elliptical (blimp- or cigar-shaped). Many smaller galaxies, in contrast, have an irregular shape.

Spiral Galaxies

Our own Galaxy and the Andromeda galaxy are typical, large spiral galaxies (see Figure 26.2). They consist of a central bulge, a halo, a disk, and spiral arms. Interstellar material is usually spread throughout the disks of spiral galaxies. Bright emission nebulae and hot, young stars are present, especially in the spiral arms, showing that new star formation is still occurring. The disks are often dusty, which is especially noticeable in those systems that we view almost edge on (Figure 26.4).

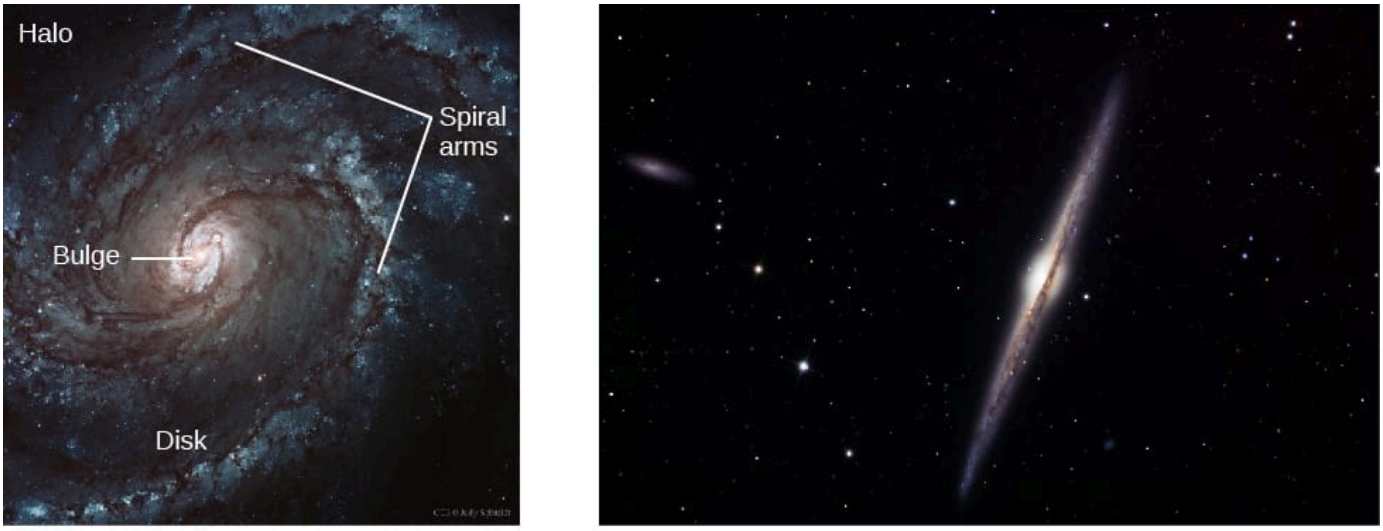


Figure 26.4 : Spiral Galaxies. (a) The spiral arms of M100, shown here, are bluer than the rest of the galaxy, indicating young, high-mass stars and star-forming regions. (b) We view this spiral galaxy, NGC 4565, almost exactly edge on, and from this angle, we can see the dust in the plane of the galaxy; it appears dark because it absorbs the light from the stars in the galaxy. (credit a: modification of work by Hubble Legacy Archive, NASA, ESA, and Judy Schmidt; credit b: modification of work by "Jschulman555"/ Wikimedia)

In galaxies that we see face on, the bright stars and emission nebulae make the arms of spirals stand out like those of a pinwheel on the fourth of July. Open star clusters can be seen in the arms of nearer spirals, and globular clusters are often visible in their

halos. Spiral galaxies contain a mixture of young and old stars, just as the Milky Way does. All spirals rotate, and the direction of their spin is such that the arms appear to trail much like the wake of a boat.

About two-thirds of the nearby spiral galaxies have boxy or peanut-shaped bars of stars running through their centers (Figure 26.5). Showing great originality, astronomers call these galaxies barred spirals.



Figure 26.5 : Barred Spiral Galaxy. NGC 1300, shown here, is a barred spiral galaxy. Note that the spiral arms begin at the ends of the bar. (credit: NASA, ESA, and the Hubble Heritage Team(STScI/AURA))

As we noted in The Milky Way Galaxy chapter, our Galaxy has a modest bar too (see Figure 25.10). The spiral arms usually begin from the ends of the bar. The fact that bars are so common suggests that they are long lived; it may be that most spiral galaxies form a bar at some point during their evolution.

In both barred and unbarred spiral galaxies, we observe a range of different shapes. At one extreme, the central bulge is large and luminous, the arms are faint and tightly coiled, and bright emission nebulae and supergiant stars are inconspicuous. Hubble, who developed a system of classifying galaxies by shape, gave these galaxies the designation Sa. Galaxies at this extreme may have no clear spiral arm structure, resulting in a lens-like appearance (they are sometimes referred to as lenticular galaxies). These galaxies seem to share as many properties with elliptical galaxies as they do with spiral galaxies

At the other extreme, the central bulge is small and the arms are loosely wound. In these Sc galaxies, luminous stars and emission nebulae are very prominent. Our Galaxy and the Andromeda galaxy are both intermediate between the two extremes. Photographs of spiral galaxies, illustrating the different types, are shown in Figure 26.6, along with elliptical galaxies for comparison.

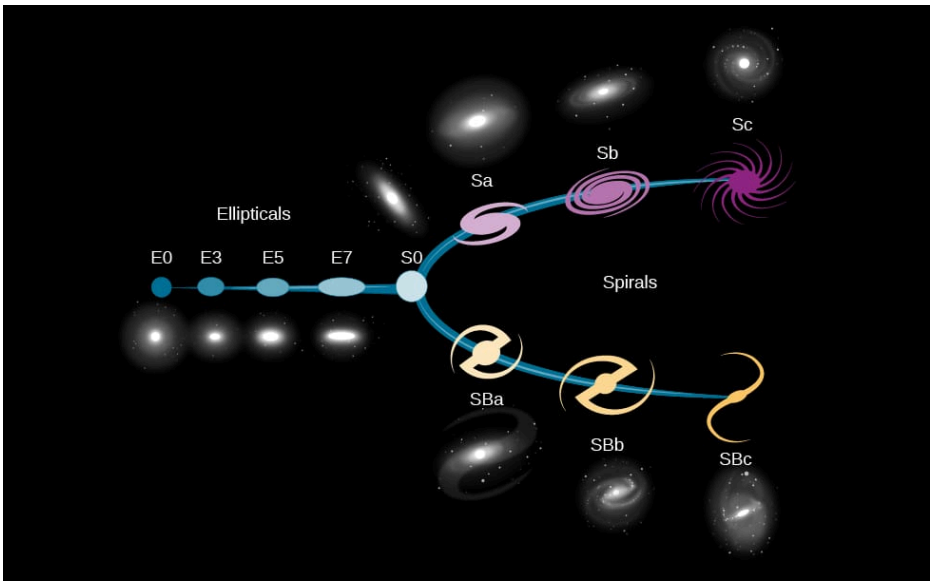


Figure 26.6 : Hubble Classification of Galaxies. This figure shows Edwin Hubble's original classification of galaxies. Elliptical galaxies are on the left. On the right, you can see the basic spiral shapes illustrated, alongside images of actual barred and unbarred spirals. (credit: modification of work by NASA, ESA)

The luminous parts of spiral galaxies appear to range in diameter from about 20,000 to more than 100,000 light-years. Recent studies have found that there is probably a large amount of galactic material that extends well beyond the apparent edge of galaxies. This material appears to be thin, cold gas that is difficult to detect in most observations.

From the observational data available, the masses of the visible portions of spiral galaxies are estimated to range from 1 billion to 1 trillion Suns (10^9 to 10^{12} M_{Sun}). The total luminosities of most spirals fall in the range of 100 million to 100 billion times the luminosity of our Sun (10^8 to 10^{11} L_{Sun}). Our Galaxy and M31 are relatively large and massive, as spirals go. There is also considerable dark matter in and around the galaxies, just as there is in the Milky Way; we deduce its presence from how fast stars in the outer parts of the Galaxy are moving in their orbits.

Elliptical Galaxies

Elliptical galaxies consist almost entirely of old stars and have shapes that are spheres or ellipsoids (somewhat squashed spheres) (Figure 26.7). They contain no trace of spiral arms. Their light is dominated by older reddish stars (the population II stars discussed in The Milky Way Galaxy). In the larger nearby ellipticals, many globular clusters can be identified. Dust and emission nebulae are not conspicuous in elliptical galaxies, but many do contain a small amount of interstellar matter.



Figure 26.7 : Elliptical Galaxies. (a) ESO 325-G004 is a giant elliptical galaxy. Other elliptical galaxies can be seen around the edges of this image. (b) This elliptical galaxy probably originated from the collision of two spiral galaxies. (credit a: modification of work by NASA, ESA, and The Hubble Heritage Team (STScI/AURA); credit b: modification of work by ESA/Hubble, NASA)

Elliptical galaxies show various degrees of flattening, ranging from systems that are approximately spherical to those that approach the flatness of spirals. The rare giant ellipticals (for example, ESO 325-G004 in Figure 26.7) reach luminosities of 10^{11} L_{Sun}. The mass in a giant elliptical can be as large as 10^{13} M_{Sun}. The diameters of these large galaxies extend over several hundred thousand light-years and are considerably larger than the largest spirals. Although individual stars orbit the center of an elliptical galaxy, the orbits are not all in the same direction, as occurs in spirals. Therefore, ellipticals don't appear to rotate in a systematic way, making it difficult to estimate how much dark matter they contain.

We find that elliptical galaxies range all the way from the giants, just described, to dwarfs, which may be the most common kind of galaxy. *Dwarf ellipticals* (sometimes called dwarf spheroidals) escaped our notice for a long time because they are very faint and difficult to see. An example of a dwarf elliptical is the Leo I Dwarf Spheroidal galaxy shown in Figure 26.8. The luminosity of this typical dwarf is about equal to that of the brightest globular clusters.

Intermediate between the giant and dwarf elliptical galaxies are systems such as M32 and M110, the two companions of the Andromeda galaxy. While they are often referred to as dwarf ellipticals, these galaxies are significantly larger than galaxies such as Leo I.



Figure 26.8 : Dwarf Elliptical Galaxy. M32, a dwarf elliptical

galaxy and one of the companions to the giant Andromeda galaxy M31. M32 is a dwarf by galactic standards, as it is only 2400 light-years across. (credit: NOAO/AURA/NSF)

Irregular Galaxies

Hubble classified galaxies that do not have the regular shapes associated with the categories we just described into the catchall bin of an irregular galaxy, and we continue to use his term. Typically, irregular galaxies have lower masses and luminosities than spiral galaxies. Irregular galaxies often appear disorganized, and many are undergoing relatively intense star formation activity. They contain both young population I stars and old population II stars.

The two best-known irregular galaxies are the Large Magellanic Cloud and Small Magellanic Cloud (Figure 26.9), which are at a distance of a little more than 160,000 light-years away and are among our nearest extragalactic neighbors. Their names reflect the fact that Ferdinand Magellan and his crew, making their round-the-world journey, were the first European travelers to notice them. Although not visible from the United States and Europe, these two systems are prominent from the Southern Hemisphere, where they look like wispy clouds in the night sky. Since they are only about one-tenth as distant as the Andromeda galaxy, they present an excellent opportunity for astronomers to study nebulae, star clusters, variable stars, and other key objects in the setting of another galaxy. For example, the Large Magellanic Cloud contains the 30 Doradus complex (also known as the Tarantula Nebula), one of the largest and most luminous groups of supergiant stars known in any galaxy.



Figure 26.9 : 4-Meter Telescope at Cerro Tololo Inter-American Observatory Silhouetted against the Southern Sky. The Milky Way is seen to the right of the dome, and the Large and Small Magellanic Clouds are seen to the left. (credit: Roger Smith/NOAO/AURA/NSF)

The Small Magellanic Cloud is considerably less massive than the Large Magellanic Cloud, and it is six times longer than it is wide. This narrow wisp of material points directly toward our Galaxy like an arrow. The Small Magellanic Cloud was most likely contorted into its current shape through gravitational interactions with the Milky Way. A large trail of debris from this interaction between the Milky Way and the Small Magellanic Cloud has been strewn across the sky and is seen as a series of gas clouds moving at abnormally high velocity, known as the Magellanic Stream. We will see that this kind of interaction between galaxies will help explain the irregular shapes of this whole category of small galaxies,

Link to Learning

View this beautiful album showcasing the different types of galaxies that have been photographed by the Hubble Space Telescope.

Galaxy Evolution

Encouraged by the success of the H-R diagram for stars (see Analyzing Starlight), astronomers studying galaxies hoped to find some sort of comparable scheme, where differences in appearance could be tied to different evolutionary stages in the life of galaxies. Wouldn't it be nice if every elliptical galaxy evolved into a spiral, for example, just as every main-sequence star evolves into a red giant? Several simple ideas of this kind were tried, some by Hubble himself, but none stood the test of time (and observation).

Because no simple scheme for evolving one type of galaxy into another could be found, astronomers then tended to the opposite point of view. For a while, most astronomers thought that all galaxies formed very early in the history of the universe and that the differences between them had to do with the rate of star formation. Ellipticals were those galaxies in which all the interstellar matter was converted rapidly into stars. Spirals were galaxies in which star formation occurred slowly over the entire lifetime of the galaxy. This idea turned out to be too simple as well.

Today, we understand that at least some galaxies have changed types over the billions of years since the universe began. As we shall see in later chapters, collisions and mergers between galaxies may dramatically change spiral galaxies into elliptical galaxies. Even isolated spirals (with no neighbor galaxies in sight) can change their appearance over time. As they consume their gas, the rate of star formation will slow down, and the spiral arms will gradually become less conspicuous. Over long periods, spirals therefore begin to look more like the galaxies at the middle of Figure 26.6 (which astronomers refer to as S0 types).

Over the past several decades, the study of how galaxies evolve over the lifetime of the universe has become one of the most active fields of astronomical research. We will discuss the evolution of galaxies in more detail in The Evolution and Distribution of Galaxies, but let's first see in a little more detail just what different galaxies are like.

This page titled 6.2: Types of Galaxies is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by OpenStax via source content that was edited to the style and standards of the LibreTexts platform.

6.3: The Expanding Universe

Learning Objectives

By the end of this section, you will be able to:

- Describe the discovery that galaxies getting farther apart as the universe evolves
- Explain how to use Hubble's law to determine distances to remote galaxies
- Describe models for the nature of an expanding universe
- Explain the variation in Hubble's constant

We now come to one of the most important discoveries ever made in astronomy—the fact that the universe is expanding. Before we describe how the discovery was made, we should point out that the first steps in the study of galaxies came at a time when the techniques of spectroscopy were also making great strides. Astronomers using large telescopes could record the spectrum of a faint star or galaxy on photographic plates, guiding their telescopes so they remained pointed to the same object for many hours and collected more light. The resulting spectra of galaxies contained a wealth of information about the composition of the galaxy and the velocities of these great star systems.

Slipher's Pioneering Observations

Curiously, the discovery of the expansion of the universe began with the search for Martians and other solar systems. In 1894, the controversial (and wealthy) astronomer Percival Lowell established an observatory in Flagstaff, Arizona, to study the planets and search for life in the universe. Lowell thought that the spiral nebulae might be solar systems in the process of formation. He therefore asked one of the observatory's young astronomers, Vesto M. Slipher (Figure 26.13), to photograph the spectra of some of the spiral nebulae to see if their spectral lines might show chemical compositions like those expected for newly forming planets.



Figure 26.13 : Vesto M. Slipher (1875–1969). Slipher spent his entire career at the Lowell Observatory, where he discovered the large radial velocities of galaxies. (credit: Lowell Observatory)

The Lowell Observatory's major instrument was a 24-inch refracting telescope, which was not at all well suited to observations of faint spiral nebulae. With the technology available in those days, photographic plates had to be exposed for 20 to 40 hours to produce a good spectrum (in which the positions of the lines could reveal a galaxy's motion). This often meant continuing to expose the same photograph over several nights. Beginning in 1912, and making heroic efforts over a period of about 20 years, Slipher managed to photograph the spectra of more than 40 of the spiral nebulae (which would all turn out to be galaxies).

To his surprise, the spectral lines of most galaxies showed an astounding redshift. By "redshift" we mean that the lines in

the spectra are displaced toward longer wavelengths (toward the red end of the visible spectrum). Recall from the chapter on Radiation and Spectra that a redshift is seen when the source of the waves is moving away from us. Slipher's observations showed that most spirals are racing away at huge speeds; the highest velocity he measured was 1800 kilometers per second.

Only a few spirals—such as the Andromeda and Triangulum Galaxies and M81—all of which are now known to be our close neighbors, turned out to be approaching us. All the other galaxies were moving away. Slipher first announced this discovery in 1914, years before Hubble showed that these objects were other galaxies and before anyone knew how far away they were. No one at the time quite knew what to make of this discovery.

Hubble's Law

The profound implications of Slipher's work became apparent only during the 1920s. Georges Lemaître was a Belgian priest and a trained astronomer. In 1927, he published a paper in French in an obscure Belgian journal in which he suggested that we live in an expanding universe. The title of the paper (translated into English) is "A Homogenous Universe of Constant Mass and Growing Radius Accounting for the Radial Velocity of Extragalactic Nebulae." Lemaître had discovered that Einstein's equations of relativity were consistent with an expanding universe (as had the Russian scientist Alexander Friedmann independently in 1922). Lemaître then went on to use Slipher's data to support the hypothesis that the universe actually is expanding and to estimate the rate of expansion. Initially, scientists paid little attention to this paper, perhaps because the Belgian journal was not widely available.

In the meantime, Hubble was making observations of galaxies with the 2.5-meter telescope on Mt. Wilson, which was then the world's largest. Hubble carried out the key observations in collaboration with a remarkable man, Milton Humason, who dropped out of school in the eighth grade and began his astronomical career by driving a mule train up the trail on Mount Wilson to the observatory (Figure 26.14). In those early days, supplies had to be brought up that way; even astronomers hiked up to the mountaintop for their turns at the telescope. Humason became interested in the work of the astronomers and, after marrying the daughter of the observatory's electrician, took a job as janitor there. After a time, he became a night assistant, helping the astronomers run the telescope and record data. Eventually, he made such a mark that he became a full astronomer at the observatory.

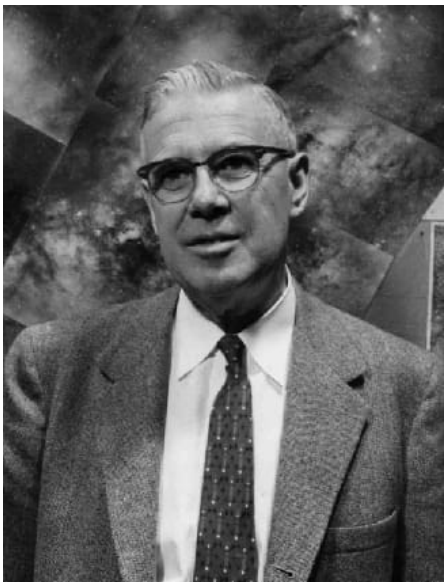


Figure 26.14 : Milton Humason (1891–1972). Humason was Hubble's collaborator on the great task of observing, measuring, and classifying the characteristics of many galaxies. (credit: Caltech Archives)

By the late 1920s, Humason was collaborating with Hubble by photographing the spectra of faint galaxies with the 2.5-meter telescope. (By then, there was no question that the spiral nebulae were in fact galaxies.) Hubble had found ways to improve the accuracy of the estimates of distances to spiral galaxies, and he was able to measure much fainter and more distant galaxies than Slipher could observe with his much-smaller telescope. When Hubble laid his own distance estimates next to measurements of the recession velocities (the speed with which the galaxies were moving away), he found something stunning: there was a relationship between distance and velocity for galaxies. *The more distant the galaxy, the faster it was receding from us*.

In 1931, Hubble and Humason jointly published the seminal paper where they compared distances and velocities of remote galaxies moving away from us at speeds as high as 20,000 kilometers per second and were able to show that the recession

velocities of galaxies are directly proportional to their distances from us (Figure 26.15), just as Lemaître had suggested.

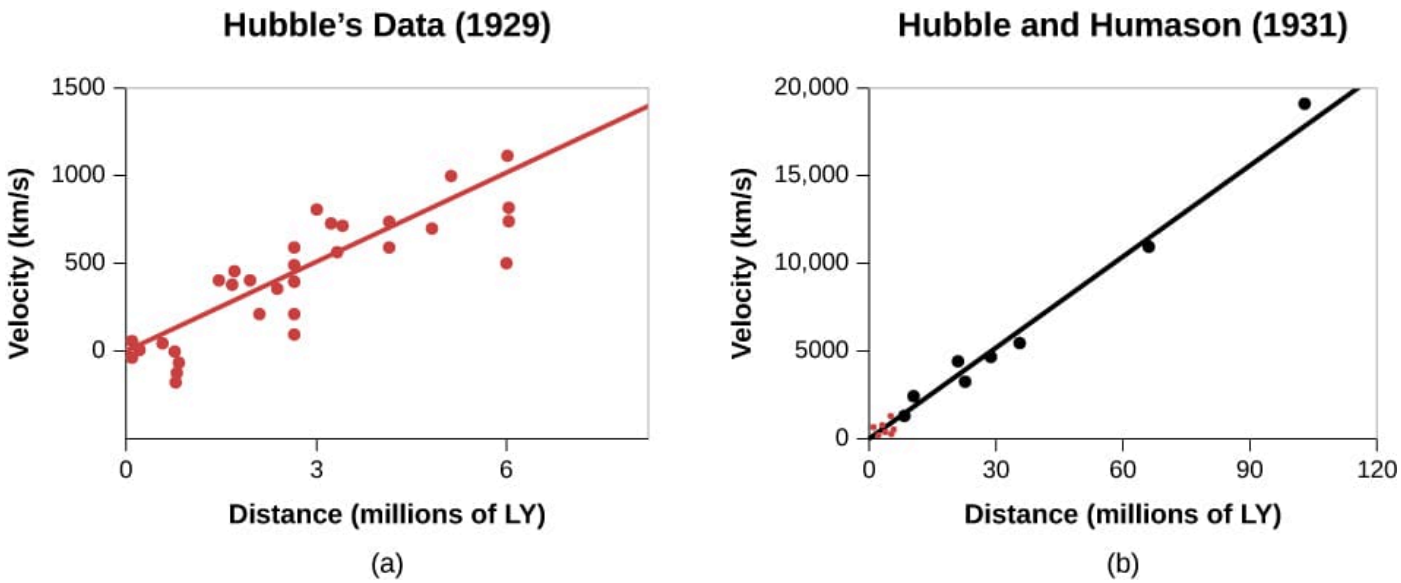


Figure 26.15 : Hubble's Law. (a) These data show Hubble's original velocity-distance relation, adapted from his 1929 paper in the *Proceedings of the National Academy of Sciences*. (b) These data show Hubble and Humason's velocity-distance relation, adapted from their 1931 paper in *The Astrophysical Journal*. The red dots at the lower left are the points in the diagram in the 1929 paper. Comparison of the two graphs shows how rapidly the determination of galactic distances and redshifts progressed in the 2 years between these publications.

We now know that this relationship holds for every galaxy except a few of the nearest ones. Nearly all of the galaxies that are approaching us turn out to be part of the Milky Way's own group of galaxies, which have their own individual motions, just as birds flying in a group may fly in slightly different directions at slightly different speeds even though the entire flock travels through space together.

Written as a formula, the relationship between velocity and distance is

v=H×dv=H×d

where *v* is the velocity, *d* is the distance, and *H* is a number called the Hubble constant. This equation is now known as Hubble's law.[1]

Astronomy Basics: Constants of Proportionality

Mathematical relationships such as Hubble's law are pretty common in life. To take a simple example, suppose your college or university hires you to call rich alumni and ask for donations. You are paid $2.50 for each call; the more calls you can squeeze in between studying astronomy and other courses, the more money you take home. We can set up a formula that connects *p*, your pay, and *n*, the number of calls

p=A×np=A×n

where *A* is the alumni constant, with a value of $2.50. If you make 20 calls, you will earn $2.50 times 20, or $50.

Suppose your boss forgets to tell you what you will get paid for each call. You can calculate the alumni constant that governs your pay by keeping track of how many calls you make and noting your gross pay each week. If you make 100 calls the first week and are paid $250, you can deduce that the constant is $2.50 (in units of dollars per call). Hubble, of course, had no "boss" to tell him what his constant would be—he *had* to calculate its value from the measurements of distance and velocity.

Astronomers express the value of Hubble's constant in units that relate to how they measure speed and distance for galaxies. In this book, we will use kilometers per second per million light-years as that unit. For many years, estimates of the value of the

Hubble constant have been in the range of 15 to 30 kilometers per second per million light-years The most recent work appears to be converging on a value near 22 kilometers per second per million light-years If H is 22 kilometers per second per million light-years, a galaxy moves away from us at a speed of 22 kilometers per second for every million light-years of its distance. As an example, a galaxy 100 million light-years away is moving away from us at a speed of 2200 kilometers per second.

Hubble's law tells us something fundamental about the universe. Since all but the nearest galaxies appear to be in motion away from us, with the most distant ones moving the fastest, we must be living in an expanding universe. We will explore the implications of this idea shortly, as well as in the final chapters of this text. For now, we will just say that Hubble's observation underlies all our theories about the origin and evolution of the universe.

Hubble's Law and Distances

The regularity expressed in Hubble's law has a built-in bonus: it gives us a new way to determine the distances to remote galaxies. First, we must reliably establish Hubble's constant by measuring both the distance and the velocity of many galaxies in many directions to be sure Hubble's law is truly a universal property of galaxies. But once we have calculated the value of this constant and are satisfied that it applies everywhere, much more of the universe opens up for distance determination. Basically, if we can obtain a spectrum of a galaxy, we can immediately tell how far away it is.

The procedure works like this. We use the spectrum to measure the speed with which the galaxy is moving away from us. If we then put this speed and the Hubble constant into Hubble's law equation, we can solve for the distance.

Example 26.1: Hubble's Law

Hubble's law ($v = H \times d$) allows us to calculate the distance to any galaxy. Here is how we use it in practice.

We have measured Hubble's constant to be 22 km/s per million light-years. This means that if a galaxy is 1 million light-years farther away, it will move away 22 km/s faster. So, if we find a galaxy that is moving away at 18,000 km/s, what does Hubble's law tells us about the distance to the galaxy?

Solution

d=vH=18,000km/s22km/s1million light-years=18,00022×1million light-years1=818million light-yearsd=vH=18,000km/s22km/s1million light-years=18,00022×1million light-years1=818million light-years

Note how we handled the units here: the km/s in the numerator and denominator cancel, and the factor of million light-years in the denominator of the constant must be divided correctly before we get our distance of 818 million light-years.

Exercise 6.3.1

Using 22 km/s/million light-years for Hubble's constant, what recessional velocity do we expect to find if we observe a galaxy at 500 million light-years?

Answer available at the end of this section.

Variation of Hubble's Constant

The use of redshift is potentially a very important technique for determining distances because as we have seen, most of our methods for determining galaxy distances are limited to approximately the nearest few hundred million light-years (and they have large uncertainties at these distances). The use of Hubble's law as a distance indicator requires only a spectrum of a galaxy and a measurement of the Doppler shift, and with large telescopes and modern spectrographs, spectra can be taken of extremely faint galaxies.

But, as is often the case in science, things are not so simple. This technique works if, and only if, the Hubble constant has been truly constant throughout the entire life of the universe. When we observe galaxies billions of light-years away, we are seeing them

as they were billions of years ago. What if the Hubble "constant" was different billions of years ago? Before 1998, astronomers thought that, although the universe is expanding, the expansion should be slowing down, or decelerating, because the overall gravitational pull of all matter in the universe would have a dominant, measureable effect. If the expansion is decelerating, then the Hubble constant should be decreasing over time.

The discovery that type Ia supernovae are standard bulbs gave astronomers the tool they needed to observe extremely distant galaxies and measure the rate of expansion billions of years ago. The results were completely unexpected. It turns out that the expansion of the universe is *accelerating* over time! What makes this result so astounding is that there is no way that existing physical theories can account for this observation. While a decelerating universe could easily be explained by gravity, there was no force or property in the universe known to astronomers that could account for the acceleration. In The Big Bang chapter, we will look in more detail at the observations that led to this totally unexpected result and explore its implications for the ultimate fate of the universe.

In any case, if the Hubble constant is not really a constant when we look over large spans of space and time, then the calculation of galaxy distances using the Hubble constant won't be accurate. As we shall see in the chapter on The Big Bang, the accurate calculation of distances requires a model for how the Hubble constant has changed over time. The farther away a galaxy is (and the longer ago we are seeing it), the more important it is to include the effects of the change in the Hubble constant. For galaxies within a few billion light-years, however, the assumption that the Hubble constant is indeed constant gives good estimates of distance.

Models for an Expanding Universe

At first, thinking about Hubble's law and being a fan of the work of Copernicus and Harlow Shapley, you might be shocked. Are all the galaxies really moving *away from us*? Is there, after all, something special about our position in the universe? Worry not; the fact that galaxies are receding from us and that more distant galaxies are moving away more rapidly than nearby ones shows only that the universe is expanding uniformly.

A uniformly expanding universe is one that is expanding at the same rate everywhere. In such a universe, we and all other observers, no matter where they are located, must observe a proportionality between the velocities and distances of equivalently remote galaxies. (Here, we are ignoring the fact that the Hubble constant is not constant over all time, but if at any given time in the evolution of the universe the Hubble constant has the same value everywhere, this argument still works.)

To see why, first imagine a ruler made of stretchable rubber, with the usual lines marked off at each centimeter. Now suppose someone with strong arms grabs each end of the ruler and slowly stretches it so that, say, it doubles in length in 1 minute (Figure 26.16). Consider an intelligent ant sitting on the mark at 2 centimeters—a point that is not at either end nor in the middle of the ruler. He measures how fast other ants, sitting at the 4-, 7-, and 12-centimeter marks, move away from him as the ruler stretches.

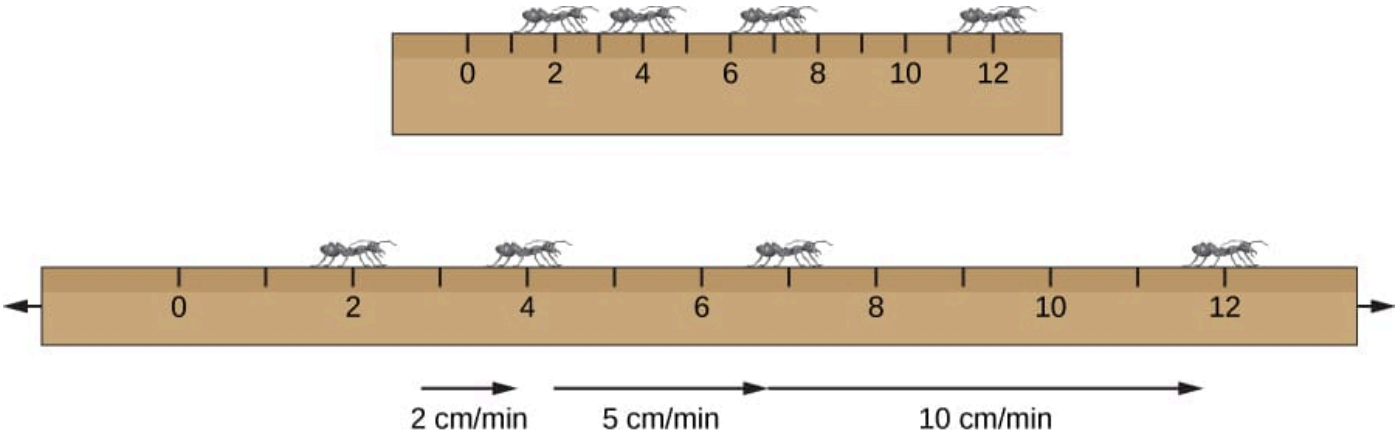


Figure 26.16 : Stretching a Ruler. Ants on a stretching ruler see other ants move away from them. The speed with which another ant moves away is proportional to its distance.

The ant at 4 centimeters, originally 2 centimeters away from our ant, has doubled its distance in 1 minute; it therefore moved away at a speed of 2 centimeters per minute. The ant at the 7-centimeters mark, which was originally 5 centimeters away from our ant, is now 10 centimeters away; it thus had to move at 5 centimeters per minute. The one that started at the 12-centimeters mark, which was 10 centimeters away from the ant doing the counting, is now 20 centimeters away, meaning it must have raced

away at a speed of 10 centimeters per minute. Ants at different distances move away at different speeds, and their speeds are proportional to their distances (just as Hubble's law indicates for galaxies). Yet, notice in our example that all the ruler was doing was stretching uniformly. Also, notice that none of the ants were actually moving of their own accord, it was the stretching of the ruler that moved them apart.

Now let's repeat the analysis, but put the intelligent ant on some other mark—say, on 7 or 12 centimeters. We discover that, as long as the ruler stretches uniformly, this ant also finds every other ant moving away at a speed proportional to its distance. In other words, the kind of relationship expressed by Hubble's law can be explained by a uniform stretching of the "world" of the ants. And all the ants in our simple diagram will see the other ants moving away from them as the ruler stretches.

For a three-dimensional analogy, let's look at the loaf of raisin bread in Figure 26.17. The chef has accidentally put too much yeast in the dough, and when she sets the bread out to rise, it doubles in size during the next hour, causing all the raisins to move farther apart. On the figure, we again pick a representative raisin (that is not at the edge or the center of the loaf) and show the distances from it to several others in the figure (before and after the loaf expands).

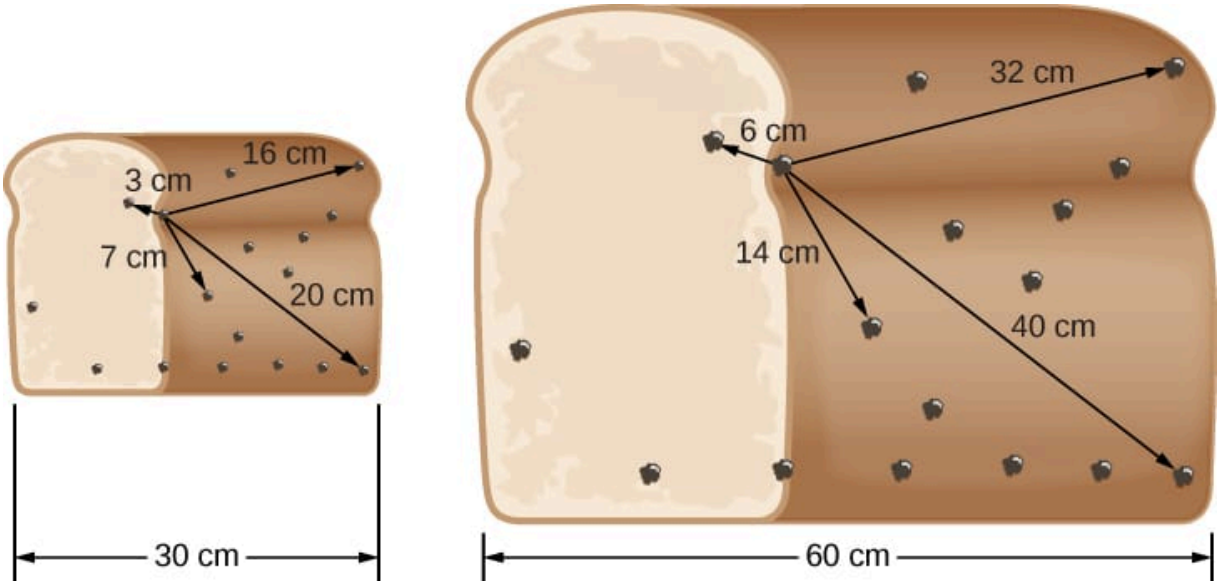


Figure 26.17 : Expanding Raisin Bread. As the raisin bread rises, the raisins "see" other raisins moving away. More distant raisins move away faster in a uniformly expanding bread.

Measure the increases in distance and calculate the speeds for yourself on the raisin bread, just like we did for the ruler. You will see that, since each distance doubles during the hour, each raisin moves away from our selected raisin at a speed proportional to its distance. The same is true no matter which raisin you start with.

Our two analogies are useful for clarifying our thinking, but you must not take them literally. On both the ruler and the raisin bread, there are points that are at the end or edge. You can use these to pinpoint the middle of the ruler and the loaf. While our models of the universe have some resemblance to the properties of the ruler and the loaf, the universe has no boundaries, no edges, and no center (all mind-boggling ideas that we will discuss in a later chapter).

What is useful to notice about both the ants and the raisins is that they themselves did not "cause" their motion. It isn't as if the raisins decided to take a trip away from each other and then hopped on a hoverboard to get away. No, in both our analogies, it was the stretching of the medium (the ruler or the bread) that moved the ants or the raisins farther apart. In the same way, we will see in The Big Bang chapter that the galaxies don't have rocket motors propelling them away from each other. Instead, they are passive participants in the *expansion of space*. As space stretches, the galaxies are carried farther and farther apart much as the ants and the raisins were. (If this notion of the "stretching" of space surprises or bothers you, now would be a good time to review the information about spacetime in Black Holes and Curved Spacetime. We will discuss these ideas further as our discussion broadens from galaxies to the whole universe.)

The expansion of the universe, by the way, does not imply that the individual galaxies and clusters of galaxies themselves are expanding. Neither raisins nor the ants in our analogy grow in size as the loaf expands. Similarly, gravity holds galaxies and clusters of galaxies together, and they get farther away from each other—without themselves changing in size—as the universe expands.

Footnotes

- 1In 2020, the International Astronomical Union suggested that it would be more fair to call it the Hubble-Lemaitre law. In telling the history in this textbook, we have acknowledged the role Lemaître played, and we urge our readers to keep his contributions in mind as they read on.

Answers

Exercise 6.3.1

v=d×H=500million light-years×22km/s1million light-years=11,000km/sv=d×H=500million light-years×22km/s1million light-years=11,000km/s

This page titled 6.3: The Expanding Universe is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by OpenStax via source content that was edited to the style and standards of the LibreTexts platform.

6.4: The Beginning of the Universe

Learning Objectives

By the end of this section, you will be able to:

- Describe what the universe was like during the first few minutes after it began to expand
- Explain how the first new elements were formed during the first few minutes after the Big Bang
- Describe how the contents of the universe change as the temperature of the universe decreases

The best evidence we have today indicates that the first galaxies did not begin to form until a few hundred million years after the Big Bang. What were things like before there were galaxies and space had not yet stretched very significantly? Amazingly, scientists have been able to calculate in some detail what was happening in the universe in the first few minutes after the Big Bang.

The History of the Idea

It is one thing to say the universe had a beginning (as the equations of general relativity imply) and quite another to describe that beginning. The Belgian priest and cosmologist Georges Lemaître was probably the first to propose a specific model for the Big Bang itself (Figure 29.11). He envisioned all the matter of the universe starting in one great bulk he called the *primeval atom*, which then broke into tremendous numbers of pieces. Each of these pieces continued to fragment further until they became the present atoms of the universe, created in a vast nuclear fission. In a popular account of his theory, Lemaître wrote, “The evolution of the world could be compared to a display of fireworks just ended—some few red wisps, ashes, and smoke. Standing on a well-cooled cinder, we see the slow fading of the suns and we try to recall the vanished brilliance of the origin of the worlds.”



Figure 29.11 : Abbé Georges Lemaître (1894–1966). This Belgian cosmologist studied theology at Mechelen and mathematics and physics at the University of Leuven. It was there that he began to explore the expansion of the universe and postulated its explosive beginning. He actually predicted Hubble’s law 2 years before its verification, and he was the first to consider seriously the physical processes by which the universe began.

Link to Learning

View a short video about the work of Lemaître, considered by some to be the father of the Big Bang theory.

Physicists today know much more about nuclear physics than was known in the 1920s, and they have shown that the primeval fission model cannot be correct. Yet Lemaître's vision was in some respects quite prophetic. We still believe that everything was together at the beginning; it was just not in the form of matter we now know. Basic physical principles tell us that when the universe was much denser, it was also much hotter, and that it cools as it expands, much as gas cools when sprayed from an aerosol can.

By the 1940s, scientists knew that fusion of hydrogen into helium was the source of the Sun's energy. Fusion requires high temperatures, and the early universe must have been hot. Based on these ideas, American physicist George Gamow (Figure 29.12) suggested a universe with a different kind of beginning that involved nuclear fusion instead of fission. Ralph Alpher worked out the details for his PhD thesis, and the results were published in 1948. (Gamow, who had a quirky sense of humor, decided at the last minute to add the name of physicist Hans Bethe to their paper, so that the coauthors on this paper about the beginning of things would be Alpher, Bethe, and Gamow, a pun on the first three letters of the Greek alphabet: alpha, beta, and gamma.) Gamow's universe started with fundamental particles that built up the heavy elements by fusion in the Big Bang.

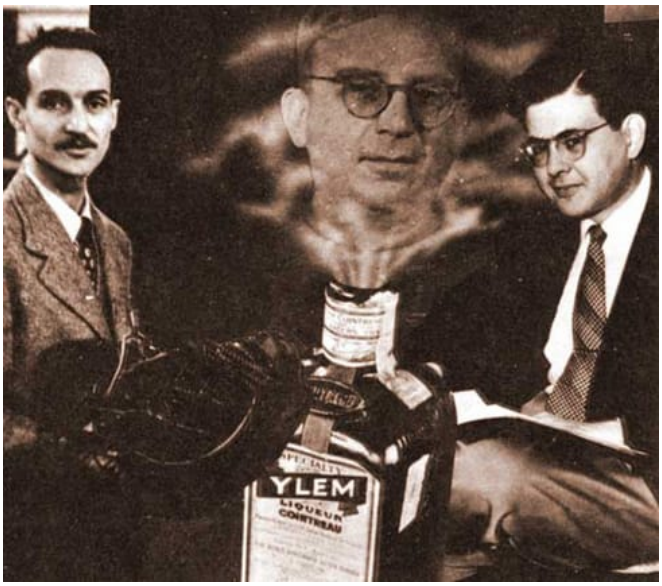


Figure 29.12 : George Gamow and Collaborators. This composite image shows George Gamow emerging like a genie from a bottle of ylem, a Greek term for the original substance from which the world formed. Gamow revived the term to describe the material of the hot Big Bang. Flanking him are Robert Herman (left) and Ralph Alpher (right), with whom he collaborated in working out the physics of the Big Bang. (The modern composer Karlheinz Stockhausen was inspired by Gamow's ideas to write a piece of music called *Ylem*, in which the players actually move away from the stage as they perform, simulating the expansion of the universe.)

Gamow's ideas were close to our modern view, except we now know that the early universe remained hot enough for fusion for only a short while. Thus, only the three lightest elements—hydrogen, helium, and a small amount of lithium—were formed in appreciable abundances at the beginning. The heavier elements formed later in stars. Since the 1940s, many astronomers and physicists have worked on a detailed theory of what happened in the early stages of the universe.

The First Few Minutes

Let's start with the first few minutes following the Big Bang. Three basic ideas hold the key to tracing the changes that occurred during the time just after the universe began. The first, as we have already mentioned, is that the universe cools as it expands. Figure 29.13 shows how the temperature changes with the passage of time. Note that a huge span of time, from a tiny fraction of a second to billions of years, is summarized in this diagram. In the first fraction of a second, the universe was unimaginably hot. By the time 0.01 second had elapsed, the temperature had dropped to 100 billion (10^{11}) K. After about 3 minutes, it had fallen to about 1 billion (10^{9}) K, still some 70 times hotter than the interior of the Sun. After a few hundred thousand years, the temperature was down to a mere 3000 K, and the universe has continued to cool since that time.

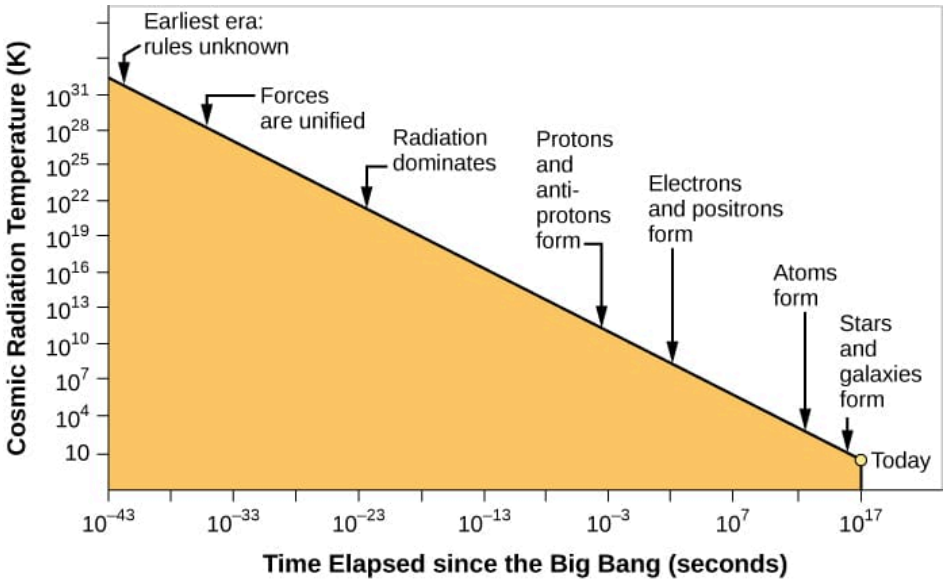


Figure 29.13 : Temperature of the Universe. This graph shows how the temperature of the universe varies with time as predicted by the standard model of the Big Bang. Note that both the temperature (vertical axis) and the time in seconds (horizontal axis) change over vast scales on this compressed diagram.

All of these temperatures but the last are derived from theoretical calculations since (obviously) no one was there to measure them directly. As we shall see in the next section, however, we have actually detected the feeble glow of radiation emitted at a time when the universe was a few hundred thousand years old. We can measure the characteristics of that radiation to learn what things were like long ago. Indeed, the fact that we have found this ancient glow is one of the strongest arguments in favor of the Big Bang model.

The second step in understanding the evolution of the universe is to realize that at very early times, it was so hot that it contained mostly radiation (and not the matter that we see today). The photons that filled the universe could collide and produce material particles; that is, under the conditions just after the Big Bang, energy could turn into matter (and matter could turn into energy). We can calculate how much mass is produced from a given amount of energy by using Einstein's formula E=mc2The Sun: A Nuclear Powerhouse).

The idea that energy could turn into matter in the universe at large is a new one for many students, since it is not part of our everyday experience. That's because, when we compare the universe today to what it was like right after the Big Bang, we live in cold, hard times. The photons in the universe today typically have far-less energy than the amount required to make new matter. In the discussion on the source of the Sun's energy in The Sun: A Nuclear Powerhouse, we briefly mentioned that when subatomic particles of matter and *antimatter* collide, they turn into pure energy. But the reverse, energy turning into matter and antimatter, is equally possible. This process has been observed in particle accelerators around the world. If we have enough energy, under the right circumstances, new particles of matter (and antimatter) are indeed created —and the conditions were right during the first few minutes after the expansion of the universe began.

Our third key point is that the hotter the universe was, the more energetic were the photons available to make matter and antimatter (see Figure 29.13). To take a specific example, at a temperature of 6 billion (6×10^9) K, the collision of two typical photons can create an electron and its antimatter counterpart, a positron. If the temperature exceeds 10^{14} K, much more massive protons and antiprotons can be created.

The Evolution of the Early Universe

Keeping these three ideas in mind, we can trace the evolution of the universe from the time it was about 0.01 second old and had a temperature of about 100 billion K. Why not begin at the very beginning? There are as yet no theories that allow us penetrate to a time before about 10^{-43} second (this number is a decimal point followed by 42 zeros and then a one). It is so small that we cannot relate it to anything in our everyday experience. When the universe was that young, its density was so high that the theory of general relativity is not adequate to describe it, and even the concept of time breaks down.

Scientists, by the way, have been somewhat more successful in describing the universe when it was older than 10^{-43} second but still less than about 0.01 second old. We will take a look at some of these ideas later in this chapter, but for now, we want to start with somewhat more familiar situations.

By the time the universe was 0.01 second old, it consisted of a soup of matter and radiation; the matter included protons and neutrons, leftovers from an even younger and hotter universe. Each particle collided rapidly with other particles. The temperature was no longer high enough to allow colliding photons to produce neutrons or protons, but it was sufficient for the production of electrons and positrons (Figure 29.14). There was probably also a sea of exotic subatomic particles that would later play a role as dark matter. All the particles jiggled about on their own; it was still much too hot for protons and neutrons to combine to form the nuclei of atoms.

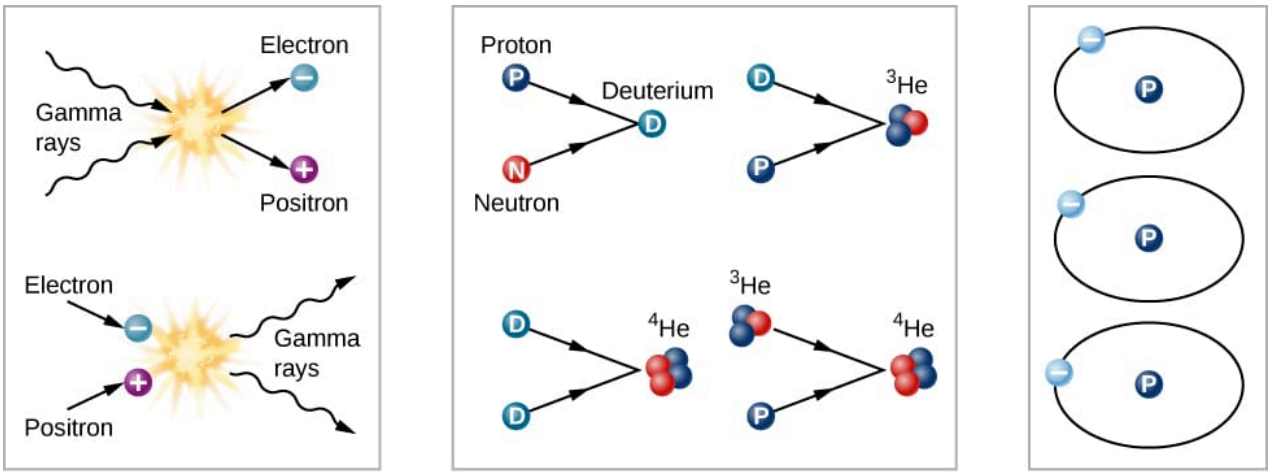


Figure 29.14 : Particle Interactions in the Early Universe. (a) In the first fractions of a second, when the universe was very hot, energy was converted into particles and antiparticles. The reverse reaction also happened: a particle and antiparticle could collide and produce energy. (b) As the temperature of the universe decreased, the energy of typical photons became too low to create matter. Instead, existing particles fused to create such nuclei as deuterium and helium. (c) Later, it became cool enough for electrons to settle down with nuclei and make neutral atoms. Most of the universe was still hydrogen.

Think of the universe at this time as a seething cauldron, with photons colliding and interchanging energy, and sometimes being destroyed to create a pair of particles. The particles also collided with one another. Frequently, a matter particle and an antimatter particle met and turned each other into a burst of gamma-ray radiation.

Among the particles created in the early phases of the universe was the ghostly neutrino (see The Sun: A Nuclear Powerhouse), which today interacts only very rarely with ordinary matter. In the crowded conditions of the very early universe, however, neutrinos ran into so many electrons and positrons that they experienced frequent interactions despite their “antisocial” natures.

By the time the universe was a little more than 1 second old, the density had dropped to the point where neutrinos no longer interacted with matter but simply traveled freely through space. In fact, these neutrinos should now be all around us. Since they have been traveling through space unimpeded (and hence unchanged) since the universe was 1 second old, measurements of their properties would offer one of the best tests of the Big Bang model. Unfortunately, the very characteristic that makes them so useful—the fact that they interact so weakly with matter that they have survived unaltered for all but the first second of time—also renders them unable to be measured, at least with present techniques. Perhaps someday someone will devise a way to capture these elusive messengers from the past.

Atomic Nuclei Form

When the universe was about 3 minutes old and its temperature was down to about 900 million K, protons and neutrons could combine. At higher temperatures, these atomic nuclei had immediately been blasted apart by interactions with high-energy photons and thus could not survive. But at the temperatures and densities reached between 3 and 4 minutes after the beginning, deuterium (a proton and neutron) lasted long enough that collisions could convert some of it into helium, (Figure 29.14). In

essence, the entire universe was acting the way centers of stars do today—fusing new elements from simpler components. In addition, a little bit of element 3, lithium, could also form.

This burst of cosmic fusion was only a brief interlude, however. By 4 minutes after the Big Bang, more helium was having trouble forming. The universe was still expanding and cooling down. After the formation of helium and some lithium, the temperature had dropped so low that the fusion of helium nuclei into still-heavier elements could not occur. No elements beyond lithium could form in the first few minutes. That 4-minute period was the end of the time when the entire universe was a fusion factory. In the cool universe we know today, the fusion of new elements is limited to the centers of stars and the explosions of supernovae.

Still, the fact that the Big Bang model allows the creation of a good deal of helium is the answer to a long-standing mystery in astronomy. Put simply, there is just too much helium in the universe to be explained by what happens inside stars. All the generations of stars that have produced helium since the Big Bang cannot account for the quantity of helium we observe. Furthermore, even the oldest stars and the most distant galaxies show significant amounts of helium. These observations find a natural explanation in the synthesis of helium by the Big Bang itself during the first few minutes of time. We estimate that *10 times more helium* was manufactured in the first 4 minutes of the universe than in all the generations of stars during the succeeding 10 to 15 billion years.

Learning from Deuterium

We can learn many things from the way the early universe made atomic nuclei. It turns out that all of the deuterium (a hydrogen nucleus with a neutron in it) in the universe was formed during the first 4 minutes. In stars, any region hot enough to fuse two protons to form a deuterium nucleus is also hot enough to change it further—either by destroying it through a collision with an energetic photon or by converting it into helium through nuclear reactions.

The amount of deuterium that can be produced in the first 4 minutes of creation depends on the density of the universe at the time deuterium was formed. If the density were relatively high, nearly all the deuterium would have been converted into helium through interactions with protons, just as it is in stars. If the density were relatively low, then the universe would have expanded and thinned out rapidly enough that some deuterium would have survived. The amount of deuterium we see today thus gives us a clue to the density of the universe when it was about 4 minutes old. Theoretical models can relate the density then to the density now; thus, measurements of the abundance of deuterium today can give us an estimate of the current density of the universe.

The measurements of deuterium indicate that the present-day density of ordinary matter—protons and neutrons—is about 5×10^{-28}–5×10^{-28} kg/m^3. Deuterium can only provide an estimate of the density of ordinary matter because the abundance of deuterium is determined by the particles that interact to form it, namely protons and neutrons alone. From the abundance of deuterium, we know that not enough protons and neutrons are present, by a factor of about 20, to produce a critical-density universe.

We do know, however, that there are dark matter particles that add to the overall matter density of the universe, which is then higher than what is calculated for ordinary matter alone. Because dark matter particles do not affect the production of deuterium, measurement of the deuterium abundance cannot tell us how much dark matter exists. Dark matter is made of some exotic kind of particle, not yet detected in any earthbound laboratory. It is definitely not made of protons and neutrons like the readers of this book.

This page titled 6.4: The Beginning of the Universe is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by OpenStax via source content that was edited to the style and standards of the LibreTexts platform.

6.5: The Cosmic Microwave Background

Learning Objectives

By the end of this section, you will be able to:

- Explain why we can observe the afterglow of the hot, early universe
- Discuss the properties of this afterglow as we see it today, including its average temperature and the size of its temperature fluctuations
- Describe open, flat, and curved universes and explain which type of universe is supported by observations
- Summarize our current knowledge of the basic properties of the universe including its age and contents

The description of the first few minutes of the universe is based on theoretical calculations. It is crucial, however, that a scientific theory should be testable. What predictions does it make? And do observations show those predictions to be accurate? One success of the theory of the first few minutes of the universe is the correct prediction of the amount of helium in the universe.

Another prediction is that a significant milestone in the history of the universe occurred about 380,000 years after the Big Bang. Scientists have directly observed what the universe was like at this early stage, and these observations offer some of the strongest support for the Big Bang theory. To find out what this milestone was, let's look at what theory tells us about what happened during the first few hundred thousand years after the Big Bang.

The fusion of helium and lithium was completed when the universe was about 4 minutes old. The universe then continued to resemble the interior of a star in some ways for a few hundred thousand years more. It remained hot and opaque, with radiation being scattered from one particle to another. It was still too hot for electrons to "settle down" and become associated with a particular nucleus; such free electrons are especially effective at scattering photons, thus ensuring that no radiation ever got very far in the early universe without having its path changed. In a way, the universe was like an enormous crowd right after a popular concert; if you get separated from a friend, even if he is wearing a flashing button, it is impossible to see through the dense crowd to spot him. Only after the crowd clears is there a path for the light from his button to reach you.

The Universe Becomes Transparent

Not until a few hundred thousand years after the Big Bang, when the temperature had dropped to about 3000 K and the density of atomic nuclei to about 1000 per cubic centimeter, did the electrons and nuclei manage to combine to form stable atoms of hydrogen and helium (Figure 29.14). With no free electrons to scatter photons, the universe became transparent for the first time in cosmic history. From this point on, matter and radiation interacted much less frequently; we say that they *decoupled* from each other and evolved separately. Suddenly, electromagnetic radiation could really travel, and it has been traveling through the universe ever since.

Discovery of the Cosmic Background Radiation

If the model of the universe described in the previous section is correct, then—as we look far outward in the universe and thus far back in time—the first "afterglow" of the hot, early universe should still be detectable. Observations of it would be very strong evidence that our theoretical calculations about how the universe evolved are correct. As we shall see, we have indeed detected the radiation emitted at this photon decoupling time, when radiation began to stream freely through the universe without interacting with matter (Figure 29.15).

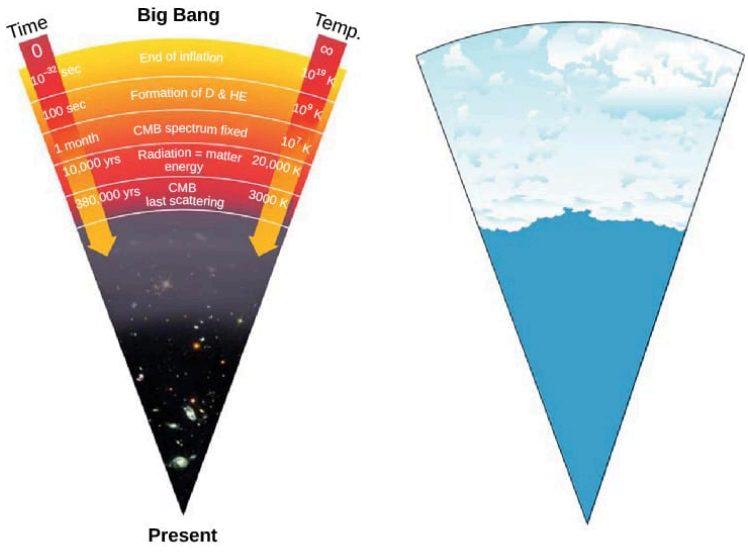


Figure 29.15 : Cosmic Microwave Background and Clouds Compared. (a) Early in the universe, photons (electromagnetic energy) were scattering off the crowded, hot, charged particles and could not get very far without colliding with another particle. But after electrons and photons settled into neutral atoms, there was far less scattering, and photons could travel over vast distances. The universe became transparent. As we look out in space and back in time, we can't see back beyond this time. (b) This is similar to what happens when we see clouds in Earth's atmosphere. Water droplets in a cloud scatter light very efficiently, but clear air lets light travel over long distances. So as we look up into the atmosphere, our vision is blocked by the cloud layers and we can't see beyond them. (credit: modification of work by NASA)

The detection of this afterglow was initially an accident. In the late 1940s, Ralph Alpher and Robert Herman, working with George Gamow, realized that just before the universe became transparent, it must have been radiating like a blackbody at a temperature of about 3000 K—the temperature at which hydrogen atoms could begin to form. If we could have seen that radiation just after neutral atoms formed, it would have resembled radiation from a reddish star. It was as if a giant fireball filled the whole universe.

But that was nearly 14 billion years ago, and, in the meantime, the scale of the universe has increased a thousand fold. This expansion has increased the wavelength of the radiation by a factor of 1000 (see Figure 29.7). According to Wien's law, which relates wavelength and temperature, the expansion has correspondingly lowered the temperature by a factor of 1000 (see the chapter on Radiation and Spectra). The cosmic background behaves like a blackbody and should therefore have a spectrum that obeys Wien's Law.

Alpher and Herman predicted that the glow from the fireball should now be at radio wavelengths and should resemble the radiation from a blackbody at a temperature only a few degrees above absolute zero. Since the fireball was everywhere throughout the universe, the radiation left over from it should also be everywhere. If our eyes were sensitive to radio wavelengths, the whole sky would appear to glow very faintly. However, our eyes can't see at these wavelengths, and at the time Alpher and Herman made their prediction, there were no instruments that could detect the glow. Over the years, their prediction was forgotten.

In the mid-1960s, in Holmdel, New Jersey, Arno Penzias and Robert Wilson of AT&T's Bell Laboratories had built a delicate microwave antenna (Figure 29.16) to measure astronomical sources, including supernova remnants like Cassiopeia A (see the chapter on The Death of Stars). They were plagued with some unexpected background noise, just like faint static on a radio, which they could not get rid of. The puzzling thing about this radiation was that it seemed to be coming from all directions at once. This is very unusual in astronomy: after all, most radiation has a specific direction where it is strongest—the direction of the Sun, or a supernova remnant, or the disk of the Milky Way, for example.



Figure 29.16 : Robert Wilson (left) and Arno Penzias (right). These two scientists are standing in front of the horn-shaped antenna with which they discovered the cosmic background radiation. The photo was taken in 1978, just after they received the Nobel Prize in physics.

Penzias and Wilson at first thought that any radiation appearing to come from all directions must originate from inside their telescope, so they took everything apart to look for the source of the noise. They even found that some pigeons had roosted inside the big horn-shaped antenna and had left (as Penzias delicately put it) “a layer of white, sticky, dielectric substance coating the inside of the antenna.” However, nothing the scientists did could reduce the background radiation to zero, and they reluctantly came to accept that it must be real, and it must be coming from space.

Penzias and Wilson were not cosmologists, but as they began to discuss their puzzling discovery with other scientists, they were quickly put in touch with a group of astronomers and physicists at Princeton University (a short drive away). These astronomers had—as it happened—been redoing the calculations of Alpher and Herman from the 1940s and also realized that the radiation from the decoupling time should be detectable as a faint afterglow of radio waves. The different calculations of what the observed temperature would be for this cosmic microwave background (CMB)[2] were uncertain, but all predicted less than 40 K.

Penzias and Wilson found the distribution of intensity at different radio wavelengths to correspond to a temperature of 3.5 K. This is very cold—closer to absolute zero than most other astronomical measurements—and a testament to how much space (and the waves within it) has stretched. Their measurements have been repeated with better instruments, which give us a reading of 2.73 K. So Penzias and Wilson came very close. Rounding this value, scientists often refer to “the 3-degree microwave background.”

Many other experiments on Earth and in space soon confirmed the discovery by Penzias and Wilson: The radiation was indeed coming from all directions (it was isotropic) and matched the predictions of the Big Bang theory with remarkable precision. Penzias and Wilson had inadvertently observed the glow from the primeval fireball. They received the Nobel Prize for their work in 1978. And just before his death in 1966, Lemaître learned that his “vanished brilliance” had been discovered and confirmed.

Link to Learning

You may enjoy watching *Three Degrees*, a 26-minute video from Bell Labs about Penzias and Wilson’s discovery of the cosmic background radiation (with interesting historical footage).

Properties of the Cosmic Microwave Background

One issue that worried astronomers is that Penzias and Wilson were measuring the background radiation filling space through Earth’s atmosphere. What if that atmosphere is a source of radio waves or somehow affected their measurements? It would be better to measure something this important from space.

The first accurate measurements of the CMB were made with a satellite orbiting Earth. Named the Cosmic Background Explorer

(COBE), it was launched by NASA in November 1989. The data it received quickly showed that the CMB closely matches that expected from a blackbody with a temperature of 2.73 K (Figure 29.17). This is exactly the result expected if the CMB was indeed redshifted radiation emitted by a hot gas that filled all of space shortly after the universe began.

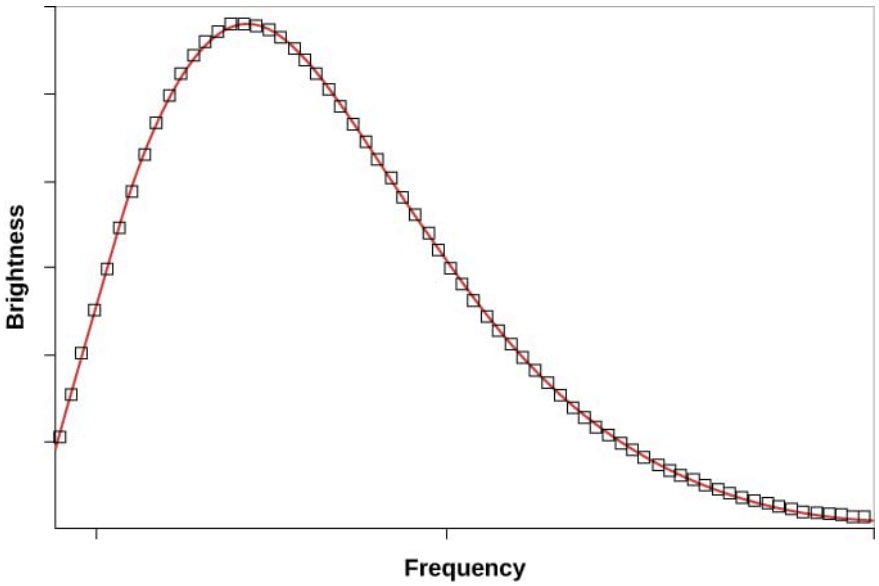


Figure 29.17 : Cosmic Background Radiation. The solid line shows how the intensity of radiation should change with wavelength for a blackbody with a temperature of 2.73 K. The boxes show the intensity of the cosmic background radiation as measured at various wavelengths by COBE's instruments. The fit is perfect. When this graph was first shown at a meeting of astronomers, they gave it a standing ovation.

The first important conclusion from measurements of the CMB, therefore, is that the universe we have today has indeed evolved from a hot, uniform state. This observation also provides direct support for the general idea that we live in an evolving universe, since the universe is cooler today than it was in the beginning.

Small Differences in the CMB

It was known even before the launch of COBE that the CMB is extremely *isotropic*. In fact, its uniformity in every direction is one of the best confirmations of the cosmological principle— that the universe is homogenous and isotropic.

According to our theories, however, the temperature could not have been *perfectly* uniform when the CMB was emitted. After all, the CMB is radiation that was scattered from the particles in the universe at the time of decoupling. If the radiation were completely smooth, then all those particles must have been distributed through space absolutely evenly. Yet it is those particles that have become all the galaxies and stars (and astronomy students) that now inhabit the cosmos. Had the particles been completely smoothly distributed, they could not have formed all the large-scale structures now present in the universe—the clusters and superclusters of galaxies discussed in the last few chapters.

The early universe must have had tiny density fluctuations from which such structures could evolve. Regions of higher-than-average density would have attracted additional matter and eventually grown into the galaxies and clusters that we see today. It turns out that these denser regions would appear to us to be colder spots, that is, they would have lower-than-average temperatures.

The reason that temperature and density are related can be explained this way. At the time of decoupling, photons in a slightly denser portion of space had to expend some of their energy to escape the gravitational force exerted by the surrounding gas. In losing energy, the photons became slightly colder than the overall average temperature at the time of decoupling. Vice versa, photons that were located in a slightly less dense portion of space lost less energy upon leaving it than other photons, thus appearing slightly hotter than average. Therefore, if the seeds of present-day galaxies existed at the time that the CMB was emitted, we should see some slight variations in the CMB temperature as we look in different directions in the sky.

Scientists working with the data from the COBE satellite did indeed detect very subtle temperature differences—about 1 part in 100,000—in the CMB. The regions of lower-than-average temperature come in a variety of sizes, but even the smallest of the colder areas detected by COBE is far too large to be the precursor of an individual galaxy, or even a supercluster of galaxies. This

is because the COBE instrument had “blurry vision” (poor resolution) and could only measure large patches of the sky. We needed instruments with “sharper vision.”

The most detailed measurements of the CMB have been obtained by two satellites launched more recently than COBE. The results from the first of these satellites, the Wilkinson Microwave Anisotropy Probe (WMAP) spacecraft, were published in 2003. In 2015, measurements from the Planck satellite extended the WMAP measurements to even-higher spatial resolution and lower noise (Figure 29.18).

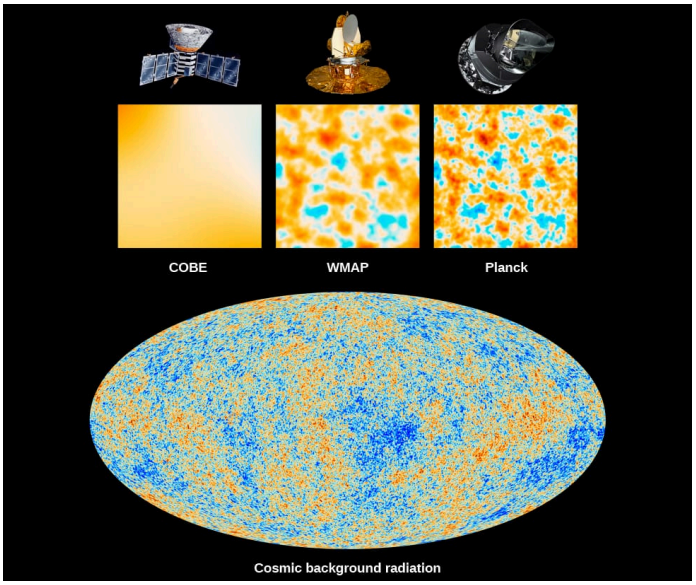


Figure 29.18 : CMB Observations. This comparison shows how much detail can be seen in the observations of three satellites used to measure the CMB. The CMB is a snapshot of the oldest light in our universe, imprinted on the sky when the universe was just about 380,000 years old. The first spacecraft, launched in 1989, is NASA’s Cosmic Background Explorer, or COBE. WMAP was launched in 2001, and Planck was launched in 2009. The three panels show 10-square-degree patches of all-sky maps. This cosmic background radiation image (bottom) is an all-sky map of the CMB as observed by the Planck mission. The colors in the map represent different temperatures: red for warmer and blue for cooler. These tiny temperature fluctuations correspond to regions of slightly different densities, representing the seeds of all future structures: the stars, galaxies, and galaxy clusters of today. (credit top: modification of work by NASA/JPL-Caltech/ESA; credit bottom: modification of work by ESA and the Planck Collaboration)

Theoretical calculations show that the sizes of the hot and cold spots in the CMB depend on the geometry of the universe and hence on its total density. (It’s not at all obvious that it should do so, and it takes some pretty fancy calculations—way beyond the level of our text—to make the connection, but having such a dependence is very useful.) The total density we are discussing here includes both the amount of mass in the universe and the mass equivalent of the dark energy. That is, we must add together mass and energy: ordinary matter, dark matter, and the dark energy that is speeding up the expansion.

To see why this works, remember (from the chapter on Black Holes and Curved Spacetime) that with his theory of general relativity, Einstein showed that matter can curve space and that the amount of curvature depends on the amount of matter present. Therefore, the total amount of matter in the universe (including dark matter and the equivalent matter contribution by dark energy), determines the overall geometry of space. Just like the geometry of space around a black hole has a curvature to it, so the entire universe may have a curvature. Let’s take a look at the possibilities (Figure 29.19).

If the density of matter is higher than the critical density, the universe will eventually collapse. In such a closed universe, two initially parallel rays of light will eventually meet. This kind of geometry is referred to as spherical geometry. If the density of matter is less than critical, the universe will expand forever. Two initially parallel rays of light will diverge, and this is referred to as hyperbolic geometry. In a critical-density universe, two parallel light rays never meet, and the expansion comes to a halt only at some time infinitely far in the future. We refer to this as a flat universe, and the kind of Euclidean geometry you learned in high school applies in this type of universe.

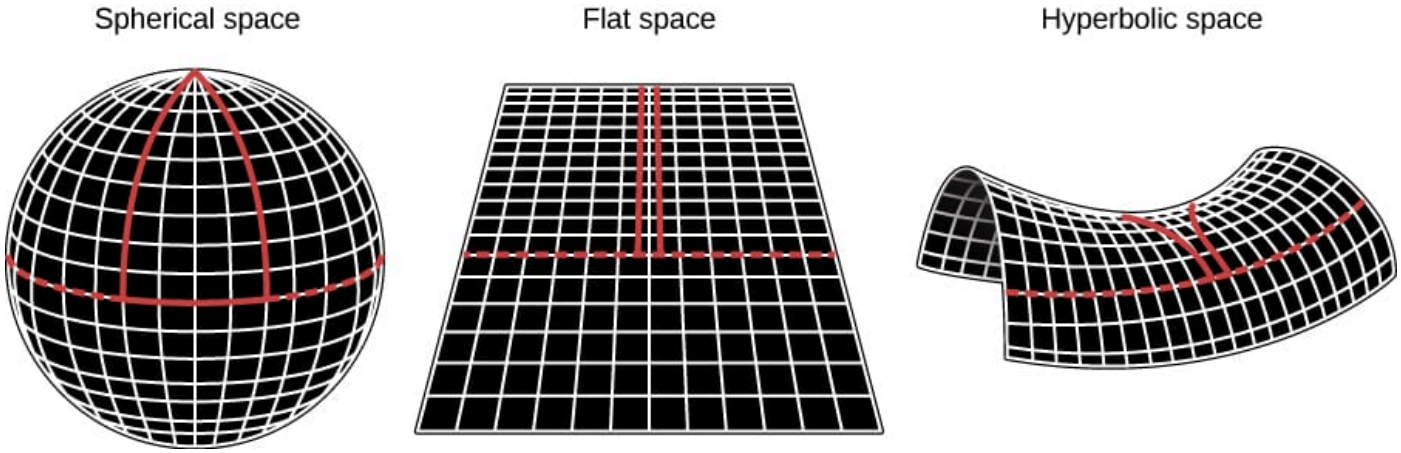


Figure 29.19 : Picturing Space Curvature for the Entire Universe. The density of matter and energy determines the overall geometry of space. If the density of the universe is greater than the critical density, then the universe will ultimately collapse and space is said to be *closed* like the surface of a sphere. If the density exactly equals the critical density, then space is *flat* like a sheet of paper; the universe will expand forever, with the rate of expansion coming to a halt infinitely far in the future. If the density is less than critical, then the expansion will continue forever and space is said to be *open* and negatively curved like the surface of a saddle (where more space than you expect opens up as you move farther away). Note that the red lines in each diagram show what happens in each kind of space—they are initially parallel but follow different paths depending on the curvature of space. Remember that these drawings are trying to show how space for the entire universe is "warped"—this can't be seen locally in the small amount of space that we humans occupy.

If the density of the universe is equal to the critical density, then the hot and cold spots in the CMB should typically be about a degree in size. If the density is greater than critical, then the typical sizes will be larger than one degree. If the universe has a density less than critical, then the structures will appear smaller. In Figure 29.20, you can see the differences easily. WMAP and Planck observations of the CMB confirmed earlier experiments that we do indeed live in a flat, critical-density universe.

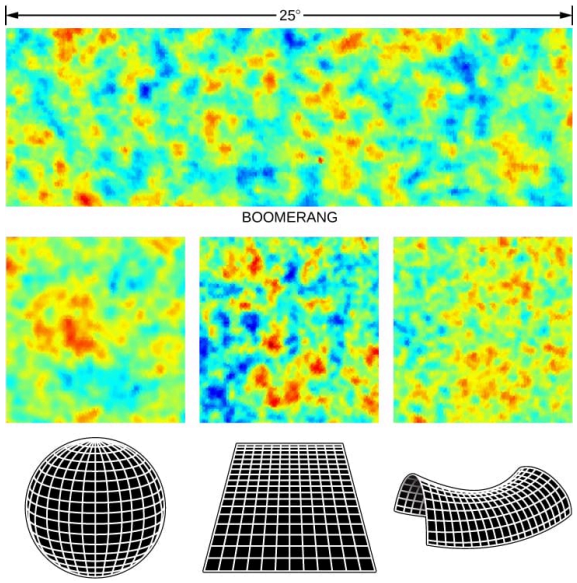


Figure 29.20 : Comparison of CMB Observations with Possible Models of the Universe. Cosmological simulations predict that if our universe has critical density, then the CMB images will be dominated by hot and cold spots of around one degree in size (bottom center). If, on the other hand, the density is higher than critical (and the universe will ultimately collapse), then the images' hot and cold spots will appear larger than one degree (bottom left). If the density of the universe is less than critical (and the expansion will continue forever), then the structures will appear smaller (bottom right). As the measurements show, the universe is at critical density. The measurements shown were made by a balloon-borne instrument called BOOMERanG (Balloon Observations of Millimetric Extragalactic Radiation and Geophysics), which was flown in Antarctica. Subsequent satellite observations by WMAP and Planck confirm the BOOMERanG result. (credit: modification of work by NASA)

Key numbers from an analysis of the Planck data give us the best values currently available for some of the basic properties of

the universe:

- Age of universe: 13.799±0.03813.799±0.038 billion years (Note: That means we know the age of the universe to within 38 million years. Amazing!)
- Hubble constant: 67.31±0.9667.31±0.96 kilometers/second/million parsecs (in the units we've been using, 20.65 kilometers/second/million light-years)
- Fraction of universe's content that is "dark energy": 68.5%±1.3%68.5%±1.3%
- Fraction of the universe's content that is matter: 31.5%±1.3%31.5%±1.3%

Note that this value for the Hubble constant is slightly smaller than the value of 70 kilometers/second/million parsecs that we have adopted in this book. In fact, the value derived from measurements of redshifts is 73 kilometers/second/million parsecs. So precise is modern cosmology these days that scientists are working hard to resolve this discrepancy. The fact that the difference between these two independent measurements is so small is actually a remarkable achievement. Only a few decades ago, astronomers were arguing about whether the Hubble constant was around 50 kilometers/second/million parsecs or 100 kilometers/second/million parsecs.

Analysis of Planck data also shows that ordinary matter (mainly protons and neutrons) makes up 4.9% of the total density. Dark matter plus normal matter add up to 31.5% of the total density. Dark energy contributes the remaining 68.5%. The age of the universe at decoupling—that is, when the CMB was emitted—was 380,000 years.

Perhaps the most surprising result from the high-precision measurements by WMAP and the even higher-precision measurements from Planck is that there were no surprises. The model of cosmology with ordinary matter at about 5%, dark matter at about 25%, and dark energy about 70% has survived since the late 1990s when cosmologists were forced in that direction by the supernovae data. In other words, the very strange universe that we have been describing, with only about 5% of its contents being made up of the kinds of matter we are familiar with here on Earth, really seems to be the universe we live in.

After the CMB was emitted, the universe continued to expand and cool off. By 400 to 500 million years after the Big Bang, the very first stars and galaxies had already formed. Deep in the interiors of stars, matter was reheated, nuclear reactions were ignited, and the more gradual synthesis of the heavier elements that we have discussed throughout this book began.

We conclude this quick tour of our model of the early universe with a reminder. You must not think of the Big Bang as a *localized* explosion *in space*, like an exploding superstar. There were no boundaries and there was no single site where the explosion happened. It was an explosion *of space* (and time and matter and energy) that happened everywhere in the universe. All matter and energy that exist today, including the particles of which you are made, came from the Big Bang. We were, and still are, in the midst of a Big Bang; it is all around us.

Footnotes

- 2Recall that microwaves are in the radio region of the electromagnetic spectrum.

This page titled 6.5: The Cosmic Microwave Background is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by OpenStax via source content that was edited to the style and standards of the LibreTexts platform.

7: EARTH'S ATMOSPHERE

7.1: Overview of Earth’s Atmosphere

7.2: Layers of the Atmosphere

7.3: Weather and Climate: What’s the Difference?

7.4: Radiation

7.5: Earth-Sun Energy Balance

7.5.1: Albedo

7.5.2: Surface Energy Balance

7.5.3: Radiation Interaction with the Atmosphere

7.6: Air Pollution

Chapter 7: Earth's Atmosphere — Introduction

In the astronomy portion of this course, we examined Earth as one world in a much larger universe. We now turn our attention toward Earth itself, beginning with the atmosphere that surrounds our planet. Although this marks the beginning of our study of meteorology, it also builds directly on ideas we have already encountered. Energy from the Sun drives many of the processes that occur in Earth's atmosphere, while the interaction of radiation with Earth's surface and atmosphere helps determine the planet's temperature.

In this chapter, we will examine the composition and structure of the atmosphere, distinguish between weather and climate, and explore how radiation moves through the Earth-Sun system. These ideas will provide the foundation for understanding the atmospheric processes and weather systems developed in the chapters that follow. We will also consider air pollution as an example of how changes in atmospheric composition can affect both the environment and human health.

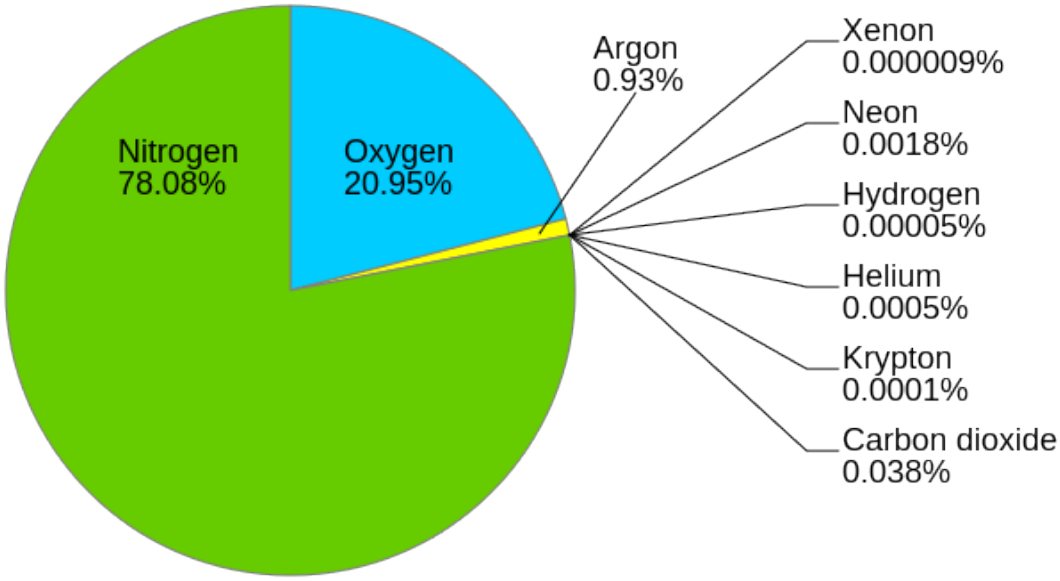
This page titled 7: Earth's Atmosphere is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Robert Wagner.

7.1: Overview of Earth's Atmosphere

Everything that happens on Earth is caused in some way by radiation from the sun, which is an average-sized star located near the edge of the Milky Way galaxy. It provides radiant energy (also called radiation or solar insolation) by converting hydrogen into helium near its core, which provides most of the Earth's warmth. Here is a classic song about the Sun heating Earth.

The Earth only receives a tiny portion of the sun's total energy output. It is this radiation that drives the atmosphere's wind and weather patterns on the Earth's surface. It is because of this solar input that the Earth can maintain an overall global average surface temperature of approximately 15°C (59°F).

Composition of the Atmosphere



Pie chart showing the gases that make up Earth's atmosphere (CC BY-SA 3.0).

Earth's atmosphere is primarily composed of nitrogen (N) and oxygen (O), with smaller quantities of other gases, as shown in the figure above. Nitrogen makes up around 78%, oxygen makes up around 21%, and Argon almost 1% of the total dry air volume in the atmosphere. Other gases (such as water vapor, the gaseous form of $\mathrm{H_2O}$) take up varying amounts depending on the location and atmospheric conditions. There is a balance of input and output of atmospheric gases at the Earth's surface both from life and surface processes. For example, when we breathe, the human body takes in oxygen ($\mathrm{O_2}$) and releases carbon dioxide ($\mathrm{CO_2}$). When plants photosynthesize, they take in carbon dioxide and release oxygen. When water ($\mathrm{H_2O}$) evaporates from the ocean, additional water vapor ($\mathrm{H_2O}$) is added to the atmosphere. When the Kilauea volcano degasses, sulfur dioxide ($\mathrm{SO_2}$) is added to the atmosphere.

Water Vapor

Water vapor concentrations vary greatly from location to location, and from time to time. In warm, moist, tropical regions, water vapor can make up nearly 4% of the atmospheric composition, while in polar regions it can be just a tiny fraction of a percent. Water vapor cannot be seen in its gaseous form, but you may see it as it *condenses* into water droplets on the side of a cool glass of water on a hot day or into water droplets that make up clouds. The process of water vapor changing phase to liquid is called *condensation*. When liquid water becomes a gaseous vapor, it is known as *evaporation*. When water becomes ice or ice becomes water, these processes are known as *freezing* and *melting,* respectively. When ice transitions directly to water vapor, the process is known as *sublimation*. The transition from water vapor to ice is known as *vapor deposition*. These terms will all be used frequently in subsequent chapters.

Water vapor is one of the most important gases in our atmosphere, because when it changes phase from vapor to liquid or ice, it releases enormous amounts of heat, called **latent heat**. Latent heat is a major source of energy in the atmosphere, particularly

for tropical storms and other types of convection. In addition, water vapor is an important greenhouse gas in the atmosphere, meaning it absorbs and re-releases a portion of the Earth's outgoing radiation, which allows our planet to remain warm.

If you don't understand convection, latent heat, and greenhouse gases yet, don't worry, all will be discussed in later chapters. The point is that despite its small fractional percentage, water vapor is arguably the most important gas in the atmosphere.

Hawaiian Focus Box

The air composition of Hawai'i occasionally faces a unique threat due to the emissions from Kīlauea, a volcano located along the southern shore of Hawai'i's Big Island. Kīlauea has been continuously erupting since 1983 and is the most active of the volcanoes that make up the Big Island. In the first few years after 1983, Kīlauea emitted up to 30,000 tons of sulfur dioxide (SO_2) a day, but that number has stabilized recently to around 5,000 tons per day. At present day with the 2018 Kilauea eruptions currently occurring, SO_2 emissions are again elevated. These emissions are known as volcanic smog, or more commonly, "vog". The sulfate particles in vog are very small, so they can effectively infiltrate human airways. In environments with high relative humidity, such as inside the human body, these particles will hydrolyze and expand, which irritates lungs and obstructs airways. Vog has been linked to respiratory disease, sinusitus, and even lung cancer. Therefore, Hawai'i's residents need to take precautions.



Water vapor and other gases escaping from the Kīlauea volcano at the Halema'uma'u vent (CC BY-SA 4.0).

Vertical Structure of the Atmosphere

Although the atmosphere extends vertically for hundreds of kilometers, almost 99% of it is within approximately 30 km of the Earth's surface. Air molecules are pulled toward Earth by the gravitational force, which pulls downward on the atmosphere. This causes the air molecules closer to Earth's surface to be more tightly compressed, meaning that there are more air molecules together in a given volume (a higher density). The greater the number of air molecules that exist above a certain altitude, the greater the effect of this compression, because the molecules are all being pulled downward together. Just as gravity has an effect on the weight of different objects (weight is the force that acts on an object due to gravity), it also gives weight to air.

$$\text{Weight} = \text{mass} \times \text{gravitational acceleration}$$

$$\text{Weight} = m \times g$$

Mass is the measure of how much matter exists in a given object or space. Mass is occasionally confused with weight. While weight increases with more mass, mass would have no weight at all without the pull of gravity. Similarly, an object's weight depends on the gravitational force: an object weighs much less on the moon than it does on Earth. Mass is typically given in grams (g) or kilograms (kg).

Note

You may want to review the International System (IS) of Units. The base units we will use include meters (m), kilograms (kg), seconds (s), and Kelvin (K). We'll also use the prefixes "kilo" (1000), "hecto" (100), and "micro" (10^{-6}) along with some derived units like hertz, Newtons, pascals, joules, and watts.

Air Density

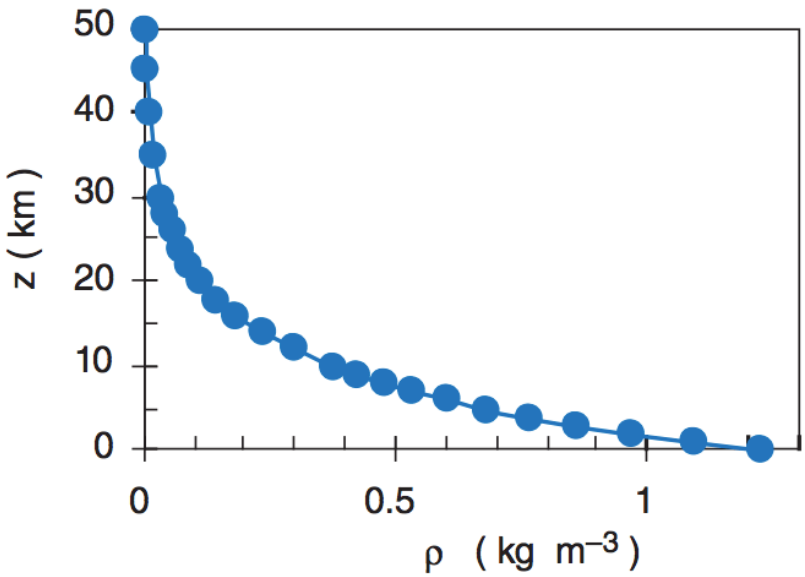
Air density is determined by the amount of mass (m) that exists in a given space, or volume (V). If you fit a lot of mass into a tiny volume, you will have higher density. If you have only a small amount of mass spread out over a large volume, the density will be lower.

$$\text{Density} = \frac{\text{mass}}{\text{volume}}$$

$$\rho = \frac{M}{V}$$

Putting these ideas together, air density is highest near the surface of Earth and decreases with height, because there are more molecules held tightly together at the surface by gravity than there are above. The standard unit for density is kg·m^{-3}.

Pro Tip: Density is often denoted by the Greek lowercase letter ρ (rho). Be careful not to confuse this with *p* or *P*, which are both used to denote pressure. When writing on paper, it may be useful to write *ρ* with a curly tail like a backwards *q* and to write *p* with a straight tail in order to differentiate the two.



Density (ρ, kg·m^{-3}) vs. height (z, km) in the atmosphere (CC BY-NC-SA 4.0).

The image above shows the distribution of density (ρ) with altitude. Atmospheric surface density is typically around 1.2 kg·m^{-3}. We call this value ρ_0 because it is the initial value at the surface. The exponential equation below approximates the distribution of density with height

$$\rho(z) = \rho_0 \times \exp(-z/H)$$

$$H = \frac{R_d \times T}{g}$$

where ρ (kg m^{-3}) is density, ρ_0 (kg m^{-3}) is the density at the Earth's surface, z (m) is altitude, T (K) is temperature (assumed to be constant through the atmosphere), R_d = 287.053 $J \cdot K^{-1} \cdot kg^{-1}$ (gas constant for dry air), and g is the acceleration due to gravity ($m \cdot s^{-2}$).

Note

Remember that the equations you see in this text are good approximations. They all involve assumptions of some kind. In the case of the exponential distribution of density with altitude, the primary issue is the assumption that temperature is constant throughout the atmosphere. This will become clear later.

"Scale height" is denoted by H and represents an e-folding distance for the drop off of density in the atmosphere. Typically the value of H is around 8000 m. This means that density at 8 km is approximately 1/e (e=2.71828) of the value at the surface. A back of the envelope calculation gives the density of the atmosphere at 8 km as 1.2 $kg \cdot m^{-3}$ divided by 3 and is approximately equal to 0.4 $kg \cdot m^{-3}$. Checking the image above, it is clear that this provides a relatively good estimate for density.

Air Pressure

You will often hear meteorologists on TV discuss the air pressure, or you might see H's and L's on a map displayed on the Weather Channel, denoting high and low pressure areas. Because air molecules are in constant motion, they collide with one another and other objects up to several billions of times a second. Each time an air molecule collides with an object, it exerts a tiny amount of force. Air pressure refers to the total force that air exerts against a given area of an object.

$$\text{Pressure} = \frac{\text{Force}}{\text{Area}}$$

$$P = \frac{F}{A}$$

A Newton (N) is the unit for force and m^2 is the unit for area, in the International System of Units (SI). Therefore, the standard unit for pressure is in Newtons per square meter — or Pascals (Pa), which are defined as 1 $N \cdot m^{-2}$.

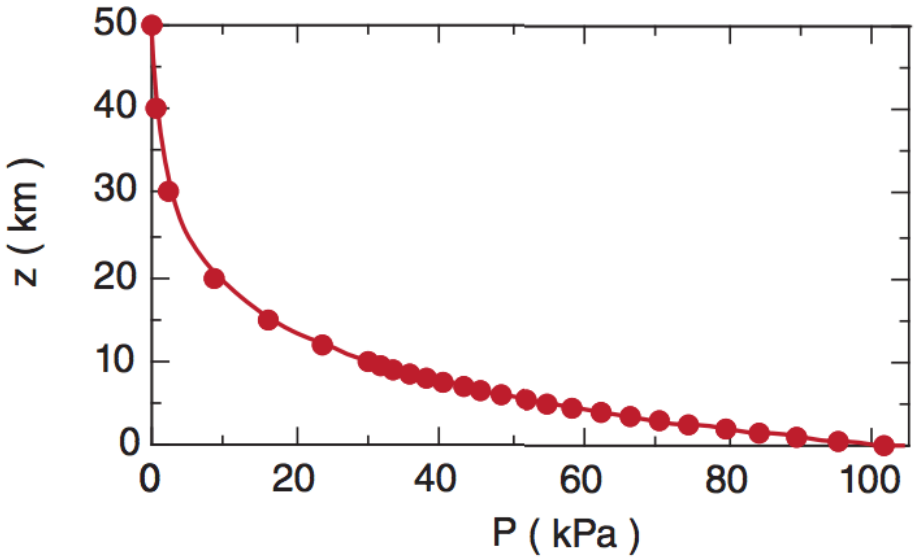
Pro Tip: You will generally find air pressure expressed in units of millibars or (mb) or hectopascals (hPa) on weather maps. These two units are equivalent to one another. In the *Practical Meteorology: An Algebra-based Survey of Atmospheric Science* textbook by Roland Stull off of which this OER is based, kPa are often used. In the aviation field, Inches of mercury (inHg) are also commonly used. At sea level, the global average for atmospheric pressure is:

101.325 kPa = 1013.25 mb = 1013.25 hPa = 29.92 in. Hg. = 1 atm (atmosphere) = 101325 Pa.

In future calculations, you will usually need to express pressure in Pascals (Pa) for your units to cancel out. Always be mindful of the units you are given and be sure you are able to convert from one type of unit to another.

You can think of the surface air pressure as the total weight of a column of air molecules extending from the surface to the top of the atmosphere.

Because there are more air molecules at the surface of the Earth and less above, air pressure is maximized at the surface and decreases with height nearly exponentially, analogous to air density.



Height (z, km) vs. pressure (P, kPa) in the atmosphere (CC BY-NC-SA 4.0).

The image above shows the distribution of pressure (P) with altitude. Atmospheric surface pressure is typically around 1013 hPa. We call this value P_0 because it is the initial value at the surface. The exponential equation below approximates the distribution of pressure with height

$$P(z) = P_0 \times \exp\left(\frac{-z}{H}\right)$$

$$H = \frac{R_d \times T}{g}$$

where P (Pa) is pressure, P_0 (Pa) is the pressure at the Earth's surface, z (m) is altitude, T (K) is temperature (assumed to be constant through the atmosphere), R_d = 287.053 $J{\cdot}K^{-1}{\cdot}kg^{-1}$ (gas constant for dry air), and g is the acceleration due to gravity ($m{\cdot}s^{-2}$).

You can get a sense of atmospheric pressure if you've ever dived more than a few feet underwater at the pool or beach. As you dive deeper, the weight of the water above you increases, and you feel increased discomfort or pressure on your head. This is why divers require special equipment to reach greater depths in the ocean. If you think of yourself at the bottom of an ocean of air, you can imagine why air pressure is highest down here at the surface. If you've ever flown on an airplane, you can also feel the changes in atmospheric pressure in your ears as you ascend and descend. Still, while pressure decreases at high elevations in an airplane, cabins are pressurized to maintain similar pressure to Earth's surface so you are not experiencing the full pressure drop with altitude.

Air Temperature

It is commonly thought that air temperature is a measure of how cold or hot an object is. We instinctively know that the higher an object's temperature, the hotter it will be, and the lower its temperature, the colder it will be. But what does temperature really measure? The answer lies in the motion of the molecules.

Air molecules are in constant motion, colliding with objects and one another. As we increase the temperature of a volume of air, the air molecules speed up, and collisions become more frequent. As we decrease the air temperature, the molecules slow down, and collisions occur less frequently. What is going on here? Simply put, air temperature is a relative measure of the average kinetic energy (kinetic meaning it relates to motion) of the molecules of a system.

Because there are more air molecules close to the Earth's surface, density and air pressure are maximized at the surface and decrease with height. Based on this, do you think air temperature will decrease or increase as you move further away from the surface?

Within the lowest 10-12 km of the atmosphere, temperature tends to decrease with height, primarily because the Earth's surface

is warmed by sunlight, which then warms the layer of air directly above it, which warms the air above that, and so on. While this is true in the lowest surface layer of the atmosphere, this is not true throughout the rest of the atmosphere. The vertical temperature profile is more complicated than the vertical profile of pressure or density. Temperature is different for each of the different layers of the atmosphere, which will be discussed later.

Temperature Scales

Surface air temperature is commonly given in degrees **Fahrenheit** (°F) in the United States. If you were born and raised in the US, this is the temperature scale that you probably use in your daily life, and what you'll see given when you look at the daily weather. You'll know that 32°F is the temperature at which water freezes, and 212°F is the temperature at which water boils.

Degrees **Celsius** (°C) is also commonly used, especially internationally. Celsius is a convenient scale to use because water freezes at 0°C and boils at 100°C. A difference in 1°C is larger than a difference of 1°F by about 1.8 times. To convert between the two, use the following equations.

$$^{\circ}C = \frac{5}{9} \times (^{\circ}F - 32)$$

$$^{\circ}F = \left(\frac{9}{5} \times {}^{\circ}C\right) + 32$$

The **Kelvin** scale (K) is almost always used as a temperature unit in scientific equations and is convenient in that it contains no negative numbers. The Kelvin scale begins at 0 K, or absolute zero, where atoms and molecules would theoretically be thermally motionless. The Kelvin scale is also sometimes known as the absolute temperature scale. The lowest possible temperature is 0 K but it does not occur naturally. The coolest naturally existing place known in the universe is the Boomerang Nebula, located in the Centaurus constellation about 5,000 light-years away from Earth. The temperature is measured at 1 K, only 1°C above absolute zero. A difference in 1 K is the same as a difference of 1°C, so a conversion is linear and simple.

$$K = {}^{\circ}C + 273.15$$

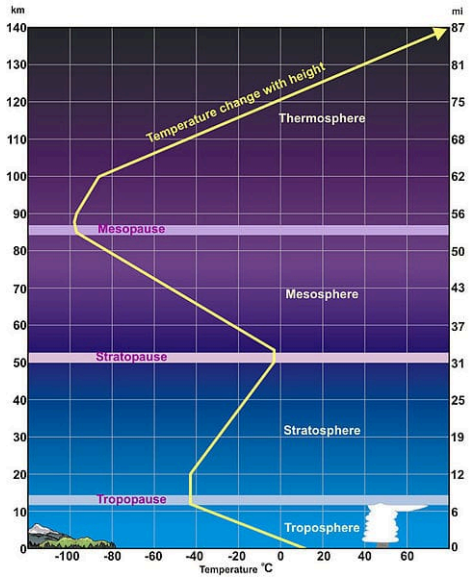
Based on this, absolute zero is -273.15 °C. Keep in mind degrees Celsius (°C) and degrees Fahrenheit (°F) always have the degree ° symbol in front, but Kelvin (K) never uses this symbol. Typical values of temperature on Earth's surface on the Kelvin scale are values around 260-310 K.

Within atmospheric sciences, the Kelvin and Celsius scale are used. This chapter is one of the only times you'll see a discussion of Fahrenheit within the course.

This page titled 7.1: Overview of Earth's Atmosphere is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by LibreTexts.

7.2: Layers of the Atmosphere

In the previous sections, we talked about how both air pressure and density decrease with height, because most of the atmosphere is held close to the Earth's surface. This change in pressure and density occurs quickly at first, but slows down at higher altitudes. Unfortunately the change in temperature with altitude is not nearly as straightforward. When we look at a vertical profile of the atmosphere, we see that it can be divided into different layers in a number of ways. We can define layers by how the air temperature varies with height, by the gas composition, and even by the electrical properties of each layer.



The vertical temperature structure of Earth's atmosphere (Public Domain) in kilometers (left axis) and miles (right axis). The layers are labeled as "spheres" and the boundaries between the layers are labeled as "pauses". For example the troposphere is the lowest layer and the tropopause is the boundary between the troposphere and the stratosphere. The yellow line shows temperature in °C, and the cloud indicates where most of the weather occurs.

As can be seen in the above figure, the air temperature decreases with height up to the tropopause, 11 km in this example. This occurs because radiation from the sun warms the Earth's surface, and the surface warms the air just above it. This will be discussed in further detail in later chapters.

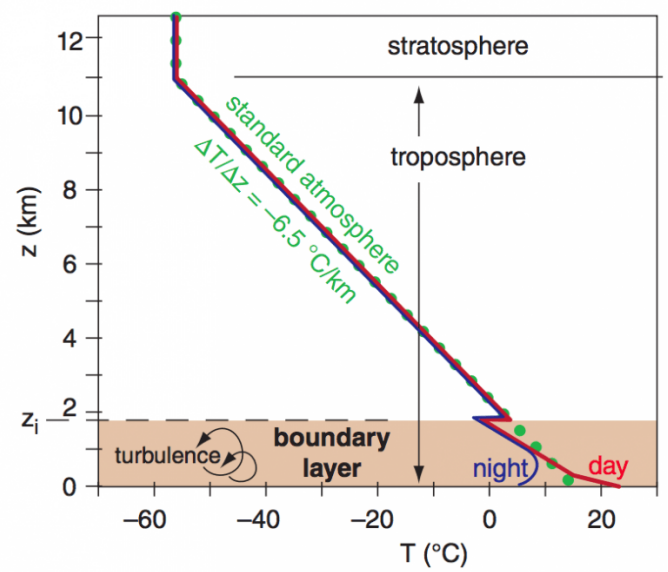
Troposphere

The layer of the atmosphere that we are most familiar with is the troposphere, which extends from the surface to around 11 km. All of the every day weather that we experience on Earth happens within the troposphere, which is characterized by frequent rising and sinking vertical air motions. The troposphere gets its name from the root word "tropo" in Greek, which means turning.

At the top of the troposphere, the air temperature stops decreasing with height in a region known as the tropopause, which separates the troposphere and the stratosphere above. The height of the tropopause varies depending on the season and location. In warmer areas near the equator, the tropopause tends to be higher (around 17 km), while in colder polar regions the tropopause is lower (around 9 km) because warm layers of air are thicker than layers of cold air. For the same reason, the tropopause is found at higher elevations in the summer, and at lower elevations in the winter. Aircraft fly at the tropopause height.

Atmospheric Boundary Layer

The **atmospheric boundary layer** (ABL) is located within the lowest 0.3 to 3 km of the troposphere and is affected in several ways due to its close proximity with the Earth's surface. Airflow closest to the surface slows down due to the effect of friction, which can be amplified due to the type of terrain or amount of vegetation. Because of this, the boundary layer experiences the greatest amount of turbulence in the atmosphere. In addition, the boundary layer is warmed by the Earth's surface during the daytime, so it is the layer that is most affected by the diurnal (day-night) heating cycle.



The change in temperature with altitude in the troposphere which is the lowest layer of Earth's atmosphere. The turbulent boundary layer near the surface is also seen (shaded tan), along with the standard atmospheric lapse rate (-6.5 $K \cdot km^{-1}$), and near-surface day (red) / night (blue) temperature differences (CC BY-NC-SA 4.0).

Stratosphere

In the *stratosphere*, the air temperature increases with height, causing a temperature inversion. This inversion layer tends to keep rising and sinking tropospheric air from mixing with stratospheric air. It also prevents rising and sinking from happening in the stratosphere itself. Because of this, it is called a *stratified* layer. The root word "strato" means layered, or spreading out, which is a good way to describe the many layers of the stratosphere. The temperature of the stratosphere increases with height because of the presence of a gas called ozone (O_3), which heats the air through the absorption of ultraviolet (UV) radiation from the sun.

At around 50 km, where the stratosphere is warmest (due to most of the UV radiation absorption occurring in the uppermost parts of the layer). The boundary known as the *stratopause* separates the stratosphere below from the mesosphere above.

Mesosphere

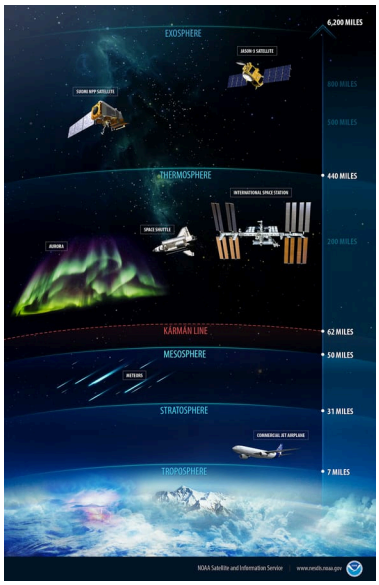
In the *mesosphere*, the air temperature once again decreases with height and the air is extremely thin with a low density. The average atmospheric pressure in this layer is about 1 mb, which means that about 99.9% of all air molecules are located below this level and only about one thousandth of the atmospheric mass is located above. You would not survive very long in the mesosphere, however, due to the extreme thinness of the air, freezing temperatures, and direct exposure to ultraviolet radiation. In addition, your blood would begin to boil at normal body temperatures if it were directly exposed to the low air pressure. The reason that temperatures decrease with height in the mesosphere is due partially to the fact that there is not a lot of ozone to absorb UV radiation. Because of this, the air molecules in the mesosphere lose more energy than they absorb.

At about 85 km near the *mesopause,* the atmosphere is at its coldest at about -90°C. The mesopause separates the mesosphere from the thermosphere above.

Thermosphere

In the *thermosphere*, the air density is extremely low so even a small amount of radiation absorption can lead to a large increase in temperature. Molecules can travel for entire kilometers without colliding into another molecule. This is the layer where auroras occur as a result of interactions between charged solar particles and air molecules.

At the top of the thermosphere, at above 500 km above the Earth's surface, molecules can actually escape the gravitational pull of the Earth. This region is known as the *exosphere*, and represents the upper limit of the atmosphere.



This figure illustrates the relative heights of different phenomena in the various vertical layers of the atmosphere. The Kármán Line at 62 miles (100 km) designates the boundary between Earth's atmosphere and outer space. Image by NOAA Satellites (Public Domain).

The previous figure shows the levels in Earth's atmosphere and the phenomena and activities that occur there. For example, commercial jet airplanes fly at the base of the stratosphere, while the International Space Station is positioned in the thermosphere.

Chapter 1 contained a vast array of topics, from defining temperature and pressure, to describing atmospheric vertical structure and components. As a reminder, these were our learning goals:

1. Convert between temperature units of Fahrenheit, Celsius, and Kelvin
2. Use mathematical formulas to define atmospheric temperature, pressure, and density
3. Compute pressure and density changes with altitude
4. Describe the vertical structure of Earth's atmosphere
5. Define and apply the ideal gas law
6. Describe hydrostatic balance
7. Discuss the difference between weather and climate
8. Note the location of terminology, coordinate systems, and units for future reference

The next section will discuss weather vs. climate and various terms and coordinate systems that are useful to review.

This page titled 7.2: Layers of the Atmosphere is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by LibreTexts.

7.3: Weather and Climate: What's the Difference?

Occasionally the terms "weather" and "climate" get confused with one another. Climate change has been a hot topic in recent decades. Knowing this, you might conclude that when the climate changes — it must be a pretty big deal, however, the weather is in a constant state of change. If it weren't, we would have little need to check the Internet or local weather station to see what the sky is going to look like tomorrow. Nobody bats an eye when the weather is different from one day to the next. What do we mean exactly by the terms weather and climate?

Weather refers to the present condition of the atmosphere at any given time and place. This includes the following elements.

Air Temperature — how hot or cold the air is.
Air Pressure — the force the air exerts on the surface.
Humidity — the amount of water vapor present in the air.
Clouds — masses of water droplets and/or ice crystals that obscure parts of the sky.
Precipitation — water (solid or liquid) that falls from clouds and reaches the ground
Visibility — the maximum horizontal distance that can be seen. This can be affected by the presence of fog or precipitation.
Wind — the horizontal flow of air, caused by local differences in air pressure.

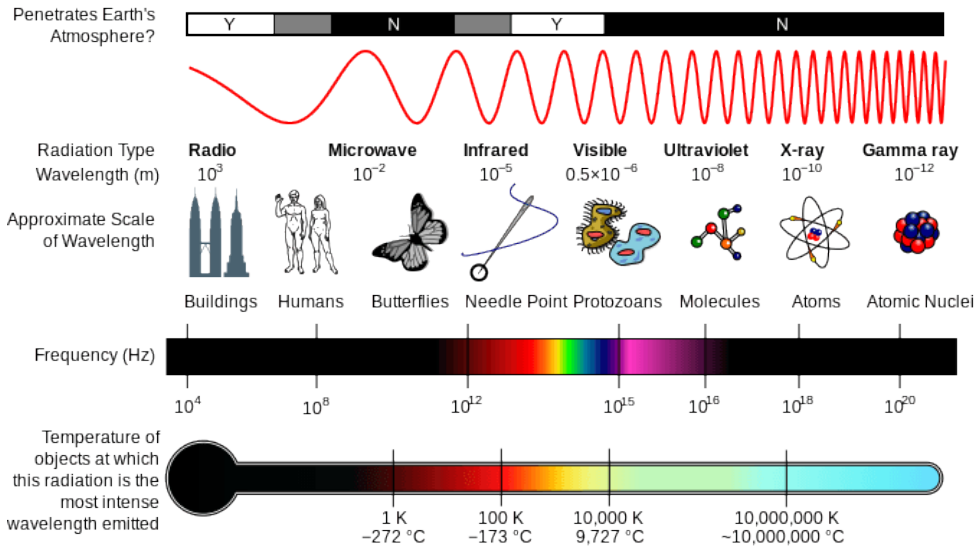
Climate represents the average range of weather events in a region over a long period of time including weather extremes such as heat waves or cold spells, as well as the frequency of these events.

One way to think of it is weather influences the kind of clothing we might wear that day, e.g., Will I need a raincoat? Is it too warm to wear long sleeves? Is it too cool to wear shorts? Climate, on the other hand, influences what clothing we buy. For example, in Hawai'i, it isn't likely that you would need to buy a heavy winter jacket.

This page titled 7.3: Weather and Climate: What's the Difference? is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by LibreTexts.

7.4: Radiation

Radiation can be thought of in two ways: electromagnetic waves or as photons. For the purpose of atmospheric science, we will generally consider radiation as a wave rather than a photon particle. **Electromagnetic radiation** is a type of energy produced by electric and magnetic fields, taking a variety of names depending on the wavelength. For example, you've probably heard of radio waves and x-rays, whereby the primary difference is the wavelength of the waves in each. The next image shows the relationship between wavelength, frequency, and temperature. We will go into further detail later, but colder objects radiate at lower frequencies and longer wavelengths, and warmer objects radiate at higher frequency and shorter wavelengths. This relationship holds true across all scales, from atomic nuclei to planets, and across temperatures from near absolute zero to 10s of millions of Kelvins.



The electromagnetic energy spectrum, including penetration of Earth's atmosphere (top), wavelength visual (red line), radiation type (name given), wavelength value (m), wavelength scale (image), frequency (Hz), and temperature of black body objects that emit at that wavelength of radiation (CC BY-SA 3.0).

This page titled 7.4: Radiation is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by LibreTexts.

7.5: Earth-Sun Energy Balance

In the prior section, the discussion turned rather technical. We went down a rabbit hole of radiation wavelengths, total emitted power per area, and the assumption of black bodies. We did this so that we could understand basic relationships about the energy balance in the Earth-Sun system. The Sun's average temperature is above 5,000 K while the Earth's average temperature is in the range 210-310 K (we will discuss this further in a later chapter). This means that the Sun and Earth radiate energy very differently.

The Sun emits solar radiation, also known as *ultraviolet radiation* or *shortwave radiation*. The Earth emits *infrared radiation* or *longwave radiation*. This follows directly from the electromagnetic energy spectrum and the respective temperatures of the Sun and Earth. The Sun emits radiation at a shorter wavelength than the Earth because it has a higher temperature, and Planck's curve for higher temperatures peaks at shorter wavelengths. It is for this reason that Earth's radiation is referred to as longwave, and the Sun's radiation is called shortwave.

We learned that a black body absorbs all incoming radiation and emits the maximum possible radiation given its temperature. In the Earth-Sun system, this means that the Earth absorbs all incoming radiation from the Sun, and emits the maximum amount given its temperature. In practice it is a little bit more complicated than this.

This page titled 7.5: Earth-Sun Energy Balance is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by LibreTexts.

7.5.1: Albedo

Not all incoming solar radiation is absorbed by the Earth because the Earth is not a perfect black body. Instead of absorbing all incoming radiation, some of it is reflected. Reflection refers to radiative energy bouncing back away from an object. We can define **albedo** (α) as the ratio of the amount of radiation reflected from an object to the amount of radiation received by an object.

$$\text{Albedo } = \alpha = \frac{\text{Reflected}}{\text{Incident}}$$

Pro Tip: The simplest way to think of albedo is based on the color of the object. Objects that are white colored are highly reflective and have a high albedo because the amount of reflected light is large compared to the amount of incoming incident light. You'll know this if you've been on a white sandy beach or a snow field and felt the brightness of the environment. Following the same logic, dark surfaces absorb light so black, brown, or dark green surfaces typically have a low albedo. You'll know this if you've walked across a black paved surface and felt the temperature difference between the blacktop and the white painted parking space demarcations.

Earth has an albedo of about 0.3 or 30%. This is an average that takes into account the high albedo and high latitude regions that are snow covered as well as the clouds and the much lower albedo oceans. Note that radiation reflected from an object does not warm the object.

This page titled 7.5.1: Albedo is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by LibreTexts.

7.5.2: Surface Energy Balance

Arguably the most important aspect to consider about the Earth-Sun system is the energy balance. In **steady-state**, the amount of incoming energy should equal the amount of outgoing energy (Net Radiative Flux=$F^* = 0$).

Let's start with the incoming solar radiation. The *solar constant* "S" is approximately equal to 1361 W·m^{-2}. This value is a rough estimate of the amount of energy per area received by the Earth from the Sun, but it is not exact. We call it a solar "constant" but it can be slightly lower or higher at times. Radiation emitted from a spherical source, like the sun, decreases by the square of the distance from the sphere's center. This is called the inverse square law.

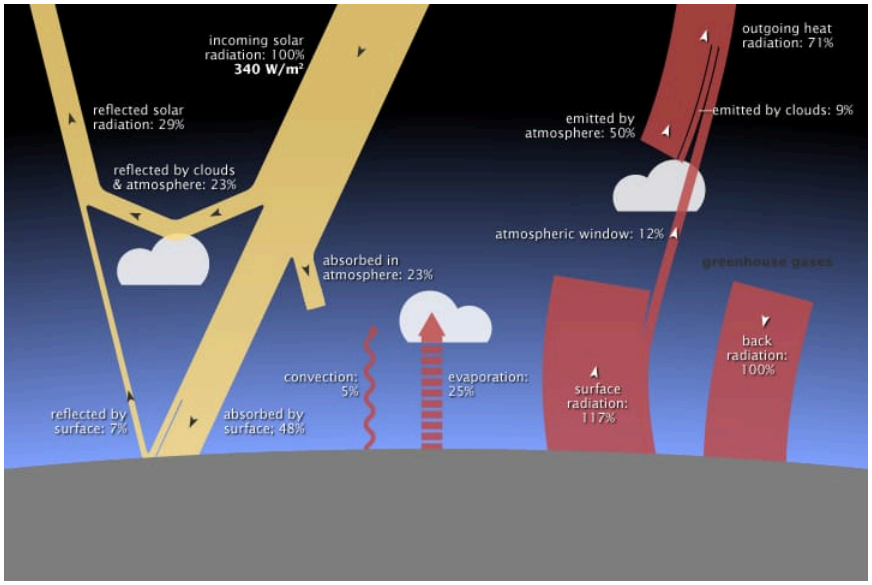
The following equation is a basic budget equation. Net radiation (F^*) is equal to the incoming solar radiation ($K\downarrow$), the reflected solar radiation ($K\uparrow$), the longwave radiation emitted by the Earth ($I\uparrow$), and the downwelling longwave radiation emitted from the atmosphere ($I\downarrow$) received by the Earth's surface.

$$F^* = K\uparrow + K\downarrow + I\uparrow + I\downarrow$$

Some of the sunlight that reaches the surface of the Earth is reflected, depending on the albedo, $K\uparrow$ is often written in the following way:

$$K\uparrow = -\alpha \times K\downarrow$$

This is a brief introduction to a surface energy balance model. As you may imagine, it can become much more complicated depending on the factors involved. It also strongly depends on the number of layers considered in the model. We will discuss this further in a later chapter.



Earth's radiative balance, incoming and outgoing radiation. (NASA Public Domain). Description in the text below.

For now, you should understand that incoming solar radiation is called shortwave radiation and is in the ultraviolet and visible portions of the electromagnetic spectrum because of the emission temperature of the Sun. When solar radiation interacts with the Earth, it is partially absorbed by the Earth's surface, and partially reflected, depending on the albedo of the surface. In the diagram above, you can see that some of the incoming solar radiation is reflected by clouds, some is reflected by the Earth's surface, but most is absorbed by the Earth's surface or the atmosphere.

You should also understand that Earth emits radiation too. However, it is at a lower intensity and a much longer wavelength, which is called the infrared portion of the electromagnetic spectrum because of the lower emission temperature of the Earth. Radiation is emitted by the Earth's surface, and by the atmosphere. We'll go into more detail on this later.

Pro Tip: Many synonyms were discussed above. When referring to radiation from the sun you may hear the following words used: solar radiation, shortwave radiation, and ultraviolet radiation. When referring to radiation from the Earth you may hear the following words used: infrared radiation, longwave radiation.

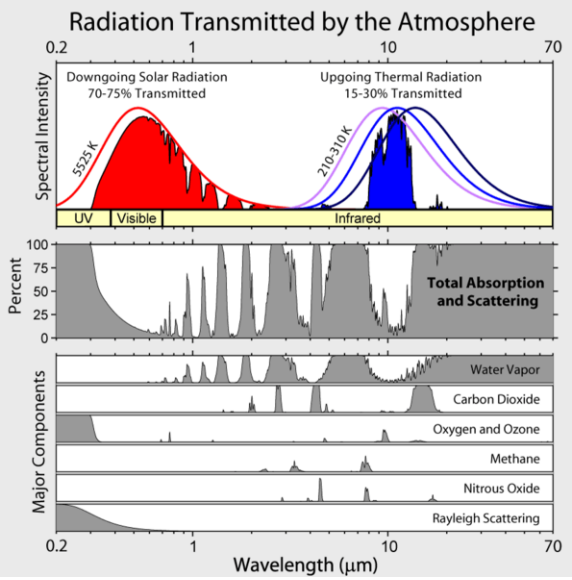
Earth's average temperature remains relatively constant as there is a balance of outgoing radiation and incoming radiation.

This page titled 7.5.2: Surface Energy Balance is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by LibreTexts.

7.5.3: Radiation Interaction with the Atmosphere

The diagram below shows the spectral intensity of downgoing solar radiation (UV and visible in red) and upgoing thermal radiation (IR in blue). The panels below the spectral intensity graph show the total percentage of radiation absorbed and scattered by the atmosphere as a function of wavelength, and divided up by the primary greenhouse gases in the atmosphere. **Greenhouse gases** are gases in the Earth's atmosphere that can absorb and emit radiation, primarily infrared radiation.

The greenhouse gas that interacts with radiation across the largest number of wavelengths is water vapor (H_2O), which interacts strongly with the atmosphere, especially in the longer wavelength portion of the spectrum. Other gases, such as Carbon Dioxide (CO_2), Oxygen (O_2), Ozone (O_3), Methane (CH_4), and Nitrous Oxide (NO) also interact with radiation.



Radiation transmitted by the atmosphere as a function of wavelength. Downgoing solar radiation is on the left, and upgoing thermal radiation is on the right, with the percentage of radiation absorbed and scattered by the Earth's atmosphere below. Finally, the percentage of radiation absorbed and scattered is then divided by the major greenhouse gases in the panels at the bottom (CC BY-SA 3.0).

In the panel of the figure above showing the total percentage of radiation that is absorbed by the Earth's atmosphere, you'll notice white gaps. These are called **Atmospheric Windows** where the wavelength of radiation is able to pass through the atmosphere without interaction with greenhouse gases. Particularly notable is the visible portion of the electromagnetic energy spectrum that can pass through the atmosphere unimpeded. Also notable is the large amount of interference in the infrared portion of the spectrum. Earth's atmosphere is practically transparent for shortwave radiation, and it strongly absorbs infrared radiation. This will be important for climate factors discussed in a later chapter.

> **Open-ended Question**: Why do you think visible light (the peak of down-going solar radiation) is centered on an atmospheric window?

The interaction of atmospheric gases in Earth's atmosphere is an important part of its energy balance. The *greenhouse effect* is the result of greenhouse gases absorbing and emitting outgoing infrared radiation emitted from Earth's surface. Because Earth's outgoing longwave radiation is partially absorbed by the atmosphere, this has a warming effect on Earth's surface making it warmer than it would be otherwise. While the greenhouse effect has a bad reputation, this process has in fact allowed Earth to be habitable.

This page titled 7.5.3: Radiation Interaction with the Atmosphere is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by LibreTexts.

7.6: Air Pollution

Air Pollution Defined

Air pollution is the release of gases that cause harm to all of Earth's other spheres (lithosphere/geosphere, biosphere, hydrosphere), which include humans, into the air. Air pollution causes millions of people to die every year. According to the World Health Organization (WHO), 99% of humans breathe air that exceeds the guideline limits for pollutants. People living in second- and third-world countries suffer the most. The Clean Air Act was established in the United States in 1970. It authorizes the U.S. Environmental Protection Agency (EPA) to safeguard public health by regulating the emissions of these harmful air pollutants. Progress has been made, but there is more to go.

Causes and Types of Air Pollution

Most air pollution comes from energy use and production. This includes driving a gasoline powered car, heating a home with natural gas, running a plant on fracked or otherwise sourced gas, etc. In all these examples a fossil fuel is burned, and gases are released into the atmosphere. More information on specific types of pollution is below.

Greenhouse Gases and Ozone

Greenhouse gases are gases that trap the Earth's heat in the atmosphere, leading to warmer temperatures, which in turn lead to things like rising sea levels and more extreme weather. Common greenhouse gases are water vapor (H_2O), carbon dioxide (CO_2), methane (CH_4), and ozone (O_3). In 2021, carbon dioxide accounted for about 80% of the US's total greenhouse gas emissions, and methane made up more than 11%. Carbon dioxide comes from the burning of fossil fuels. Methane comes from natural (cows produce methane as they digest food) and industrial (decomposition of waste) sources, and include large amounts released during oil and gas drilling. While carbon dioxide emissions are very high, methane is more potent, so it is very dangerous.

Ozone is ozone no matter where it is in the atmosphere. Good ozone is good only because it is in the stratosphere where we cannot breathe it and it absorbs ultraviolet (UV) rays. Bad ozone is near the Earth's surface where we can breathe it. It also absorbs solar ultraviolet light, but but the dangers outweigh the benefits. For UV protection, we are interested in the total number of ozone molecules between us and the Sun. 90% of ozone molecules are in the stratosphere and 10% are in the troposphere - some down near Earth's surface where we can breathe them. There are important issues affecting human and ecological health for both good ozone and bad ozone. For good ozone, the most important issues are the reduction of ozone globally, the Antarctic Ozone Hole, and Arctic ozone loss that is caused by chlorofluorocarbons. Chlorofluorcarbons were widely used in various applications like refrigerants, aerosol propellants, and foam blowing agents, but are now largely phased out due to their significant contribution to ozone depletion in the Earth's atmosphere. Reduced ozone means more solar UV gets to the ground causing more skin cancer. For bad ozone, the most important issues include the production of too much ozone in cities and nearby regions that is caused by too many pollutants from traffic, industrial processes, power generation, and other human activities. Increased ozone means more people have respiratory and heart problems. It is important to note that in October 2016 (2022 for the US), more than 140 countries signed the Kigali Agreement to reduce the use of chlorofluorocarbons and develop greener alternatives over time.

Smog and Soot

Smog is the bad ozone written about above. **Soot** is made from tiny particles of chemicals, soil, smoke, dust, or allergens carried in the air. They are the two most prevalent types of air pollution. The sources of smog and soot are similar as they occur with emissions from cars and trucks, factories, power plants, incinerators, and engines generated by combusting fossil fuels.

Smog irritates eyes and throats and damages the lungs. Children, elderly, and people who work or exercise outdoors are at higher risk. It's even worse for people who have asthma or allergies; these extra pollutants can intensify their symptoms and trigger asthma attacks. The tiniest airborne particles in soot are especially dangerous because they can penetrate the lungs and bloodstream and worsen bronchitis, lead to heart attacks, and even hasten death.

Because some sources of these pollutants have been preferentially located in or next to low-income neighborhoods, the negative effects of this pollution are experienced at a higher rate by the people who live in these communities. This has become an

important environmental justice issue.

Hazardous Pollutants

Several pollutants, even in tiny amounts, pose sometimes fatal health risks. Common pollutants include mercury, lead, dioxins, and benzene which are emitted as fossil fuels combust or are found in gasoline. Large amounts of lead can damage children's brains and kidneys, and small amounts can affect children's ability to learn. This is why pediatricians regularly test children for lead exposure. Mercury damages the central nervous system. Dioxins are usually found in food but can also be found in the air and is a carcinogen that affects the liver, immune system, nervous system, endocrine system, and reproductive functions. Benzene is also a carcinogen and causes eye, skin, and lung irritation, as well as blood disorders.

Another category of pollutants, polycyclic aromatic hydrocarbons (PAHs), are by-products of smoke generated by wildfires and exhaust by traffic. PAHs have been correlated to eye and lung irritation, blood and liver issues, and cancer.

Air Pollution in the United States

According to the American Lung Association (ALA), more than 33% of the people in the U.S. live in counties with unhealthy air. The Clean Air Act, enacted in 1970, has been critical in reducing air pollution by reducing harmful emissions from transportation, power plants, and manufacturing. Fossil fuel interests, often supported by industry-friendly lawmakers, have continuously sought to weaken its protections but maintaining the integrity of this law is crucial to improving and safeguarding air quality.

The latest findings reveal that climate change-driven wildfires and extreme heat are exacerbating the challenges to public health. Recent studies by the ALA focus on ozone levels, year-round particle pollution, and short-term particle pollution, revealing that people of color are 61% more likely than white people to live in counties with a failing grade in at least one of these categories, and three times more likely to live in areas where all three pollution types fail to meet health standards.

In the rankings for these pollution categories, central California cities (Bakersfield, Visalia, Fresno) occupy the top three spots for the worst air quality, despite the state's progress in reducing emissions over the past five decades. On the other end of the spectrum, cities like Burlington, Vermont; Honolulu; and Wilmington, North Carolina, consistently rank among the best for air quality in the country.

The most effective strategy to reduce air pollution is promoting the transition to cleaner fuels and industrial practices. By shifting to renewable energy sources like wind and solar, increasing fuel efficiency in vehicles, and replacing gasoline-powered cars with electric vehicles (assuming the electricity is not generated by fossil fuels), we can address pollution at its source while also mitigating the global warming that exacerbates its harmful health effects.

As for the economic costs of controlling air pollution, a report commissioned by the Natural Resources Defense Council (NRDC) shows that the benefits of cleaner air far outweigh the costs. Each year, the Clean Air Act's regulations result in benefits that are up to 32 times greater than the cost of enforcement. These benefits include preventing up to 370,000 premature deaths, reducing hospital admissions for respiratory and cardiac diseases by 189,000, and generating up to $3.8 trillion in net economic gains for the U.S. economy annually.

This page titled 7.6: Air Pollution is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by William Brune (John A. Dutton: e-Education Institute) via source content that was edited to the style and standards of the LibreTexts platform.

8: ENERGY & HUMIDITY

8.1: Energy Transfer
8.2: Latent Heat
8.3: Saturation
8.4: Humidity
8.5: Why Do We Care So Much About Moisture?
8.6: Atmospheric Stability and Lapse Rates

Chapter 8: Energy & Humidity — Introduction

The atmosphere is constantly exchanging energy. Sunlight warms Earth's surface, the surface transfers energy to the air above it, and moving air carries energy from one place to another. Water is an especially important part of these exchanges because it can exist in the atmosphere as a solid, liquid, or gas. As water changes from one phase to another, energy is absorbed or released, connecting the movement of water through the atmosphere directly to the movement of energy.

In this chapter, we will begin by examining the ways energy is transferred and the role of **latent heat** during changes of phase. We will then turn our attention to water vapor in the atmosphere. The amount of water vapor that can remain in the gas phase depends strongly on temperature, leading to the important concept of **saturation**. We will also examine several ways atmospheric moisture can be described, including relative humidity and dew point, and consider what these measurements tell us about the state of the atmosphere.

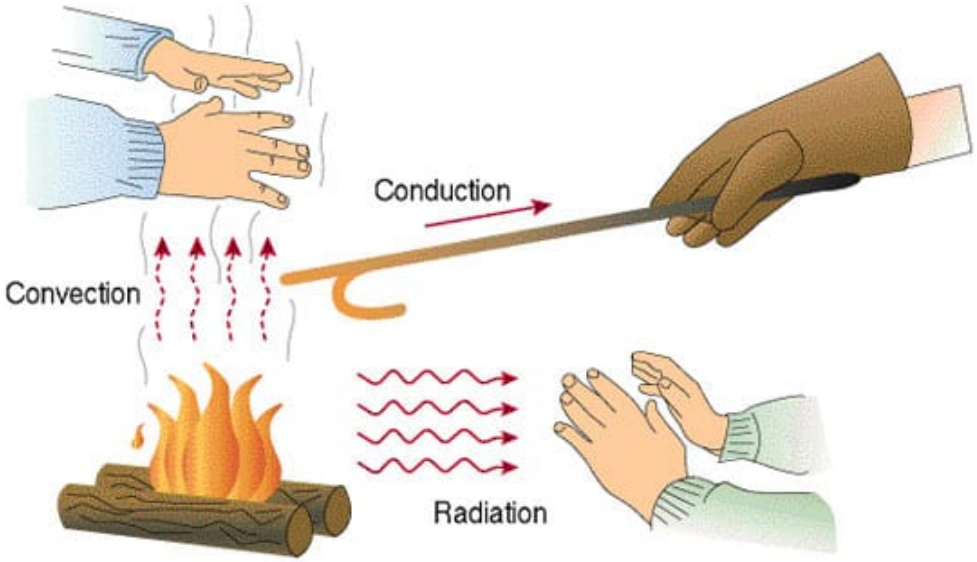
Finally, we will bring energy and moisture together to investigate **atmospheric stability**. Whether an air parcel rises, sinks, or remains at the same level depends on its temperature relative to the surrounding atmosphere, and condensation can change that behavior through the release of latent heat. These relationships between energy, moisture, and vertical motion provide the physical foundation for many of the weather processes we will encounter in the chapters ahead, including cloud formation, precipitation, thunderstorms, and larger weather systems.

This page titled 8: Energy & Humidity is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Robert Wagner.

8.1: Energy Transfer

One important example of kinetic energy is **thermal energy**, which comes from the tiny movement of many molecules in a system. In Chapter 1, we discussed how *temperature* is a measure of the average speed of atoms and molecules in a system. Here we can further describe temperature as being proportional to the average kinetic energy of the random motions of the molecules in a system. The faster the molecules move, the higher the temperature.

The transfer of thermal energy due to the temperature difference between two objects is what is known as **heat**. Heat is a form of energy in transit, and once transferred, it is stored as internal energy. There are four main methods of heat transfer: *conduction*, *convection*, *radiation*, and the absorption or release of *latent heat*.

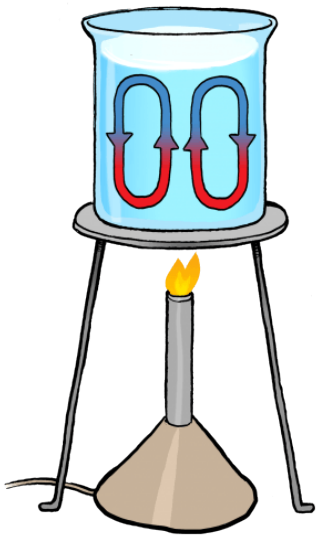


Conduction, convection, and radiation are all methods of heat transfer (CC BY-SA 4.0).

When you directly touch a hot object, such as the stovetop, the energy from the hot stove top is immediately transferred to your cool hand due to a difference in the speed of the molecules, causing you to feel a burn. This is an example of **conduction**: energy directly transferred through a substance without the movement of material. Certain materials are better conductors of heat than others. Metal, for example, conducts heat very efficiently, while air, which acts as an insulator, is a very poor conductor of heat.

Another type of energy transfer is convection, which occurs in liquids and gases (both fluids) because molecules can move freely and currents can naturally occur in them. This happens constantly in the atmosphere. **Convection** refers to movement within a fluid due to the tendency of lower density fluid to rise over higher density fluid, which sinks due to the force of gravity resulting in heat transfer within the fluid.

In the figure below, a beaker is being heated from the bottom by a flame. The arrows show upward movement in the center where the fluid is heated and therefore is less dense and buoyant. The cooler fluid at the top of the beaker is more dense and sinks toward the bottom under the influence of gravity.



Arrows show the motion of convection within a fluid that is heated from the bottom (CC BY 2.0).

Finally, when it comes to the atmosphere, the most important form of energy is the energy we get from the Sun, which is called *radiant energy* or radiation. **Radiation** is another type of heat transfer that was covered in Chapter 2. Remember that radiant energy can be transferred to an object without the space in between being heated. This is the result of electromagnetic waves from the Sun, which are then absorbed by the Earth's surface or atmosphere and converted to thermal energy. Electromagnetic waves do not need any matter to travel through and propagate at the speed of light in a vacuum which is about 300,000 $km \cdot s^{-1}$.

Heat, or the transfer of thermal energy due to the difference in temperature between two objects, is defined as Δq and is given in standard units of joules per kilogram ($J \cdot kg^{-1}$). One joule (J) is the standard unit of energy or work ($kg \cdot m^{2} \cdot s^{-2}$).

This page titled 8.1: Energy Transfer is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by LibreTexts.

8.2: Latent Heat

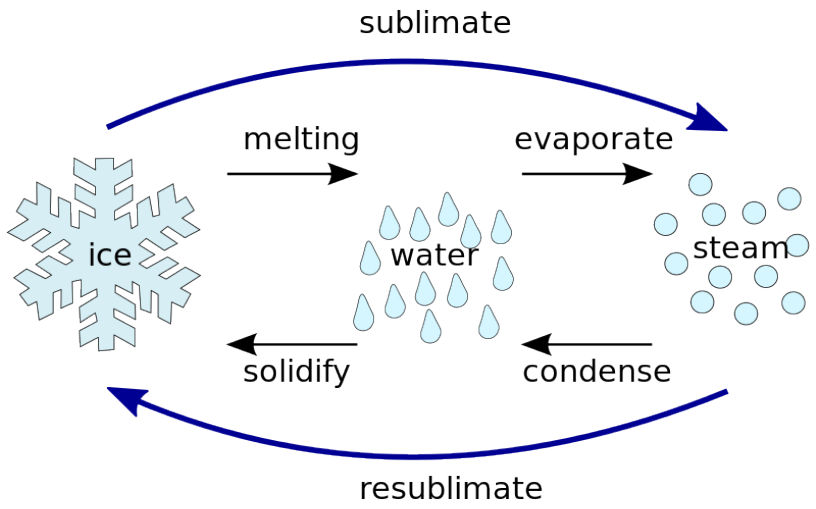
Latent heat refers to the amount of heat that is added or removed from a system when a change of phase takes place. In atmospheric science, this almost always refers to the phase changes of water (H_2O).

When heat is added to an ice cube, the temperature of the ice will increase until it reaches its melting point (0°C). After this point, any further addition of heat will cause the ice to melt to liquid water, but the temperature of the ice-water system will not change until the phase change is complete. The heat that is absorbed by the ice-water system from the environment in order for the phase change to occur is known as latent heat.

You have experienced this as your ice melts in your water glass. It stays cold until all of the ice is gone and then rapidly warms. The absorption of latent heat is also the reason you feel cold right after leaving your shower or a body of water. The excess water on your skin starts to evaporate, but it requires additional energy in order to transition from liquid to vapor. Energy is absorbed from the environment (your skin) to evaporate the remaining liquid, causing your skin to cool. This is known as evaporative cooling.

Processes that take heat from the environment (melting, evaporation, sublimation) are considered cooling processes. On the other hand, processes that add heat to the environment (condensation, freezing, deposition) are warming processes because during the phase change energy is released. This is an important distinction, so be sure this is clear before moving on. I'll repeat it below for emphasis.

> **Pro Tip:** Melting, evaporation, and sublimation are considered cooling processes because they take heat from the environment. Condensation, freezing, and deposition are considered warming processes because during the phase change energy is released and they add heat to the environment.



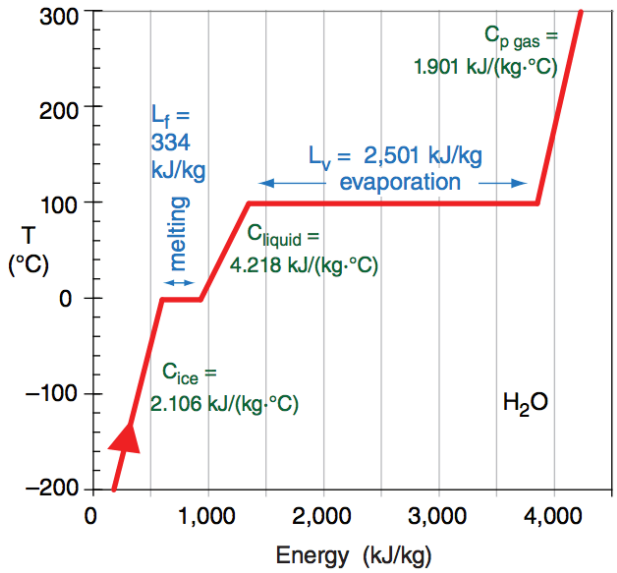
Phase changes of water (CC BY-SA 3.0). The lower energy state is on the left, and the higher energy state is on the right. Going from left to right takes energy (cools the environment), going from right to left releases energy (warms the environment).

The reason phase changes of water are so important in the atmosphere is because it acts as the energy to fuel convection and thunderstorms. When water vapor condenses to form the liquid water found in clouds, this process releases enormous amounts of latent heat to the environment.

Latent heat (Q_E) is the latent energy that is possessed by an object of mass m, but usually we want to know the change of latent heat that is caused by a phase change process.

$$\Delta Q_E = L * \Delta m_{\text{water}}$$

The change in latent heat is equal to the latent heat factor multiplied by the mass of water undergoing a phase change.



The y-axis of this figure shows the state of water at different temperatures. The x-axis shows the amount of energy applied to the water. When water reaches 0°C and 100°C, the melting/freezing and boiling points of water, respectively, the temperature stops increasing while energy is still being applied. Instead of increasing the water's temperature, this energy is instead going into the processes of melting and evaporation. From here, you can see that the latent heat of vaporization is much greater than the latent heat of melting, because much more energy is applied while the temperature remains the same (CC BY-NC-SA 4.0).

Different phases of water have different latent heat factors (L), meaning the change in phase of a mass of water corresponding to the change in latent heat is different depending on what process is taking place. This is well displayed in the image above by the horizontally flat regions where energy is going into the system but no temperature change is occurring.

For evaporation or condensation (phases changes between liquid water and vapor), the latent heat of vaporization, L_v = 2.5 * 10^6 J·kg^{-1} is used. For melting or freezing (phase changes between ice and liquid water), the latent heat of fusion or freezing, L_f = 3.34 * 10^5 J·kg^{-1} is used.

This page titled 8.2: Latent Heat is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by LibreTexts.

8.3: Saturation

Imagine a closed jar filled halfway with water. At the initial time, more water molecules evaporate from the water surface than the number that return. However, after some time, the number of molecules evaporating from the surface will be equal to the number of molecules condensing back into the water surface. When condensation and evaporation are equal, this is called **saturation**.

Saturation occurs when air contains the maximum amount of water vapor possible for its given temperature. That is why condensation equals evaporation. If evaporation occurs, the air cannot contain more water vapor, so some must condense. Now let's get quantitative.

Vapor pressure at saturation

Every gas in the atmosphere exerts pressure, for example, vapor pressure makes up a fraction of the total atmospheric pressure. In the following equation, all of the gases in Earth's atmosphere contribute to the total atmospheric pressure $P_{\text{atmosphere}}$.

$$P_{\text{atmosphere}} = P_{N_2} + P_{O_2} + P_{Ar} + \ldots..P_{H_2O}$$

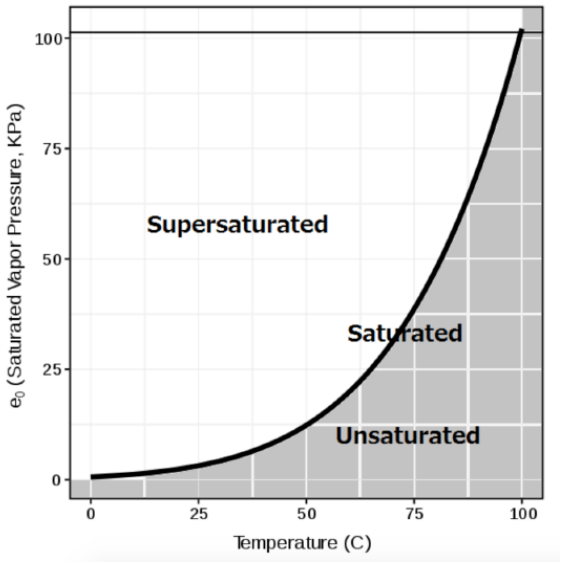
Specifically for water vapor, the more water vapor that is added to the atmosphere, the higher the vapor pressure P_{H_2O}. The units for vapor pressure are the same as pressure and can be in Pascals, hectoPascals, or kiloPascals. Because we are staying consistent with Roland Stull's *Practical Meteorology* textbook, we will use kiloPascals (kPa) throughout this chapter.

The amount of water vapor that the atmosphere can contain depends on temperature. Lower temperature air cannot contain as much water vapor as higher temperature air. If we think of this quantitatively in terms of pressure, *saturation vapor pressure* refers to the pressure exerted by the movement of water vapor molecules exerted over a surface of liquid water. When the partial pressure exerted by water vapor is equal to the saturation vapor pressure, the air is saturated.

The **Clausius-Clapeyron equation** gives the approximate relationship between saturation vapor pressure (e_s) and temperature in the atmosphere

$$e_s \approx e_0 \cdot \exp\left[\frac{L_v}{R_v} \cdot \left(\frac{1}{T_0} - \frac{1}{T}\right)\right]$$

where the water-vapor gas constant $\mathfrak{R}_v$ is 461 J·K^{-1}·kg^{-1}, T_0 is 273.15 K, e_0 is 0.6113 kPa, and L_v is the latent heat of vaporization, 2.5×10^6 J·kg^{-1}. This results in $L_v/\mathfrak{R}_v$ being equal to 5423 K. In this equation, units for temperature must be in Kelvin. Note that in the equation above, $\exp[x]$ implies the exponential function e^x, but it is written on one line for visual purposes.



A graph of the saturation vapor pressure as a function of temperature showing the exponential relationship between the two from the Clausius-Clapeyron equation (Modified from CC BY-SA 4.0).

The image shows the relationship between temperature and saturation vapor pressure based on the Clausius-Clapeyron equation. Lower temperatures are saturated with respect to water vapor at lower vapor pressures, while higher temperatures need higher vapor pressures to be saturated. Temperature is the primary factor determining water vapor saturation.

> In the graph of saturation vapor pressure vs. temperature notice the saturation vapor pressure value at the boiling temperature, 100 °C. The saturation vapor pressure value e_s(100 °C)=101.325 kPa, is the same value as the atmospheric surface pressure. Water boils at the Earth's surface when the saturation vapor pressure is equal to the atmospheric pressure, which is why water boils at 100 °C. Will water boil at the same temperature at the top of Mount Everest?

This page titled 8.3: Saturation is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by LibreTexts.

8.4: Humidity

Humidity is a measure of the water vapor content of the air. The amount of water vapor in the air depends on the controls over evaporation discussed earlier. There are several ways in which a meteorologist can express the humidity of the air. Each humidity measure is controlled to some degree by air temperature.

Absolute humidity is the weight of water vapor per unit volume of air, usually measured in units of grams of water vapor per cubic meter of air. Absolute humidity is not often used to express the moisture content of air because it is sensitive to changes in both the temperature of the air and atmospheric pressure. For instance, let's say that a 1 cubic meter parcel of air at the surface has 2 grams of water in it. Now lift the parcel of air upwards into the atmosphere. As the air rises upward the decrease in atmospheric pressure on the parcel allows it to expand outward occupying more space. Let's say that the parcel doubles in size as a result of uplift. Before rising, the absolute humidity was 2 gm/m^3. As the air doubles in volume the new absolute humidity is 1 gm/m^3. In actuality the parcel still has the same weight of water in it, 2 grams. But given the way absolute humidity is calculated it appears the amount of water in the air has decreased.

Instead of absolute humidity, we use a measure that is not sensitive to volume changes in the air. ***Specific humidity*** is measured as the weight of water vapor in the air per unit weight of air, which includes the weight of water vapor. The units of measurement are grams of water vapor per kilogram of air. Given that weight is not significantly influenced by temperature or atmospheric pressure, specific humidity is much more useful as a measure of humidity. Another measure very similar to specific humidity is the mixing ratio. The ***mixing ratio*** is the weight of water vapor per unit weight of dry air. Because the atmosphere is made up of so little moisture by volume, the mixing ratio is virtually the same as the specific humidity.

Humidity is not only measured as a weight, but also by the pressure it creates. ***Vapor pressure*** is the partial pressure created by water vapor. Vapor pressure, like atmospheric pressure, is measured in millibars and is relatively insensitive to volumetric expansion or temperature. The ***saturation vapor pressure*** is simply the pressure that water vapor creates when the air is fully saturated.

When we think of air as being saturated with moisture we often say that the air is "holding all the moisture it can". This implies that once the air has reached saturation it won't "accept" anymore water by evaporation. This is wrong. So long as there is water available evaporation will continue even when the air is fully saturated. Let's examine the concept of saturation in more detail.

Imagine a beaker filled halfway with water. Let's put a top on it to constrain the movement of water molecules and eliminate the influence of wind on evaporation. As the water absorbs heat it begins to change phase and enter the air as water vapor. Above the surface, water vapor molecules dart about suspended in the air. However, near the surface water molecules are attaching themselves back the surface, thus changing back into liquid water (condensation) (A). As evaporation occurs the water level in the beaker decreases (B). This occurs because evaporation exceeds condensation of water back onto the surface. After some time, the amount of water entering the air from evaporation is equal to that condensing (C). When this occurs the air is said to be saturated.

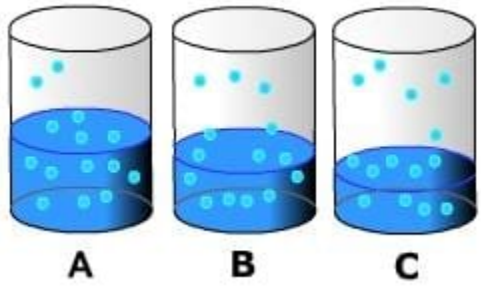


Figure 8.4.1: Evaporation and condensation in an enclosed beaker of water.

The saturation level of the air is directly related to the air's temperature. As air temperature increases, more water can remain in a gas phase. As temperature decreases, water molecules slow down and there is a greater chance for them to condense on to surfaces. The graph below shows the relationship between air temperature and vapor pressure, a measure of the humidity, at

saturation.

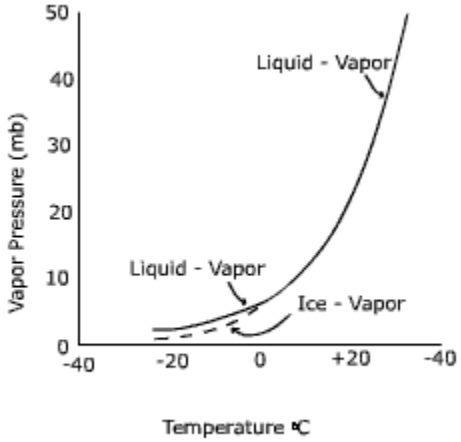


Figure 8.4.2: Relationship Between Air Temperature and Vapor Pressure

Note that below zero degrees Celsius the curve splits, one for the saturation point above a liquid surface (liquid-vapor) and one for a surface of ice (ice - vapor). The first thing you might be wondering is how water can exist as a liquid below the freezing point. Water that is not frozen below 0° C is called "super-cooled water". For water to freeze, the molecules must become properly aligned to attach to one another. This is less likely to occur especially with small amounts of water, like cloud droplets. Thus in clouds where temperatures are below freezing it is common to find both super-cooled liquid water and ice crystals.

Notice that the saturation vapor pressure at -20° C is lower for ice than for a liquid surface. Why would this be so? You may recall that to convert water from a liquid to a gas requires about 600 calories per gram. To convert water from a solid to a gas requires about 680 calories, hence it is more difficult to "liberate" a molecule of water from ice than water. Therefore, when the air is saturated, there are more molecules above a water surface (i.e. more vapor pressure) than an ice surface (i.e. less vapor pressure).

Dew point temperature is the temperature at which condensation takes place and is used as a measure of moisture content. The dew point temperature depends on the amount of moisture in the air, the more moisture in the air, the higher the dew point temperature. It gets its name "dew point" because dew will form on surfaces when the air reaches saturation.

Have you ever noticed that even though it's 100% relative humidity out, it feels a lot drier during the winter than the summer? To see why, we have to examine relative humidity. ***Relative humidity*** is the ratio of the amount of water vapor in the air to its saturation point. Often relative humidity is defined as the amount of water vapor in the air to "how much it can hold" at a given temperature. The notion of a holding capacity is dispelled when one considers what saturation really means. Regardless, what we do know is that the saturation level of the air with respect to water vapor depends on the air's temperature. We know that as air temperature increases, the ability for the air to keep water in its vapor state is easier. That is, as the air temperature increases it can keep more water in the vapor state. So why does saturated cold air feel drier than warm air at saturation? Let's look at an example.

Continental polar air (cP) has an average temperature of 5° C (41° F). Its saturation mixing ratio is 6 g/kg. So continental air at 100% relative humidity is

RH = 6 g/kg (in the air) / 6 g/kg (saturation)

The average temperature of maritime tropical air (mT) is 22o C (71.6o F) with a saturation mixing ratio of 16 g/kg. So maritime tropical air at 100% relative humidity is

RH = 16 g/kg (in the air) / 16 g/kg (saturation)

Therefore, polar air is drier at 100% relative humidity because it has much less moisture in it at saturation than the warmer maritime tropical air.

This page titled 8.4: Humidity is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Michael E. Ritter (The Physical Environment) via source content that was edited to the style and standards of the LibreTexts platform.

- 7.2.3: Humidity by Michael E. Ritter is licensed CC BY-SA 4.0. Original source: https://www.thephysicalenvironment.com.

8.5: Why Do We Care So Much About Moisture?

You may be wondering at this point why we care so much about moisture and why we need so many definitions of (almost) the same thing. The reason is that moisture is an extremely important atmospheric property. Water can exist in three phases (vapor, liquid, ice) within the atmosphere at typical pressures and temperatures. It has an especially large impact on the human experience—think about a humid day, foggy conditions, rain, snow, or even hail! Less obvious is its impact on atmospheric stability, which drives the aforementioned conditions.

For now, let's think about the process of water vapor condensing to form liquid water. There is one final definition of humidity that will be helpful.

Lifting Condensation Level

The lifting condensation level, z_{LCL}, is the altitude where clouds form. At the LCL, temperature equals the dew point temperature, resulting in saturation and therefore condensation. The height (z) of the LCL is

$$z_{LCL} = a \cdot (T - T_d)$$

where a is 0.125 km °C^{-1}. We can also define the temperature at the LCL as follows.

$$T_{LCL} = T - \Gamma_d \cdot z_{LCL}$$

Moist Adiabatic Lapse Rate

In the last chapter, we discussed how temperature changes as a dry parcel of air is lifted in the atmosphere. You will recall that as an air parcel is lifted, the temperature drops by 9.8 K every km due to the work the air parcel must do to the environment as it expands. Let's add moisture to the discussion and see how this changes things.

If the air parcel reaches saturation (100% relative humidity) and water vapor condenses to liquid water within the parcel, latent heat will be released. In the case of a rising air parcel that is cooling from adiabatic expansion, this added heat from condensation counterbalances some of the cooling. Hence, the air parcel will no longer cool at the dry adiabatic lapse rate but at the smaller moist adiabatic lapse rate (Γ_m). Unlike the dry adiabatic lapse, the moist adiabatic lapse rate is not constant and varies based on the temperature and moisture of the air parcel.

We will approximate the moist adiabatic lapse rate with the following value.

$$\Gamma_m = 4.5 \frac{K}{km} = 4.5 \frac{^\circ C}{km}$$

The difference between the dry adiabatic lapse rate (Γ_m) and the moist adiabatic lapse rate (Γ_m) is significant and has profound influence on atmospheric stability, the topic of the following chapter.

Chapter 4: Questions to Consider

1. Explain the conditions needed for saturation to occur.
2. What is the saturation vapor pressure of air at 26°C?
3. Explain the difference between specific humidity and relative humidity.
4. If the temperature is 10°C and the pressure is 700 hPa, calculate the saturation specific humidity and the saturation mixing ratio.
5. Explain why the moist adiabatic lapse rate is less than the dry adiabatic lapse rate.

Selected Practice Question Answers:

Query 8.5.1

This interactive element is not available in this format. Visit the link or scan the QR code to access it online.

View this Interactive Content online

This page titled 8.5: Why Do We Care So Much About Moisture? is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by LibreTexts.

8.6: Atmospheric Stability and Lapse Rates

Adiabatic Processes

When discussing stability in atmospheric sciences, we typically think about air parcels, or imaginary blobs of air that can expand and contract freely, but do not mix with the air around them or break apart. The key piece of information is that movement of air parcels in the atmosphere can be estimated as an adiabatic process. **Adiabatic processes** do not exchange heat and they are reversible.

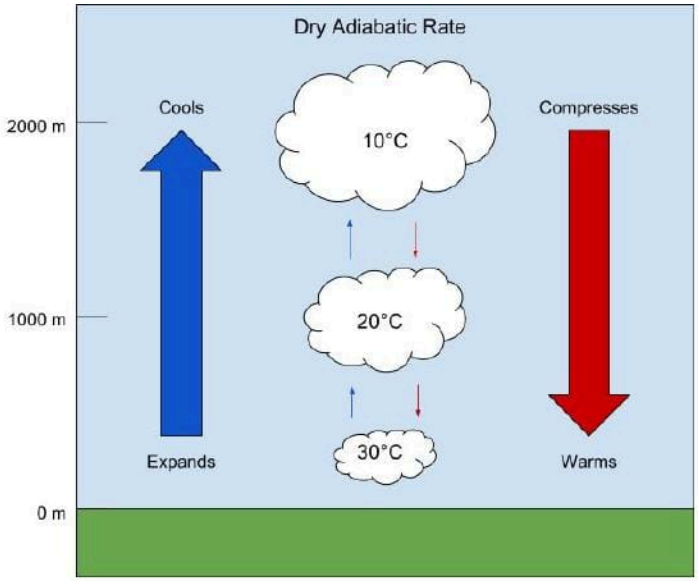
Imagine you have a parcel of air at the Earth's surface. The air parcel has the same temperature and pressure as the surrounding air, which we will call the environment. If you were to lift the air parcel, it would find itself in a place where the surrounding environmental air pressure is lower, because we know that pressure decreases with height. Because the environmental air pressure outside the parcel is lower than the pressure inside the parcel, the air molecules inside the parcel will effectively push outward on the walls of the parcel and expand adiabatically. The air molecules inside the parcel must use some of their own energy in order to expand the air parcel's walls, so the temperature inside the parcel decreases as the internal energy decreases. To summarize, rising air parcels expand and cool adiabatically without exchanging heat with the environment.

Now imagine that you move the same air parcel back to Earth's surface. The air parcel is moving into an environment with higher air pressure. The higher environmental pressure will push inward on the parcel walls, causing them to compress, and raise the inside temperature.

The process is adiabatic, so again, no heat is exchanged with the environment. However, temperature changes in the air parcel can still occur, but it is not due to mixing, it is due to changes in the internal energy of the air parcel.

Dry Adiabatic Lapse Rate

As long as an air parcel is unsaturated (relative humidity < 100%), the rate at which its temperature will change will be constant. A decrease in temperature with height is called a **lapse rate** and while the temperature decreases with altitude, it is defined as positive because it is a *lapse* rate. Recall from chapter 3 that the **dry adiabatic lapse rate**, Γ_d, **is equal to 9.8 K·km^{-1} = 9.8 °C·km^{-1}**. This drop in temperature is due to adiabatic expansion and a decrease in internal energy.



Air rises, expands, and cools at the dry adiabatic lapse rate, approximated as a 10°C decrease per km (created by Britt Seifert).

Let's get back to the topic of atmospheric stability. Stability in the atmosphere refers to a condition of equilibrium. As discussed with the example of the boulder on a hill or valley, some initial movement resulted in either more (unstable), less (stable), or no

change (neutral). Given some initial change in the elevation of an air parcel, if the air is in stable equilibrium, the parcel will tend to return back to its original position after it is forced to rise or sink. In an unstable equilibrium, an air parcel will accelerate away from its initial position after being pushed. The motion could be upward or downward, but generally unstable atmospheres favors vertical motions. Finally, in a neutral equilibrium, some initial change in the elevation of an air parcel will not result in any additional movement.

Determining Stability

How do you know if an air parcel will be stable after some initial displacement? Stability is determined by comparing the temperature of a rising or sinking air parcel to the environmental air temperature. Imagine the following: at some initial time, an air parcel has the same temperature and pressure as its environment. If you lift the air parcel some distance, its temperature drops by 9.8 K·km^{-1}, which is the dry adiabatic lapse rate. If the air parcel is colder than the environment in its new position, it will have higher density and tend to sink back to its original position. In this case, the air is stable because vertical motion is resisted. If the rising air is warmer and less dense than the surrounding air, it will continue to rise until it reaches some new equilibrium where its temperature matches the environmental temperature. In this case, because an initial change is amplified, the air parcel is unstable. In order to figure out if the air parcel is unstable or not we must know the temperature of both the rising air and the environment at different altitudes.

One way this is done in practice is with a weather balloon. We can get a vertical profile of the environmental lapse rate by releasing a radiosonde attached to a weather balloon. A radiosonde sends back data on temperature, humidity, wind, and position, which are plotted on a thermodynamic diagram. This vertical plot of temperature and other variables is known as a **sounding**.

Dry Stability

If an air parcel is dry, meaning unsaturated, stability is relatively straightforward. An atmosphere where the environmental lapse rate is the same as the dry adiabatic lapse rate, meaning that the temperature in the environment also drops by 9.8 K·km^{-1}, will be considered neutrally stable. After some initial vertical displacement, the temperature of the air parcel will always be the same as the environment so no further change in position is expected.

If the environmental lapse rate is less than the dry adiabatic lapse rate, some initial vertical displacement of the air parcel will result in the air parcel either being colder than the environment (if lifted), or warmer than the environment (if pushed downward). This is because if lifted, the temperature of the air parcel would drop more than the temperature of the environment. This is a stable situation for a dry air parcel and a typical scenario in the atmosphere. The **global average tropospheric lapse rate is 6.5 K·km^{-1}**, which is stable for dry lifting.

Finally, if the environmental lapse rate is greater than the dry adiabatic lapse rate, some initial vertical displacement of the air parcel will result in the air parcel either being warmer than the environment (if lifted), or colder than the environment (if pushed downward). This is because if lifted, the temperature of the air parcel would drop less than the temperature of the environment. This is an unstable situation for a dry air parcel.

In general for a dry air parcel, the following is true.

$$\Gamma_d = \Gamma_{env} \quad \text{NEUTRAL}$$

$$\Gamma_d < \Gamma_{env} \quad STABLE$$

$$\Gamma_d > \Gamma_{env} \quad UNSTABLE$$

Moist Adiabatic Lapse Rate

When moisture is added, everything gets more complicated. In Chapter 4 we learned that whether or not an air parcel is saturated depends primarily on its temperature and, of course, its moisture content. The graph of the Clausius-Clapeyron relationship shows us that given the same amount of moisture, air is more likely to be saturated at a lower temperature.

We know that as an air parcel is lifted, its temperature drops according to the dry adiabatic lapse rate. So what happens when the air parcel is cold enough that the air becomes saturated with respect to water vapor? The short answer is that if it continues to

cool, water vapor will condense to liquid water to form a cloud.

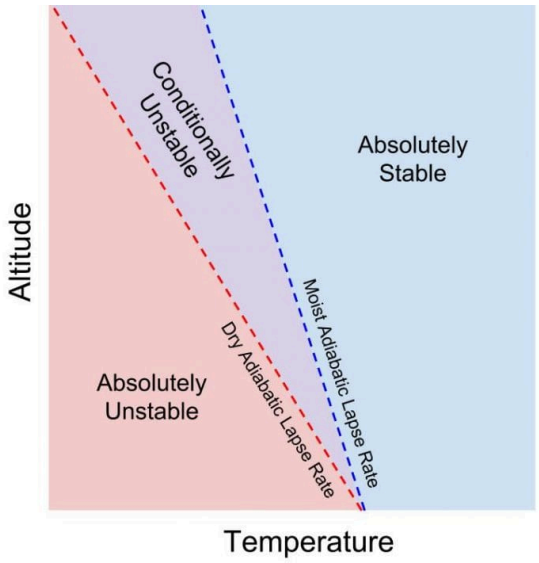
When water vapor condenses, it goes from a higher energy state to a lower energy state. Energy is never created nor destroyed, especially in phase changes, so what happens to all that excess energy? The energy gets released in the form of latent heat. The latent heat of condensation is approximately equal to 2.5 * 10^6 $J{\cdot}kg^{-1}$, which means that for every kg of water vapor that condenses to form liquid water, 2.5 *10^6 Joules of energy are released.

This has large consequences for the lapse rate of an air parcel and distinguishes the dry adiabatic lapse rate from the moist adiabatic lapse rate. As latent heat is added from the process of condensation, it offsets some of the adiabatic cooling from expansion. Because of this, the air parcel will no longer cool at the dry adiabatic lapse rate, but will cool as a slower rate, known as the **moist adiabatic lapse rate**. To summarize, a parcel will cool at the dry adiabatic rate until it is saturated, after which it won't cool as quickly due to condensation. The moist adiabatic lapse rate varies a little by temperature, but in this class we will consider it a constant for simplicity:

$$\Gamma_m = 4.5K \cdot km^{-1} = 4.5^{\circ}C \cdot km^{-1}$$

Moist Stability

The effects of moisture change the lapse rate of the air parcel and, therefore, affects stability. However, the concepts are still the same and we still compare the air parcel temperature to the environmental temperature. We have just one added complication to worry about—we need to know whether the air parcel is dry or moist. Some definitions are included below, which take into account both dry and moist adiabatic lapse rates.



A thermodynamic diagram showing the stability of the atmosphere based on the dry (Γ_d = 9.8 K km^{-1}) and moist (Γ_m = 4.5 K km^{-1}) adiabatic lapse rates (Created by Britt Seifert).

The atmosphere is said to be **absolutely stable** if the environmental lapse rate is less than the moist adiabatic lapse rate. This means that a rising air parcel will always cool at a faster rate than the environment, even after it reaches saturation. If an air parcel is cooler at all levels, then it will not be able to rise, even after it becomes saturated (when latent heating will counteract some cooling).

The atmosphere is said to be **absolutely unstable** if the environmental lapse rate is greater than the dry adiabatic lapse rate. This means that a rising air parcel will always cool at a slower rate than the environment, even when it is unsaturated. This means that it will be warmer (and less dense) than the environment, and allowed to rise.

The atmosphere is said to be **conditionally unstable** if the environmental lapse rate is between the moist and dry adiabatic lapse rates. This means that the buoyancy (the ability of an air parcel to rise) of an air parcel depends on whether or not it is saturated. In a conditionally unstable atmosphere, an air parcel will resist vertical motion when it is unsaturated, because it will cool faster than the environment at the dry adiabatic lapse rate. If it is forced to rise and is able to become saturated, however, it will cool at

the moist adiabatic lapse rate. In this case, it will cool slower than the environment, become warmer than the environment, and will rise.

Hawaiian Focus Box

Around Hawaii, the atmosphere is almost always conditionally unstable, meaning that the environmental lapse rate lies somewhere between the dry and moist adiabatic lapse rates. For this reason, Hawaii almost always has convective clouds. Convective clouds are clouds where the edges are bumpy and cumuliform, like cauliflower. The clouds are convective because the atmosphere is stable to dry lifting and unstable to moist lifting. Once the air is saturated, instability sets in and vertical motion takes off. This is especially common as air is lifted over our mountainous islands. The forced lifting from the terrain creates clouds and rain right over the mountains! In scientific terms, the initial lifting of the stable low level dry air by the terrain causes the air to adiabatically expand and reach saturation, at which point the environment is unstable to moist lifting and convection is the result.

This page titled 8.6: Atmospheric Stability and Lapse Rates is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by LibreTexts.

9: CLOUDS AND PRECIPITATION

9.1: Cloud Condensation Nuclei
9.2: Cloud Naming Conventions
9.3: Cloud Droplets
9.4: Precipitation Types

Chapter 9: Clouds and Precipitation — Introduction

Clouds are among the most visible signs of activity in Earth's atmosphere. Their formation brings together many of the ideas introduced in the previous chapter. Water evaporates from Earth's surface and enters the atmosphere as water vapor. As air rises and cools, it may approach saturation, creating conditions in which water vapor can condense. But reaching saturation is only the beginning of the story. To understand how clouds and precipitation form, we need to examine what happens to water once condensation begins.

Cloud formation depends on tiny particles in the atmosphere known as **cloud condensation nuclei**. Water can condense onto these particles to form the microscopic droplets that make up a cloud. Depending on atmospheric temperature and other conditions, clouds may contain liquid water droplets, ice crystals, or a mixture of both. The altitude, atmospheric stability, and movement of air also help determine the shapes and types of clouds that develop. Learning to recognize these cloud types can therefore provide clues about what is happening in the atmosphere and what weather may follow.

Cloud droplets and ice crystals are initially far too small to fall to the ground as precipitation. They must grow through interactions with other droplets and crystals before becoming large enough to overcome the upward motions of the surrounding air. Different processes dominate this growth in warm and cold clouds, eventually producing **rain, snow, sleet, or hail** depending on the conditions within the cloud and the atmosphere below it.

In this chapter, we will follow water from its first condensation onto microscopic particles through the development of clouds and, ultimately, precipitation. In doing so, we will connect the concepts of humidity, saturation, and atmospheric stability from Chapter 8 to the clouds and precipitation we observe in everyday weather.

This page titled 9: Clouds and Precipitation is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Robert Wagner.

9.1: Cloud Condensation Nuclei

When air becomes saturated, excess water vapor in the atmosphere can condense to form liquid water. However, water vapor requires a surface on which to condense. The surface is provided by tiny particulates in the atmosphere known as aerosols, or more generally, cloud condensation nuclei (**CCN**). Aerosols are everywhere in the atmosphere. They can be composed of earth matter like dust, clay, or soot. They can also be composed of sea salt from the ocean, black carbon from fires, or sulfate from volcanic activity. They can be composed of material from plant or organism life, like sulfates and nitrates or volatile organic compounds, or even phytoplankton. Anything that is small enough to be lofted and suspended in the atmosphere has the potential to be an aerosol. An aerosol that is hygroscopic, or "water-liking", has the potential to be a cloud condensation nuclei.

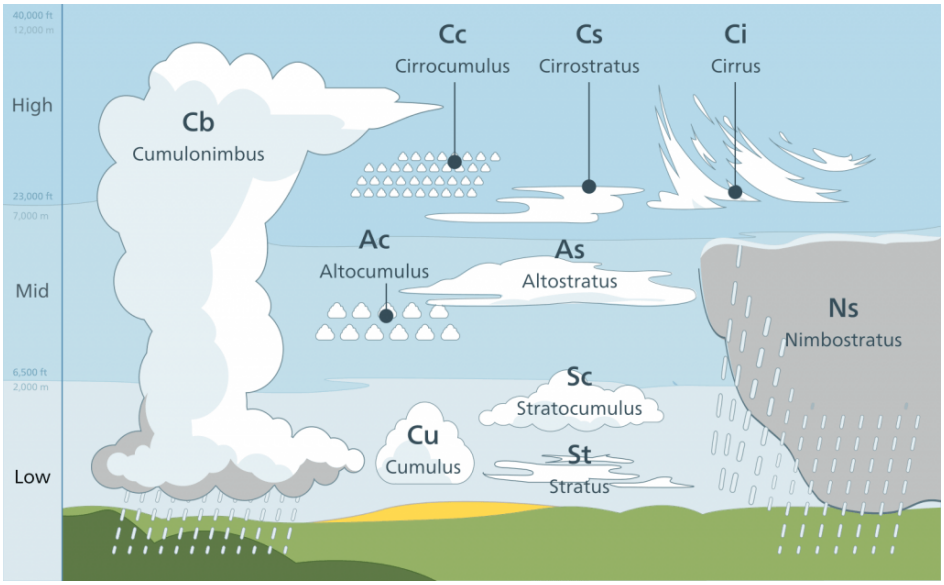
If there were no aerosols in the atmosphere, relative humidities well over 100% would be needed for clouds to form. In laboratory studies with a clean atmosphere (with no aerosols) relative humidities up to 400% have been measured. In Earth's atmosphere, aerosols are abundant, so water vapor does not struggle to find CCN to condense onto.

Another type of aerosol that helps with the formation of cold clouds, or clouds composed of ice crystals, are called Ice Nuclei (IN). Dust is a particularly good ice nuclei. Materials with a hexagonal shape, like ice crystals, also make particularly good surfaces for ice to form on. A common one used in cloud seeding is called silver iodide, AgI. We'll discuss this more in the following chapter on precipitation.

This page titled 9.1: Cloud Condensation Nuclei is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by LibreTexts.

9.2: Cloud Naming Conventions

Once a cloud forms, how do we know what to call it? Clouds form in many environments and look different depending on those environments, and depending on whether they're composed of liquid droplets or ice crystals. The following diagram gives a brief overview of the many cloud types, along with their common abbreviations.



This diagram shows the common cloud types and associated altitudes. The cloud naming convention involves dividing the atmosphere into Low, Middle, and High level clouds, and typically cumuliform and stratiform clouds (CC BY-SA 3.0).

Clouds can be identified based on their appearance, shape, and altitude. The two primary cloud categories are cumuliform and stratiform clouds.

By Appearance and Shape

Cumuliform clouds develop as a result of vertical motions by atmospheric instability. They are *convective* clouds meaning that they form in air parcels that are buoyant and are undergoing convection, which is the transfer of heat or mixing within a fluid due to warm air rising and cool air sinking. Examples of cumuliform clouds include cumulus, cumulus congestus, and cumulonimbus.

Stratiform clouds are horizontally layered clouds. They tend to spread into wide regions, and take on an appearance of a sheet or blanket. They typically form when a layer of air is brought to saturation but is thermodynamically stable, or when a convective cloud meets a stable layer and spreads out in a layered fashion. Examples of stratiform clouds include nimbostratus, stratus, altostratus, cirrostratus, and cirrus.

By Water Phase

Clouds composed of only liquid water droplets are called "warm clouds", and typically have clearly defined edges. Low altitude clouds are usually warm clouds.

Clouds made up of only ice crystals are called "cold clouds" and typically have fuzzy looking edges. The edges look fuzzy and not well defined because it takes longer to go from the ice phase to the vapor phase as compared to warm clouds. This transition time scale results in a larger cloud to dry air transition region. High altitude clouds are always cold clouds.

Clouds composed of liquid water and ice crystals are called "mixed phase clouds". It is difficult to distinguish by eye whether a cloud is mixed phase. Often the decision comes down to the height of the cloud, knowing that clouds that extend from the surface up high in the atmosphere likely have a mixture of both liquid droplets and ice crystals.

By Altitude

Clouds at a high altitude have the prefix “cirro” or “cirrus”. Due to the high altitude, the cirrus and cirrostratus are made out of ice crystals.

Clouds at mid-altitude have the prefix “alto”. They are usually made out of liquid droplets, but can be a mixture between liquid droplets and ice crystals.

Low level clouds don’t have a particular prefix.

By Characteristics

The prefix “nimbo” or suffix “nimbus” indicates a precipitating cloud. Nimbostratus usually have light to moderate precipitation, whereas cumulonimbus or thunderstorm clouds have heavy precipitation and sometimes even hail.

Other

The above clouds represent the primary cloud types, but many other cloud names and cloud types are used in atmospheric sciences. A few of the more common cloud names are given below.

Lenticular clouds are stationary lens-shaped clouds with a smooth appearance that usually forms over the summit of a mountain or on the lee wave crest. They are also referred to as lee-wave clouds or mountain-wave clouds.

Mammatus clouds are formed from downward convection, typically in the anvil portion of a cumulonimbus. They have a distinctive look that makes them easy to spot, especially at sunset. They are made up of hanging pouches that look a little bit like udders.

Fog

A cloud that forms at or near the ground is called fog. The main types of fog are upslope, radiation, advection, precipitation or frontal, and steam fog.

Upslope, radiation, and advection fogs develop *due to cooling*. Upslope fog develops as stable, moist air rises over topography like a hill or mountain. Radiation fog occurs over land as radiational cooling lowers the air temperature to its dew point temperature. Radiation fog usually occurs during calm, clear nights. Advection fog forms as warm, moist air flows over a colder surface and the air cools to its dew point temperature.

Frontal and steam fogs are formed by the *addition of water vapor*. Frontal fog develops when warm raindrops evaporate in a colder airmass. Steam fog or sea smoke forms as cold air moves over warmer water.

This page titled 9.2: Cloud Naming Conventions is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by LibreTexts.

9.3: Cloud Droplets

Recall from the previous chapter on clouds that **cloud condensation nuclei** (CCNs) are required for water vapor to condense onto. Many CCNs are hygroscopic, meaning they tend to absorb moisture, so condensation may start before the relative humidity reaches 100%. For example, when condensation occurs on salt particles, which are extremely hygroscopic, condensation can begin at 80% relative humidity or lower.

Imagine there are many CCNs of differing sizes in a body of humid, but unsaturated air. If the air were to be lifted by a mountain or in a rising thermal, it would cool, and the relative humidity would increase. As the air nears saturation, condensation will begin to occur on the largest and most hygroscopic CCN. At some later time, a cloud of many small cloud droplets, far too small to fall at any significant speed, will form.

The terminal velocity of a falling cloud droplet (with radius "r" less than 40 μm) is given by the following equation from Stoke's Drag Law:

$$w_{T,\text{ cloud}} = \left(-1.19 * 10^{8} m^{-1} s^{-1}\right) * r^{2}$$

where "r" needs to be expressed in meters. A simple calculation will show that it takes hours, if not days, for a cloud droplet to fall from even low altitudes. The friction provided by the air or even tiny upward air currents will keep cloud droplets suspended for long periods.

Terminal velocity implies that a steady state has been reached in the fall velocity with a balance between the downward gravitational force on the drops and the upward frictional drag of air.

Raindrops

For raindrops, a different equation is used to approximate the fall speed. For spherical raindrops,

$$w_{T,\text{ rain}} = \left(-220 m^{1/2} s^{-1}\right) * \left(\frac{\rho_0}{\rho_{\text{air}}} * r\right)^{1/2}$$

where ρ_o is a reference value of density, typically 1.2 kg m^{-3} and ρ_{air} is the density of air where the raindrop exists. Again, "r" is the radius of the drop, given in meters. High up in the atmosphere when ρ_{air} is small, the speed of a falling raindrop, $w_{T,\text{ rain}}$,will be faster than near the surface when ρ_{air} is similar magnitude to ρ_o. As the air density increases, the frictional drag on a drop also increases.

Note that this equation for raindrops is a vast simplification because raindrops are not typically spherical shaped. As they fall, the passing air deforms them into pancake shaped drops. However, this equation provides an approximation for fall velocity.

So how can cloud droplets grow to form raindrops? The condensation process is not enough and is far too slow to produce raindrops. From a volumetric view, it takes 1 million cloud droplets (10 μm radius) to combine together to make one single raindrop (1000 μm = 1 mm radius). There has to be another faster process by which cloud droplets can grow or combine together to become large and heavy enough to fall.

We will discuss two primary rain formation theories.

1. Collision-Coalescence Process
2. Ice Phase Processes (Wegener-Bergeron-Findeisen)

Collision-Coalescence Process

In warm clouds, where all of the cloud droplets are liquid, the **collision-coalescence** process is the primary mechanism responsible for producing precipitation. This is thought to be the case especially over tropical oceans. The collision-coalescence process is exactly as it sounds: cloud droplets collide and coalesce or stick together. Larger cloud droplets have slightly higher terminal velocities, because they have a smaller surface-area-to-weight ratios. This advantage allows them to fall faster and collide

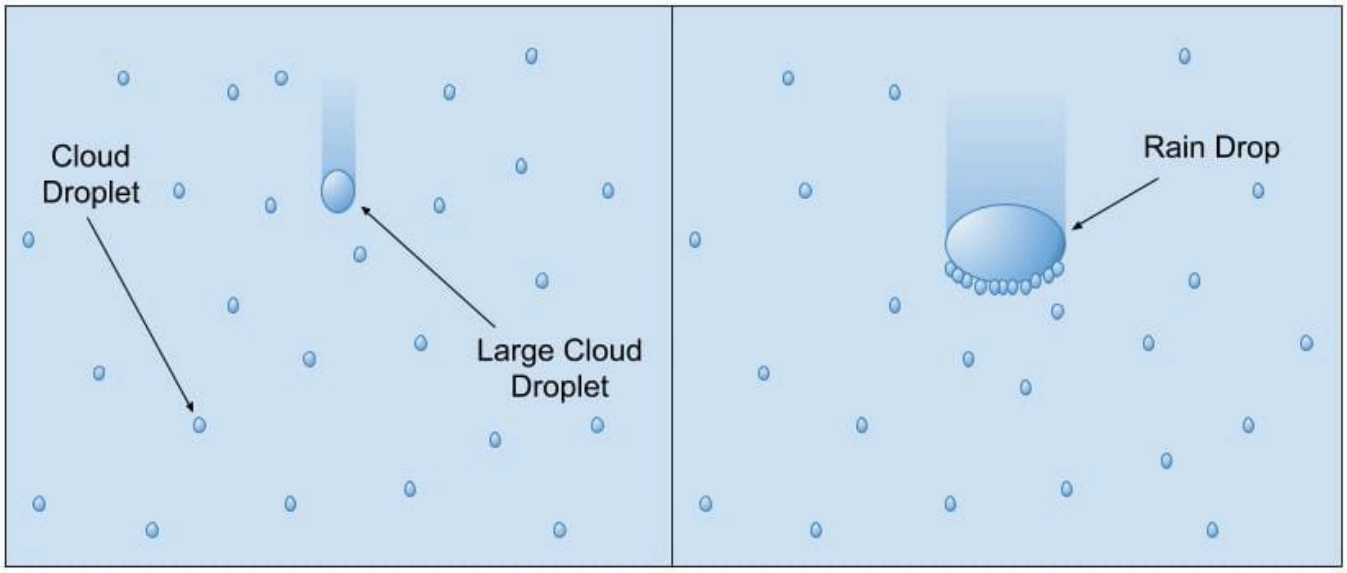
with smaller cloud droplets. Sometimes the cloud droplets will stick together and coalesce to form a larger droplet. This begins a positive feedback where these larger droplets then fall even faster, collide with even more smaller droplets in their path, and aggregate more and more cloud droplets together. However, note that collision between cloud droplets does not always mean that coalescence will occur. Sometimes droplets will bounce apart during collision if their surface tensions are too strong. For collision-coalescence to begin, a cloud needs to have a wide distribution of cloud droplet sizes. This can occur from a variation in CCN type—for example, sea salt aerosols are particularly large—or from random collisions between droplets.

The total amount of liquid water in a cloud as well as the time that a cloud droplet spends inside of a cloud influences how large it can grow through the collision-coalescence process. The cloud height is of course a factor here, but its a little more complicated than that. Rising motion in a cloud will slow the downward speed of a falling droplet. This can act to increase the amount of a time a cloud droplet spends in a cloud, as well as the size it will grow. Let's cover a few examples.

Deep cumulus clouds with convective updrafts tend to produce larger raindrops because upward motion is strong and droplets have a long time in the cloud to grow. In fact, the droplets need to become sufficiently large in order for their fall velocity to overcome the updraft velocity.

On the other hand, stratus clouds are typically not very thick and have weak updrafts, so droplets in these clouds don't spend a long time in the cloud itself, and therefore are not be able to grow very large. If there is moist air below the stratus cloud, the drops may reach the ground as a light drizzle. However, if there is dry air below the stratus cloud, the drops may evaporate before they are able to reach the ground.

To summarize, in warm clouds, cloud droplets grow to precipitation sized drops through the collision-coalescence process. The most important factor in raindrop production is the liquid water content of a cloud. Assuming the cloud has sufficient water, other factors that affect raindrop production are: thickness of the cloud; strength of updrafts within the cloud; cloud drop distribution of sizes; and difference in electric charge of the droplets and the cloud itself. Thin stratus clouds with weak vertical motion may produce weak drizzle, if any, while tall cumulus clouds with strong updrafts can produce heavy rain showers. The following image illustrates the collision-coalescence process of raindrop production.



The collision-coalescence process occurring in warm clouds. The left image shows the importance of a range of cloud droplet sizes to initiate the collision-coalescence process while the image at right shows the acceleration of the process once a raindrop forms (Image Created by Britt Seifert).

Ice Phase Process

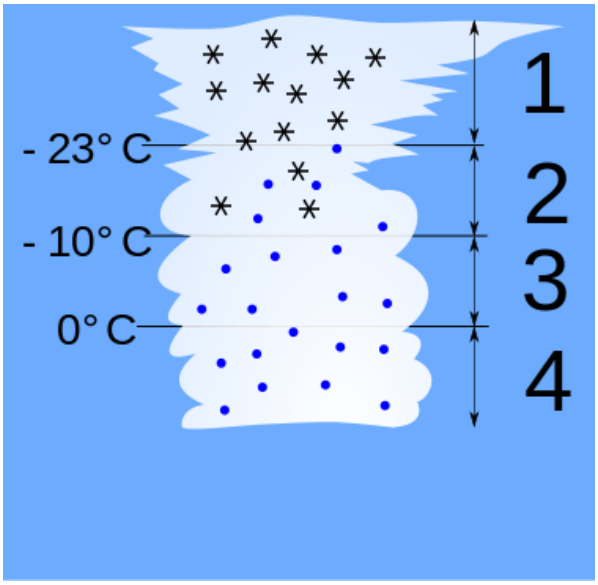


Ice crystals forming on a window at 30,000 feet (CC BY-SA 3.0).

Outside of the tropics, the ice phase process of rain formation is the primary mechanism producing most of the worlds precipitation. The ice phase process occurs in cold clouds or clouds with temperatures below 0°C. To understand why, we need to know something about freezing of liquid water droplets.

Supercooled Water and Ice Nuclei

Supercooled water is liquid water that exists below the freezing point of water (below 0°C). Similar to how cloud droplets need a surface on which to condense, ice crystals also need a nucleus or ice embryo to freeze. Without an ice nucleus, liquid water drops can remain liquid in temperatures as low as -40°C. Once beyond -40°C, all hydrometeors (water particles) will exist in the solid state. Typically the distribution of liquid and solid hydrometeors in a cloud looks like the following image.



A simple diagram showing the distribution of liquid and solid phase hydrometeors in a mixed phase cloud (CC BY-SA 3.0).

At low elevations above freezing (region 4), the hydrometeors in the cloud exist as liquid droplets. Above the freezing level (region 3), supercooled liquid droplets exist. Above that, some liquid droplets begin to freeze, and both liquid and ice phase hydrometeors co-exist (region 2). Finally, above some level where the temperature is cold enough, all hydrometeors will exist in their solid state

(region 1).

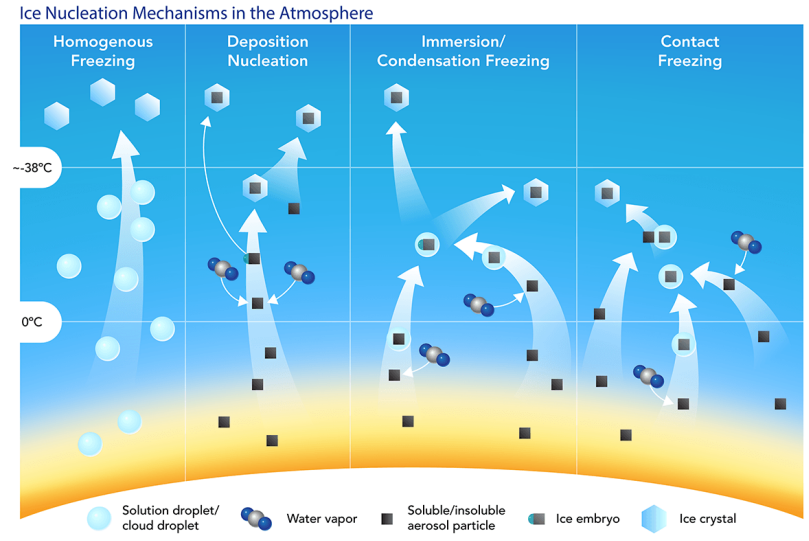
When liquid water droplets freeze without any sort of nucleus, this is known as **homogenous** or **spontaneous freezing**. While this occurs within a large body of freshwater at temperatures slightly below 0°C, cloud droplets will not freeze spontaneously until temperatures are -40°C or lower.

For droplets to freeze spontaneously, enough molecules within the droplet must form a rigid pattern and become a tiny ice structure known as an *ice embryo*. When this embryo grows sufficiently large, at a certain size it will act as an ice nucleus, which are described below. The other molecules in the droplet then become attached to the ice structure and the entire droplet freezes.

Tiny ice embryos are able to form when water drops just below freezing, but typically at these temperatures there is enough thermal agitation to weaken their structure and break them apart. At lower temperatures, there is less thermal motion, and ice embryos have a better chance of growing large enough to freeze the surrounding water. When you have larger volumes of water, ice embryos have a better chance of growing large enough to freeze the surrounding liquid before being broken up, but this becomes more and more difficult with smaller volumes of water. Only the largest cloud droplets can freeze spontaneously without a nucleus at temperatures below -40°C. In most cases, ice nuclei are required for ice crystals to form in sub-freezing clouds.

Just as CCNs are required for liquid cloud droplets to form, ice crystals form on particles called **ice nuclei (IN)**. Particles serve as effective IN if they have similar geometry to an ice crystal, for example, ice itself is an effective IN. There are not many IN in the atmosphere, especially at temperatures above -10°C, but certain types of particles become active IN with lower temperatures. For example, dust can be an effective IN. Ice nuclei are rare compared to hygroscopic cloud condensation nuclei.

Some IN allow water vapor to immediately become solid ice when they come in contact together. These are known as *deposition nuclei* because the water vapor changes phase directly into solid ice without becoming liquid first (phase change from gas to solid is called deposition). IN that are effective at causing the freezing of supercooled liquid droplets are called **freezing nuclei**. Some freezing nuclei must be immersed in a liquid drop in order to freeze it, while others are effective at inducing condensation and then freezing. Many freezing nuclei will cause supercooled droplets to freeze as soon as they collide, which is called **contact freezing**, and these nuclei are referred to as **contact nuclei**. These different freezing methods are outlined in the figure below.



Four primary mechanisms that form ice in the atmosphere: homogeneous freezing; deposition nucleation; immersion freezing; and contact freezing (CC BY-SA 4.0).

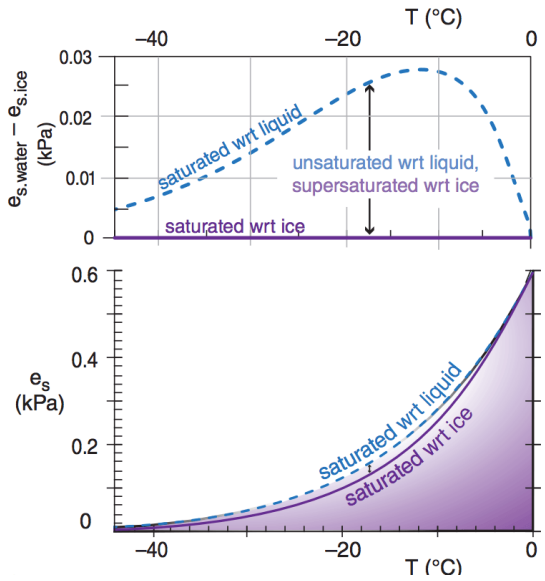
To summarize, cloud droplets may freeze spontaneously, but only at very low temperatures. Ice nuclei can help the growth of ice crystals, but they are not naturally abundant.

Saturation Vapor Pressure

So, we have cold clouds that contain many more liquid cloud droplets than ice crystals, even at sub-freezing temperatures, and these particles are not large/heavy enough to precipitate out of the cloud. How do we get rain and snow out of the ice-crystal

process then?

Imagine a cloud with super cooled liquid water and saturated air. At saturation, the liquid droplets are in equilibrium with the water vapor in the air. The number of water molecules leaving and entering the surface of the liquid droplets are equivalent. Now imagine that an ice crystal forms by one of the processes described above. In the below-freezing portion of a cloud, this ice crystal is surrounded by many liquid supercooled droplets. Because the saturation vapor pressures with respect to liquid and ice are slightly different, the presence of this new ice crystal has a big impact on the cloud.



The saturation vapor pressure with respect to liquid (blue) and with respect to ice (purple). The difference is shown at the top (CC BY-NC-SA).

With respect to liquid, the liquid droplets were at saturation. But now, with respect to ice, the ice crystal is in an environment that is supersaturated. You can think of this as the following: it is easier for water molecules to escape a liquid surface through evaporation than to escape a solid surface. This means that there will be many more molecules escaping the liquid water surface at a given temperature and will require more water vapor around it in order to keep the droplet in saturation. At the same temperature, the saturation vapor pressure above a surface of water is greater than the saturation vapor pressure above a surface of ice.

This difference in saturation vapor pressure with respect to water and ice causes water vapor molecules to deposit from the environment onto the ice crystal. Because the vapor molecules are being removed from the environment around the liquid droplet, the vapor pressure with respect to the water surface decreases. This throws the water droplets out of equilibrium with their surroundings, causing them to evaporate, replenishing the removed water vapor from the environment. This provides additional moisture for the ice crystal, allowing it to grow at the expense of the liquid droplets.

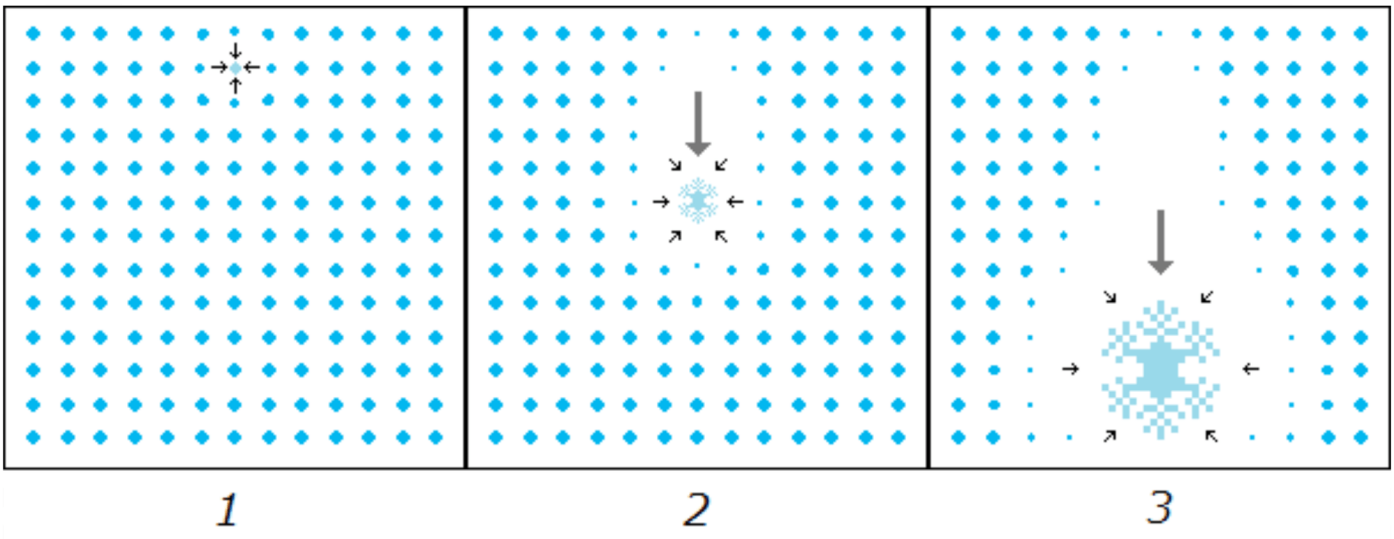
This process is called the **Wegener-Bergeron-Findeisen** (or more generally, the ice phase) process. Ice crystals in a sub-freezing region of a cloud will grow larger at the expense of surrounding water droplets.

Falling Ice Crystals

Imagine this happening throughout a large cloud. The water vapor within the cloud as well as the water vapor from evaporating supercooled liquid droplets provides a continuous source of moisture for ice crystals, allowing them to grow rapidly. Eventually, these ice crystals become large enough to fall. The same issues exist with updrafts and rising air, but at some point the crystals will fall faster than the updrafts within the cloud.

Sometimes, ice crystals collide with nearby supercooled droplets in the cloud, causing them to freeze onto the crystal as ice. The ice crystal will grow larger and larger as it collides into more droplets, this is called **riming or accretion**. This forms an icy clump called **graupel** (also known as snow pellets), which may break apart into tiny ice particles as it collides with more droplets. These splinters may form graupel of their own as they collide into other droplets, which in turn may also splinter, causing a chain reaction.

In clouds that are colder, ice crystals may collide together and break apart into smaller ice particles, which act as tiny seeds that can freeze supercooled droplets on contact. This could also cause a chain reaction that produces many ice crystals. As these ice crystals fall, they can collide and stick together. This process of collision and sticking is called **aggregation**. A fully grown ice crystal is what we call a **snowflake**.



The growth of ice crystals occurs during its descent through aggregation (CC 0).

There must be many times more water droplets in a cloud than ice crystals for ice crystals to get large enough to fall as snow—on the order of 100,000:1 to 1,000,000:1.

This page titled 9.3: Cloud Droplets is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by LibreTexts.

9.4: Precipitation Types

The ice phase process accounts for most global precipitation. The term **precipitation** refers to all types of precipitation—from fog, rain, snow, hail, etc. Anything composed of water that is falling in the atmosphere can be called precipitation. However, we know that not everywhere in the world gets snow all the time. While the ice phase process helps with precipitation formation, many things can happen to falling drops along their journey from the cloud to the ground. Here are a few examples.

Rain: Ice crystals melt before they hit the ground.

Snow: Ice crystals collide and stick, forming a fully grown ice crystal, and falling to the ground as a snowflake.

Graupel: Ice crystals collide and stick with other ice crystals forming clumps of snow called graupel.

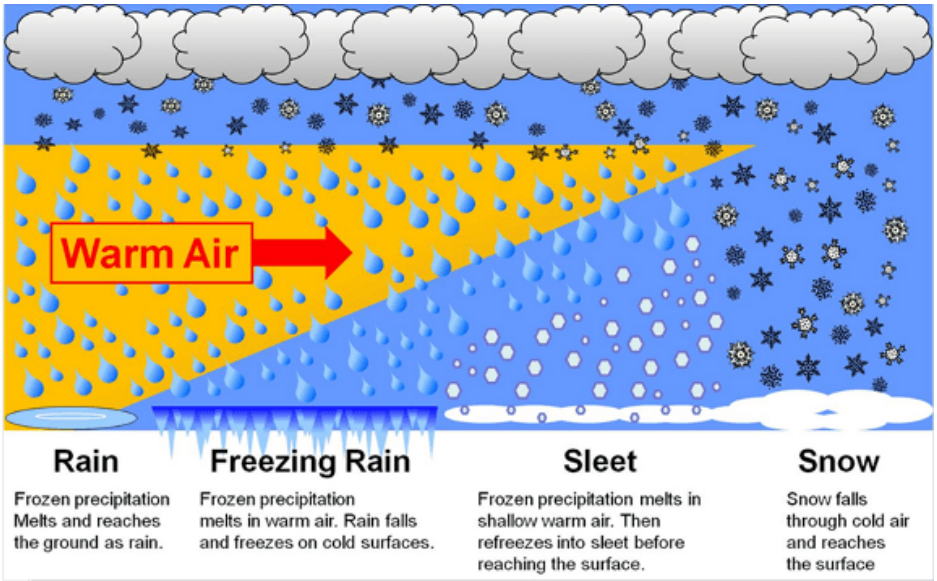
Sleet: A mixture of rain and snow, formed by partial melting.

Freezing Rain: Supercooled liquid rain that freezes on impact with the surface.

Ice Pellets: Ice crystals melt before they hit the ground, refreeze in a cold layer, usually just above the surface, and end up falling as frozen rain drops.

Hail: Ice crystals that repeatedly pass through a supercooled liquid region of cloud where riming builds up on the hydrometeor. The formation of hail requires strong updrafts and a relatively long time inside of a cloud.

As you can see, the journey of an ice crystal after formation is not always straightforward and depends strongly on the environmental conditions, especially temperature. In the following chapters (especially chapter 12), we will learn how airmasses and fronts combine together regularly in Earth's atmosphere to create conditions that are conducive to all types of precipitation.



Precipitation changes associated with the passage of a warm front (Public Domain).

The above image provides an example of how a warm front creates temperature gradients in the atmosphere that produces precipitation types from rain, freezing rain, sleet, and snow depending on your location with respect to the front.

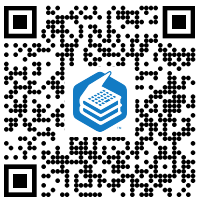
Chapter 7: Questions to Consider

1. Which rain formation process produces most of the worlds precipitation? What type of cloud does this process occur in?
2. What factors determine how large a cloud droplet can grow through collision-coalescence?
3. Match the ice nucleation methods:

Query 9.4.1

This interactive element is not available in this format. Visit the link or scan the QR code to access it online.

View this Interactive Content online



Query 9.4.2

4. *This interactive element is not available in this format. Visit the link or scan the QR code to access it online.*

View this Interactive Content online



Selected Practice Question Answers:

Query 9.4.3

This interactive element is not available in this format. Visit the link or scan the QR code to access it online.

View this Interactive Content online



This page titled 9.4: Precipitation Types is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by LibreTexts.

10: WEATHER OBSERVATION AND MAPS

10.1: Meteorological Reports and Observations
10.2: Synoptic Weather Maps
10.3: Map Analysis

Chapter 10: Weather Observation and Maps - Introduction

Weather is something we experience at a particular place and time, but meteorologists need to understand what is happening across much larger areas. To do this, observations of temperature, pressure, wind, cloud cover, and other atmospheric conditions are collected at many locations and reported using standardized methods. These observations provide the basic information needed to describe and analyze the atmosphere.

In this chapter, we will examine how weather observations are collected, reported, and displayed. We will look at common meteorological reports and learn how individual observations can be represented using station models. When observations from many locations are combined on synoptic weather maps, patterns in temperature, pressure, wind, and other variables begin to emerge.

Finally, we will explore how meteorologists analyze these maps using features such as isobars and isotherms. Identifying these larger patterns allows us to move beyond describing weather at a single location and begin understanding weather systems across entire regions. These skills will provide an important foundation for our study of atmospheric circulation, air masses, fronts, and severe weather in the chapters that follow.

This page titled 10: Weather Observation and Maps is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Robert Wagner.

10.1: Meteorological Reports and Observations

The **World Meteorological Organization (WMO)** is a branch of the United Nations and it sets global weather observation standards to be used by countries across the globe. Weather systems span countries and continents so weather observations have to be synchronized to get an accurate big-picture view (a **synoptic** view) of the weather at a given time. Upper air and surface observations are taken at specific times in **Coordinated Universal Time (UTC)** so that they can be coordinated simultaneously across different time zones. For example, most upper-air observations are taken at 00 and 12 UTC, but surface observations are typically taken more frequently. Synoptic weather maps put together these observations.

Observations are reported using internationally standardized codes. One of the most commonly used codes is **METAR**, which comes from the French *MÉTéorologique Aviation Régulière*, or *Meteorological Terminal Aviation Routine*. It is a summary of surface weather conditions reported at hourly or half hourly intervals depending on the station. METAR is provided by airport terminals for the purposes of aviation meteorology. In general, METAR includes reports of wind speed and direction, visibility, cloud layers at different levels of the atmosphere, surface temperature and dew point temperature, air pressure, as well as weather conditions such as precipitation or thunderstorms near the station. **SPECI** is a special non-routine form of METAR, provided when airport weather conditions change significantly. You are not required to translate or memorize METAR for the purposes of this class, but it is useful to recognize it because it is very commonly used. Translations for parts of METAR code can be easily found online.

METAR: PHNL 072153Z 05012KT 10SM FEW025 FEW050 BKN200 31/17 A3006 RMK AO2 SLP180 T03060167

Above is an example of METAR, taken from Honolulu International Airport. “PHNL” is the ICAO airport code for Honolulu International. The “072153Z” indicates that the report was given on the 7th of the month (August 7th) at 2153Z (Universal Time), which is 11:53 AM local time. The “05012KT” is the wind report, and indicates that winds are blowing at 12 knots from 50°, which is roughly from the northeast. The “10SM” indicates that visibility is 10 statute miles, which is another way of saying that visibility on the runway is clear. If visibility is 10 statute miles or greater, it is always reported as just “10SM”. The “FEW025 FEW050 BKN200” are reports of different cloud heights, which are important for aviation purposes. This report says that there are is a layer of a “few” clouds at 2,500 and 5,000 feet, with a layer of “broken” clouds at 20,000 feet. The “31/17” is telling us that the temperature is 31°C with a dew point temperature of 17°C. The “A3006” gives the station air pressure at 30.06 in Hg (inches of Mercury, this unit is still primarily used in aviation). The “RMK” denotes the remarks section of a METAR report where additional remarks and information about the weather are provided. The “AO2” is a code that indicates that the airport observing site is automated and contains a precipitation sensor. Some sites are not automated or do not have a precipitation sensor so there are different codes to denote this. Automated reports can also have non-automated remarks added to them. The “SLP180” gives the sea level pressure as 1018.0 mb. And the “T03060167” gives a more accurate temperature and dew point temperature reading. It reports that the temperature is actually 30.6°C and the dew point temperature is actually 16.7°C.

Weather Observation Locations

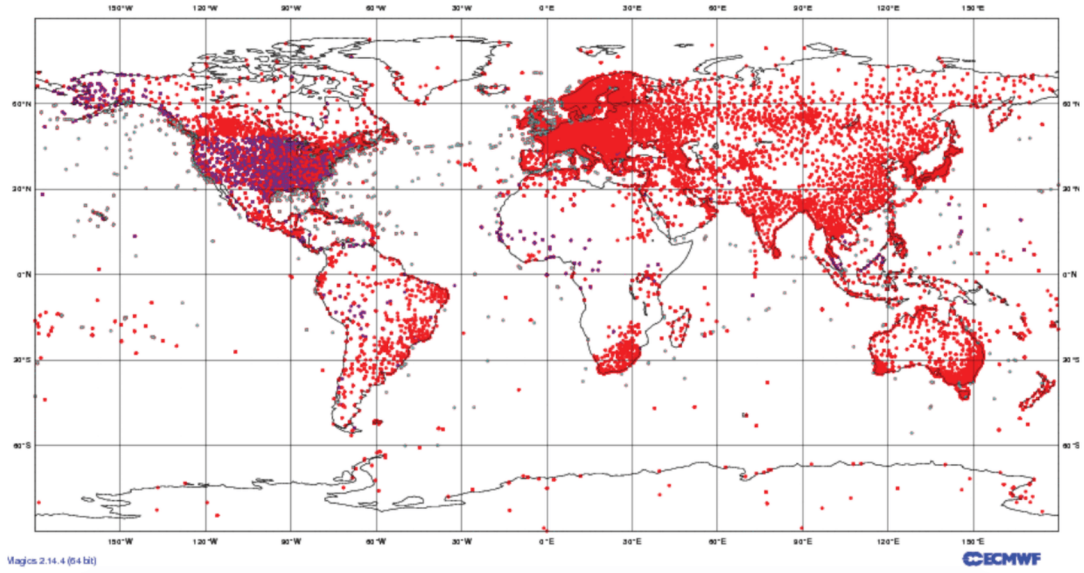
Surface weather observations include automated observations from Automated Surface Observing System (ASOS) sites in the United States, as well as hourly observations from airports around the world, reported as METAR. Manual and ship observations are also made at specific times. Surface observations of the weather are important for providing information about the weather conditions that we experience here on the ground. These conditions include temperature, dew point temperature, wind speed and direction, air pressure, cloud cover, visibility, and weather conditions such as precipitation or thunderstorms.

Although surface weather observations are important, they only tell a part of the full story. Just as ripples on the surface of a river can be a sign of what is happening below, surface observations can give an idea of what is happening above. Much of the weather that happens at or near the ground is strongly affected by conditions higher up in the atmosphere.

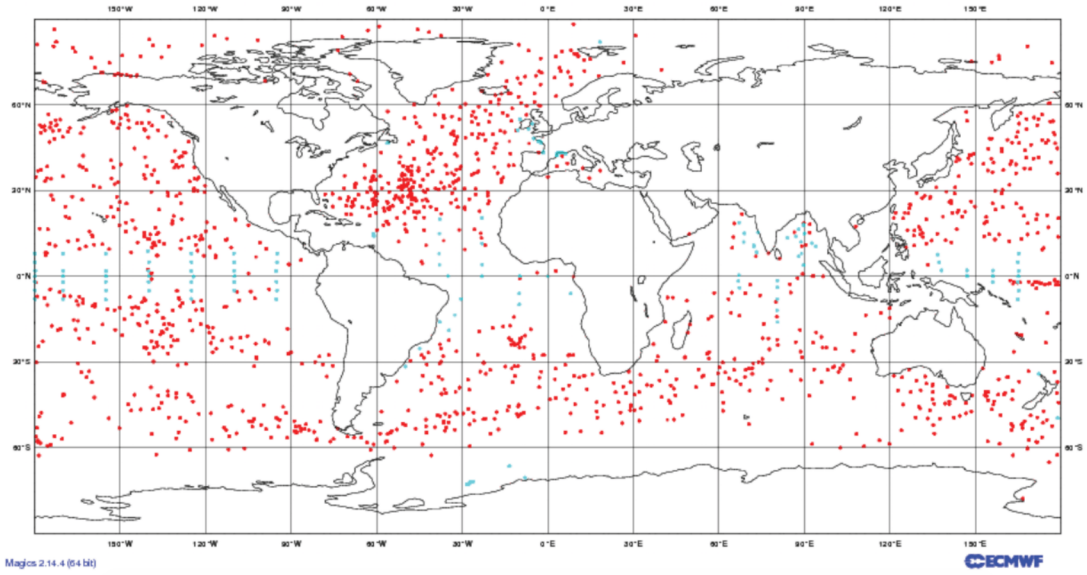
Observations of the upper atmosphere are taken by **radiosonde** packages, which are carried aloft by weather balloons. These radiosonde observations (**RAOBs**) provide **soundings** of the upper-air environment, giving information about temperature, humidity, and pressure at vertical levels throughout the atmospheric column. Recall from Chapter 5 that temperature and humidity aloft are usually plotted against pressure in a thermodynamic diagram called a Skew-T. Some radiosondes can infer wind data at different heights—these are called **rawinsondes**. When these instrument packages are dropped from an aircraft, they are called **dropsondes**. All of this data is accumulated, organized, tested, and stored in computers at governmental centers such as the **European Centre for Medium-Range Weather Forecasts (ECMWF)**. Other data collecting systems include weather radar (such

as the **NEXRAD** network in the United States) and satellite, as well as vertical **wind profilers** and Radio Acoustic Sounding Systems (**RASS**).

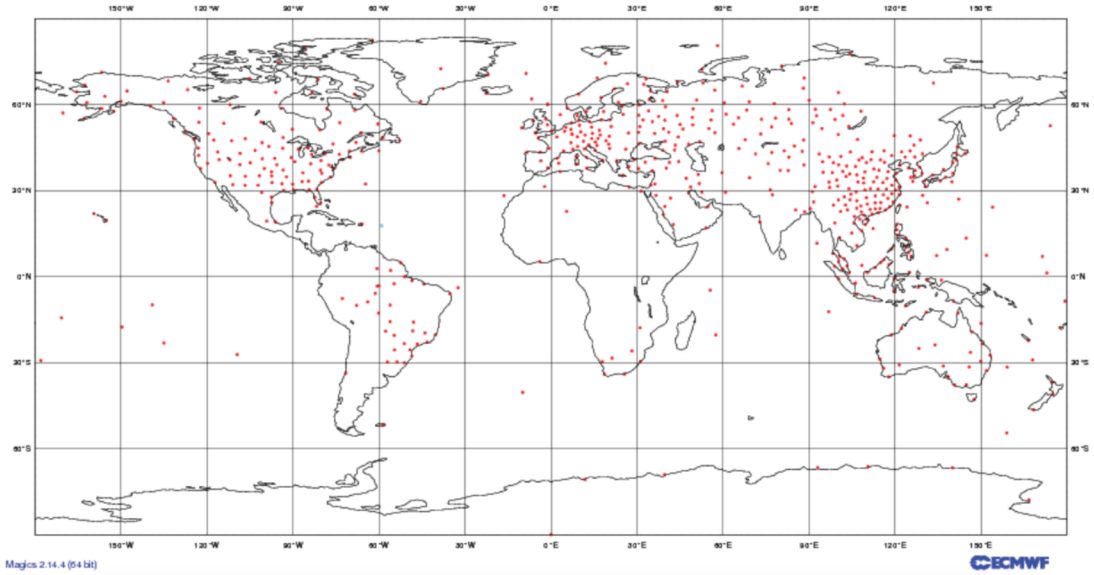
The following figures show locations of data collected by ECMWF over a 6-hour time period. The first figure shows surface observation locations located on land, the second shows surface observations over the ocean, and the third shows the upper air sounding observational network.



Surface observation locations for temperature, humidity, winds, clouds, precipitation, pressure, and visibility (CC BY-NC-SA 4.0).



Buoy observation platforms for surface temperature and winds over the ocean (CC BY-NC-SA 4.0).



Upper air observation and sounding locations for temperature, pressure, and humidity aloft. Stull Figure 9.4 (CC BY-NC-SA 4.0).

Sea-Level Pressure Adjustment

Weather stations exist at many altitudes. Because air pressure decreases with height, weather stations in cities at high elevations will report far lower air pressures than cities at lower elevations. Air pressure varies vertically far more than it does horizontally. To make air pressure comparable, the station pressure needs to be adjusted to the pressure it would have if the station were at sea level. If this were not done, pressure differences between nearby surface stations would be dominated by their difference in elevation, and a surface pressure map would more closely resemble a map of elevation rather than a map of atmospheric pressure. By correcting for altitude differences, synoptic maps show mean sea level pressure, which is used to show low and high pressure centers near the surface.

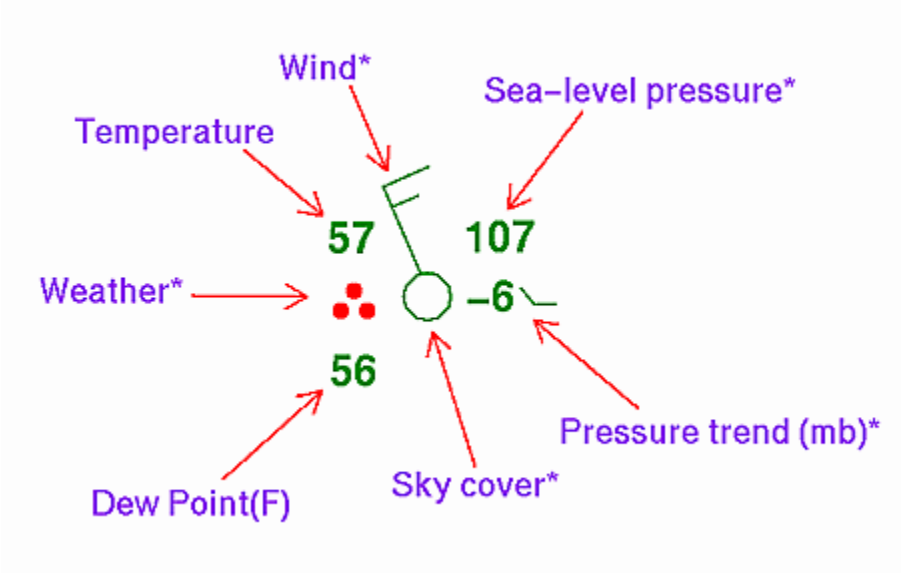
This page titled 10.1: Meteorological Reports and Observations is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by LibreTexts.

10.2: Synoptic Weather Maps

Surface weather observations taken at many weather stations at a given time can be shown on a weather map through the use of a **station plot model**, an example of which is given below. Station plot models typically provide wind information through wind barbs, as well as current weather conditions, sea level pressure, temperature, dew point temperature, visibility, and cloud cover. Sometimes additional information such as pressure tendency, cloud type, and precipitation amounts are also given.

When reading wind barbs, keep in mind that the flag always points in the direction that the wind is coming from. You can think of it as an arrow flying from a bow, the flag represents the feathers, which are at the back of the arrow in the direction the arrow is coming from. The tip is the direction the arrow is flying toward. More information on wind barb and cloud cover symbols are provided below.

Sea level pressure is typically given in three digits, with the last digit being the nearest tenth. The initial 9 or 10 is left out. So, “147” is actually 1014.7 mb and “998” would be 999.8 mb. When reading station plot pressure, mentally place a decimal point to the left of the third digit, and use your intuition and the pressure of nearby stations in order to decide if the pressure has a leading 9 or 10.



Example station plot model (Public Domain).

The table below provides information for wind barbs. Two concentric circles with no lines indicates calm wind and a line with no barbs is 1-2 speed units. A half line is 5, a full line is 10, and a triangle is 50. On many weather maps, the unit for winds is given in knots, although different weather maps can use different units so it is always good to check. The lines and triangles are read a bit like roman numerals—just add them up!

Symbol	Wind Speed	Description
◎	calm	two concentric circles
——	1 - 2 speed units	shaft with no barbs
⟍——	5 speed units	a half barb (half line)
\——	10 speed units	each full barb (full line)
◣——	50 speed units	each pennant (triangle)

• The total speed is the sum of all barbs and pennants. For example, ◣\\\— indicates a wind from the west at speed 75 units. Arrow tip is at the observation location.

• *CAUTION: Different organizations use different speed units, such as knots, m s^{-1}, miles h^{-1}, km h^{-1}, etc. Look for a legend to explain the units. When in doubt,* assume knots *— the WMO standard. To good approximation, 10 knots ≈ 5 m s^{-1} .*

Wind barb chart (CC BY-NC-SA 4.0).

The table below provides sky cover information, which is typically shown in the circular center of the station plot model. Depending on the symbol type, it can indicate clear sky or various levels of overcast skies.

Sky Cover (oktas)	Sym-bol	Name	Abbr.	Sky Cover (tenths)
0	○	Sky **Clear**	SKC	0
1	◴	**Few*** Clouds	FEW*	1
2	◔			2 to 3
3	◔	**Scattered**	SCT	4
4	◑			5
5	◕	**Broken**	BKN	6
6	◕			7 to 8
7	◕			9
8	●	**Overcast**	OVC	10
(9)	⊗	Sky **Obscured**		un-known
(/)	⊖	Not Measured		un-known

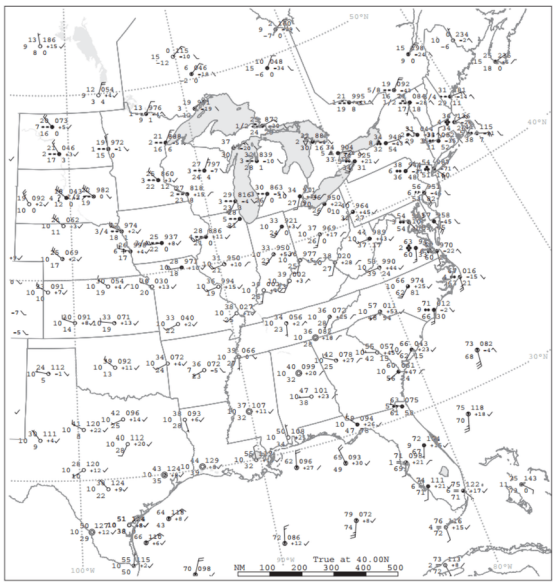
* "Few" is used for (0 oktas) < coverage ≤ (2 oktas).

Sky cover chart. Stull Table 9-1θ (CC BY-NC-SA 4.0).

This page titled 10.2: Synoptic Weather Maps is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by LibreTexts.

10.3: Map Analysis

Surface observations as described above are plotted on a map in the form of station plot models, as in the figure below. When looking at the map of raw data below, you can get an idea of the prevailing wind direction in different parts of the US, which areas are experiencing widespread rain, which areas are overcast, and from the wind direction patterns you can even infer centers of high and low pressure. However, much of this information is not very clear without doing some analysis on the map. Map analysis can either be done by hand or by computer, and involves the drawing of **contours**, or **isopleths** (lines of equal value) to connect areas of constant air pressure and temperature at the surface. Aloft, contours may be drawn to show areas of constant height, constant humidity, constant wind speed, and other parameters of interest at constant pressure levels. Map analysis also includes drawing and labeling boundaries in the atmosphere, such as **fronts** or **dry lines** to show the locations and movement of air masses. Fronts will be covered in a later chapter. Here, we will focus on lines of constant pressure, **isobars**, and lines of constant temperature, **isotherms**.

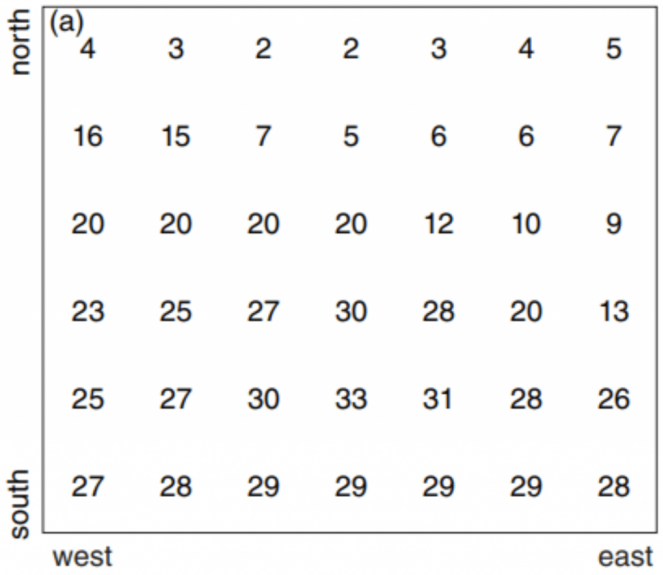


Map with station plot models (CC BY-NC-SA 4.0).

Isopleths

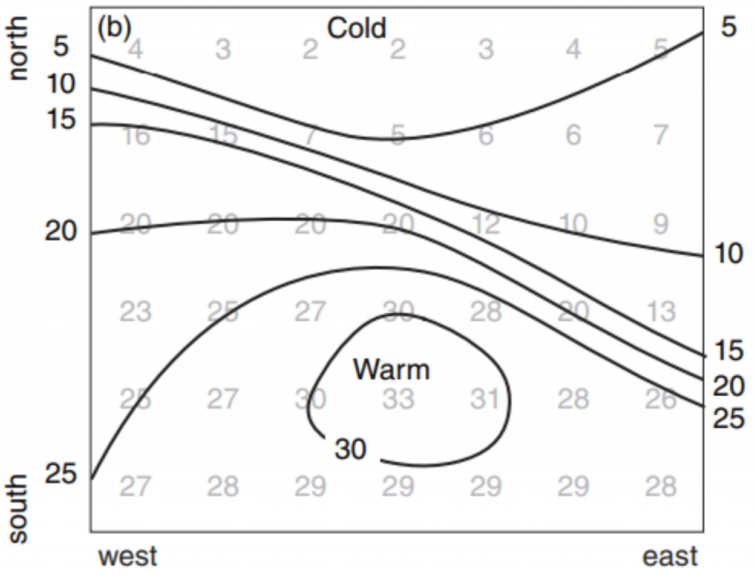
The below figure shows a grid of equally spaced temperature values in °C. When analyzing the temperature on a map, you draw isopleths of constant temperature: isotherms. The main purpose of this is to separate areas of warmer air from cooler air, and to get an idea of how quickly the temperature changes across a horizontal distance, which is also known as the temperature gradient. Isotherms are typically drawn every 5°C on the 5's – for example, 5°C, 10°C, 15°C, 20°C, 25°C, and so on. Standard conventions are used for contouring on different map surfaces, for example, isotherms are usually dashed and shown in red. You will not need to know these conventions for this class, but it may still come in handy to be aware.

In looking at the below temperature field, you can see that there are many values that do not line up exactly with the lines you need to draw. Many values lie in between contours, such as 4°C and 16°C. You will need to interpolate data, meaning you will have to infer where a datapoint is based on the data around it. Starting from the upper left corner, imagine that you are going to draw your 5°C isotherm. Five lies much closer to 4 than 16, so you will start your isotherm close to the 4. 5°C will also lie in between 3 and 15 °C, but slightly farther away from the 3 value, so you will slope your isotherm down slightly. Five degrees centigrade will also lie in between 2 and 7 °C, so you will continue your isotherm to the right. However, it will slope down even further, because 5 lies closer to 7 than to 2. Your isotherm will then cross directly through the 5°C data point and slope back upward slightly between the 3 and 6 and directly through the middle between the 4 and 6 and back up through the 5 in the upper right hand corner.



Blank temperature field (CC BY-NC-SA 4.0).

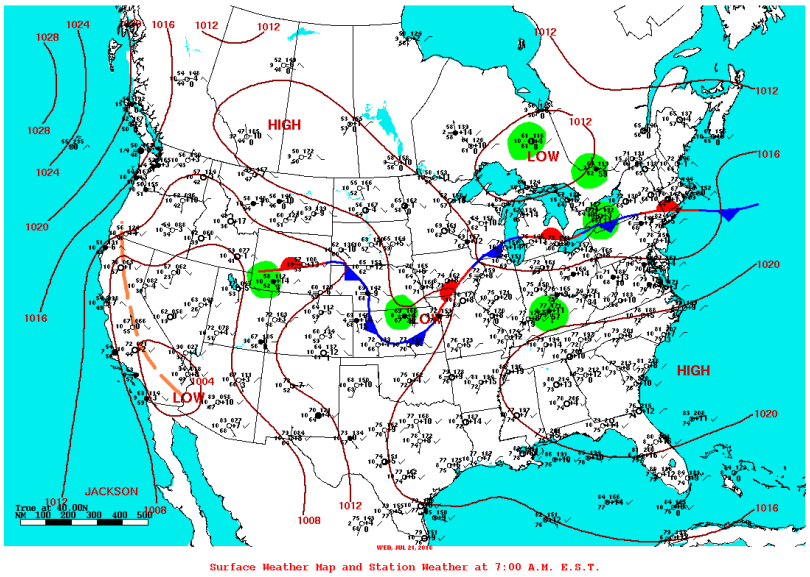
The complete contoured example of this temperature field is given below. The atmosphere is a continuous fluid, so any fields that you are contouring (pressure, temperature) will never have values that jump or suddenly end. Contours will either be closed (both ends will connect) or they will extend to the edge of the map. They will *never* cross or end suddenly. If two contours were to cross, it would mean that one place has two different air temperatures at one time, which is impossible. Areas where isotherms are close together indicate a strong temperature gradient, which may be indicative of a frontal zone.



Completed temperature field with isotherms (CC BY-NC-SA 4.0).

Isobars

Pressure fields are analyzed by drawing **isobars**, or lines of constant air pressure. The surface mean sea level pressure is analyzed in much the same way as the isotherms above. However, the conventions are different. Typically, pressure is analyzed every 4 mb and centered on 1000 mb. Isobars are typically drawn as solid black lines. The mean sea level pressure field is analyzed in order to identify areas of high and low pressure, as shown in the figure below. Areas where isobars are drawn close together indicate places where the pressure gradient force is strong. A stronger pressure gradient force is indicative of stronger wind speeds. As you will learn later, winds tend to flow counterclockwise around areas of lower pressure, and clockwise around areas of high pressure in the Northern Hemisphere.



Surface map with isobars and station plot models (Public Domain).

Because weather information is shared across continents and across country borders, it is important to maintain standards such that everyone effectively speaks the same language. Weather is one of the greatest international collaborations ever. If it weren't for China and Russia sharing their weather information with the US, our long range forecasts would be more inaccurate because we wouldn't know the initial conditions of the atmosphere upstream of the North American continent.

This page titled 10.3: Map Analysis is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by LibreTexts.

11: ATMOSPHERIC CIRCULATION & AIR MASSES

11.1: Atmospheric Circulation

11.1.1: Main Forces

11.1.2: Differential Heating

11.1.3: Single-Cell Model

11.1.4: Three-Cell Model

11.1.5: Global Surface Winds

11.1.6: Jet Streams

11.2: Air Masses and Fronts

11.2.1: Air Masses

11.2.2: Surface Fronts

Chapter 11: Atmospheric Circulation & Air Masses — Introduction

The atmosphere is constantly in motion. Differences in solar heating create differences in temperature and pressure, setting air into motion and producing winds. Earth's rotation then modifies that motion, helping create the large-scale circulation patterns that transport heat and moisture around the planet. In this chapter, we will build from the forces that influence individual winds to the global circulation of the atmosphere.

We will begin by examining the major influences on atmospheric motion, including pressure differences, the Coriolis effect, and friction. **Some of this material is presented mathematically in the text, but our primary goal will be to understand what these forces do and how they affect the speed and direction of the wind. Unless specifically discussed in lecture or assigned in an activity, you do not need to reproduce or perform calculations with the more advanced equations presented in this chapter.** We will then use these ideas to understand global wind patterns, atmospheric circulation cells, and jet streams.

Finally, we will examine air masses and fronts. Large bodies of air acquire temperature and moisture characteristics from the regions where they form, and their movement brings changing weather to different locations. Where contrasting air masses meet, fronts develop and can produce clouds, precipitation, and rapid changes in weather. These concepts provide an important bridge to the next chapter, where we will examine cyclones, thunderstorms, and severe weather.

This page titled 11: Atmospheric Circulation & Air Masses is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Robert Wagner.

11.1: Atmospheric Circulation

This section explains how differences in atmospheric pressure, combined with Earth's rotation and surface friction, produce winds and organize atmospheric circulation. It introduces the major forces that affect wind speed and direction and examines how these forces interact to produce characteristic wind patterns. These ideas are then applied to local winds, global circulation, and upper-level winds such as the jet streams.

This page titled 11.1: Atmospheric Circulation is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Robert Wagner.

11.1.1: Main Forces

There are **five forces** that influence the speed or direction of horizontal winds.

1. Pressure gradient force
2. Advection
3. Centrifugal force
4. Coriolis force
5. Turbulent drag

Remember from Chapter 1 that according to the Cartesian coordinates, x points **east**, y points **north**, and z points **upwards**. To define wind, we use wind components u, v, and w which correspond to the x, y, and z directions. These wind components are used in equations of motion used to predict wind to designate direction in a three-dimensional plane. We'll go through each of the five forces one by one to discuss how they affect wind speed and/or direction.

Pressure Gradient Force

A **pressure gradient** (PG) is a change in pressure over a distance. Therefore, the units of the pressure gradient force are Pascals per meter (P m^{-1}). The pressure gradient can be calculated simply as the change in pressure divided by the distance over which that change occurs. The size or strength of the pressure gradient determines the size or strength of the force that results from it.

$$PG = \frac{\Delta P}{\Delta x}$$

The **pressure gradient force** (PGF) is a force from high to low pressure over a distance. Without differences in pressure, there would be no wind because there would be nothing to accelerate airflow. In Chapter 9 we learned about isobars, or lines of constant pressure. When isobars are tightly packed, we know that there is a strong pressure gradient or large change in pressure over a relatively short distance. A strong pressure gradient results in a large pressure gradient force and, as we'll see, higher wind speeds. Depending on the text you read, you'll see a number of different forms of the pressure gradient force. Two equivalent representations are given below.

$$\frac{PGF_x}{m} = -\frac{1}{\rho} \cdot \frac{\Delta P}{\Delta x}$$

The PGF as defined above is simply a change in pressure divided by the distance and air density. This gives the force per unit mass, hence PGF/m . Above we define PGF in the x-direction, but it can act in any horizontal plane. The units of the PGF will be in units of acceleration above because Force = Mass * Acceleration, but because we're dividing the left-hand side by Mass the units will be m s^{-2}. The negative signs in the above equation is due to the fact that the PGF acts from high pressure to low pressure.

Another equivalent way to represent the PGF is as follows.

$$|PGF| = \text{ Volume } \cdot \frac{\Delta P}{L}$$

Because *Mass = Volume * Density*, we then see that *Volume = Mass/Density*. The second equation is identical to the first, except that *Mass/Density* is on the right hand side in the form of *Volume* instead of being split between left and right sides of the equals sign. Also, the variable *L* is used for length or distance instead of Δ**x** so that it isn't for a specific direction. Similarly, the absolute value of the PGF is calculated, and the direction is determined, by the location of High and Low pressure systems. Typically in class we won't calculate the PGF, only the PG. However, it'll be widely used in force balance equations.

Advection

While **advection** is included in the list of forces, it is actually not a true force. Still, advection can result in a change of wind speed in some locations. Wind moving through a point carries specific momentum, which is defined as momentum per unit mass. Recall

that momentum is mass times velocity. Specific momentum then is simply equal to the velocity or wind speed. Therefore, as wind moves by a point, the wind can move or advect variations in winds to the fixed location.

Centrifugal Force

The **centrifugal force** is an apparent force that includes the effects of inertia for winds moving along a curved path. The directionality of the centrifugal force points *outward* from the center of the curve. The centrifugal force is the opposite of the centripetal force. As we know, inertia is the physical tendency to remain unchanged. Therefore inertia causes an air parcel to "want" to move along a straight line. Turning the air parcel along a curved path requires a centripetal force that pulls inward to the center of rotation. As a result, a net imbalance of other forces occurs.

You have felt the centrifugal force many times in your life. The centrifugal force is easily felt as you travel in a moving vehicle around a corner. The force that you feel pulling you outwards is the centrifugal force.

Coriolis Force

The **Coriolis force** (CF) is another apparent force that occurs due to the rotation of Earth. The Coriolis force is a deflecting force. It acts only on objects already in motion. Therefore it cannot create wind, but it can change the wind direction by deflecting it. The Coriolis force acts perpendicular to the direction of motion, but whether the Coriolis force acts 90° to the right or left of the motion vector depends on the hemisphere on Earth. In the Northern Hemisphere, the Coriolis force acts 90° to the right of the motion vector while in the Southern Hemisphere, the force acts 90° to the left of the motion vector. The equation below gives the Coriolis force

$$CF = m \cdot 2 \cdot u \cdot \Omega \cdot \sin(\phi)$$

where m is for the mass of the object in kg, u is the speed of the object in m s^{-1}, the symbol Φ denotes latitude in degrees, and the angular rotation rate, Ω, is found from the rotation rate of Earth. Earth turns 2*pi radians over 24 hrs, so 2*pi/24 hrs gives Ω = 7.27E-5 radians·s^{-1}. Based on this, we can see that the Coriolis parameter will be 0 at the equator when $\sin(0) = 0$ and maximized at the poles when $\sin(90) = 1$.

We can also see that the Coriolis force is strongly dependent on the speed of the object. If we assume the "object" is actually wind, stronger winds will be more strongly deflected by the Coriolis force.

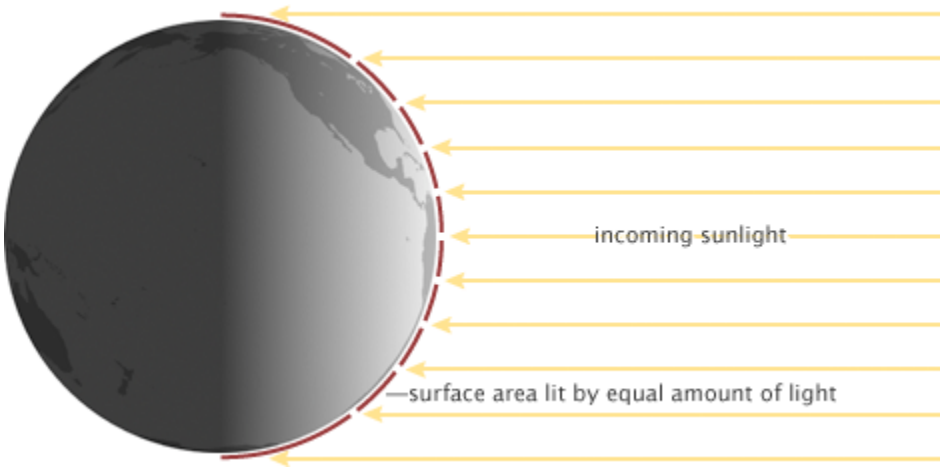
Turbulent Drag

Turbulent drag occurs when Earth's surface or objects on it cause resistance to airflow and reduce the wind speed. Any object on Earth's surface can cause drag, such as grass, trees, and buildings, which block and decelerate wind. The bottom layer of the troposphere around 0.3 to 3 km thick is called the **atmospheric boundary layer (ABL)**. Turbulence in the ABL mixes the extremely slow movement of air near the surface with the faster movement of air in the ABL and slows the wind speed in the entire ABL.

This page titled 11.1.1: Main Forces is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by LibreTexts.

11.1.2: Differential Heating

Because the Earth is round, solar radiation is not equally spread at all latitudes. Near the equator where sunlight shines directly on Earth, more solar radiation per square meter is received as compared to near the poles where sunlight shines at sharp angles to the surface (see image below). Toward Earth's poles, the same solar radiation is spread over a larger surface area such that each square meter of Earth's surface gets less radiation at the poles. As Earth rotates, the incoming solar radiation is zonally spread along latitude lines.

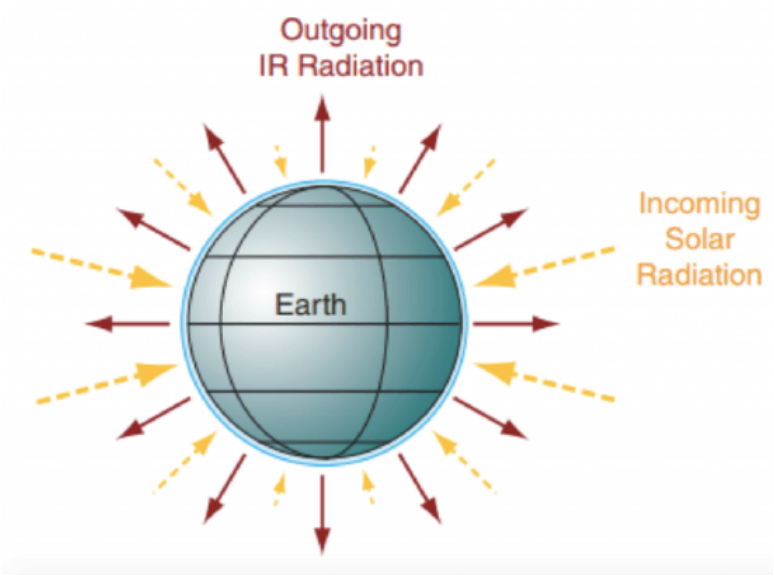


Earth's uneven heating by the sun due to the curvature of its surface NASA (Public Domain).

In this way, incoming solar radiation depends on latitude. The sun shines more directly on tropical regions at lower latitudes than at higher latitudes all year-round. Solar radiation adds heat to the Earth-atmosphere-ocean system, and thus lower latitudes get heated more than higher latitudes. This should be as expected because we know the tropics are warmer than the polar regions.

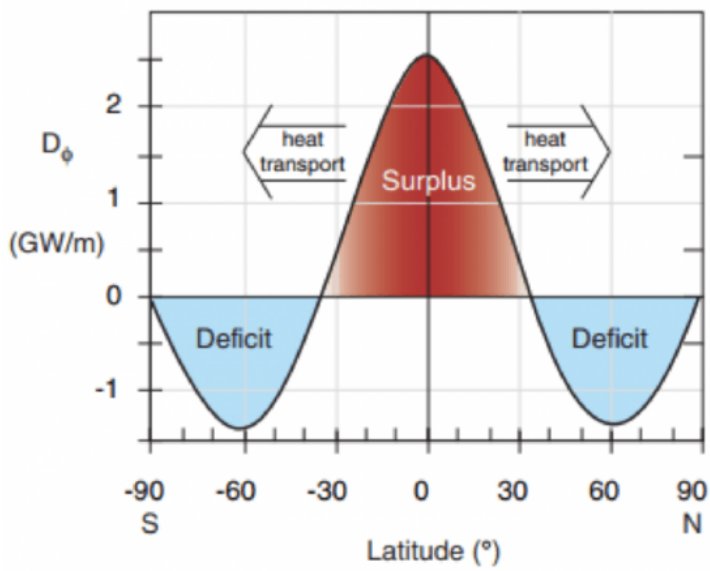
While Earth is continually heated by the sun, it is also continually losing energy by emitting outgoing longwave infrared (IR) radiation at all latitudes, and at all times, both on the light and the dark side of the globe. You'll recall from Chapter 2 that IR radiation is emitted by Earth according to the Stefan-Boltzman law, which is highly dependent on temperature.

When averaged over the globe and over long time scales, incoming UV radiation exactly balances outgoing IR radiation. But, latitude by latitude, incoming UV and outgoing do not perfectly balance. More solar energy is received by the Earth in the tropics, and while the cooling by outgoing IR radiation helps to offset this, there is still a net gain of radiative energy in the tropics. However, near Earth's poles, incoming solar radiation is less direct and too weak to offset the cooling by outgoing IR radiation, so there is net cooling at the poles. This causes warmer air at the equator, and cold air at the poles and drives Earth's atmospheric general circulation.



Incoming solar radiation (yellow dashed) and outgoing infrared radiation (red solid) diagram (CC BY-NC-SA 4.0).

The above image illustrates this more directly. Incoming solar radiation is focused near the equator, while outgoing IR radiation is relatively evenly spread across all latitudes. This results in the below energy surplus near the equator and deficit toward the poles.



Heat transport balances the net radiative imbalances (CC BY-NC-SA 4.0).

Earth's general circulation attempts to redistribute heat around the globe and rebalance the energy imbalances inherent in an unevenly heated, rotating planet. However, the general circulation cannot instantly balance global temperature, especially when the uneven heating is continuous. Therefore, a meridional temperature gradient always remains.

In an attempt to balance out Earth's incoming and outgoing energy, warm air is transported toward the poles, while cool air flows back toward the equator. This seems simple enough. However, this seemingly simple flow is complicated by many factors, including Earth's rotation, the position of continents, interactions with the oceans and many others. In order to build an understanding of this process, we'll start with some simplified models and build complexity.

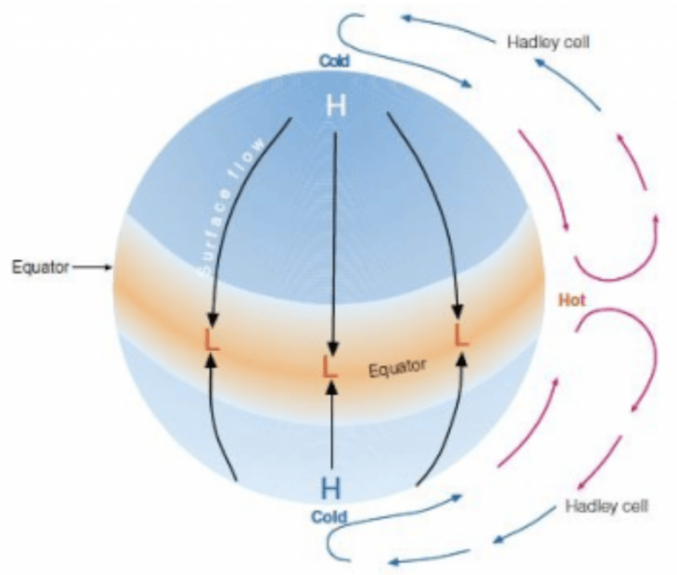
This page titled 11.1.2: Differential Heating is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by LibreTexts.

11.1.3: Single-Cell Model

The first model we'll examine is the single-cell model. With this model, we make the following assumptions.

1. The earth is entirely covered with water. This is to remove any land-sea interactions.
2. There are no seasons and the sun is always shining directly over the equator. This removes seasonal wind shifts.
3. There is no Coriolis force. While the Earth rotates to spread heat along latitudinal lines, this allows us to only be concerned with the pressure gradient force.

With these assumptions in place, Earth's global circulation would like the figure below, with one giant vertically overturning cell in each hemisphere. The excess heating at the equator is transported poleward by rising warm air, which is replaced by cold sinking polar air moving equatorward. This circulation is known as the **Hadley cell**. The Hadley cell is known as a *thermally direct* circulation because in it, warm air is rising and cold air is sinking.



The single-cell model of Hadley cells on a planet (CC BY-SA 4.0).

The circulation can be thought of in two ways. In the first, hot air at the equator rises because it is warm and buoyant. It reaches the tropopause, spreading laterally north and south at high elevations. To compensate for the rising air, surface air flows toward the equator, resulting in convergence and further uplift. Continuity of this circulation results in a global circulation with rising air at the equator and sinking air at the poles.

A second way to view global circulation is that the excess heating of air at the equator creates a large area of low pressure at the surface of the planet, while excess cooling at the poles creates high pressure at the surface. This global horizontal pressure gradient causes air to flow from high to low at the surface (pole to equator), where the air subsequently rises at the equator and flows back to the poles and sinks.

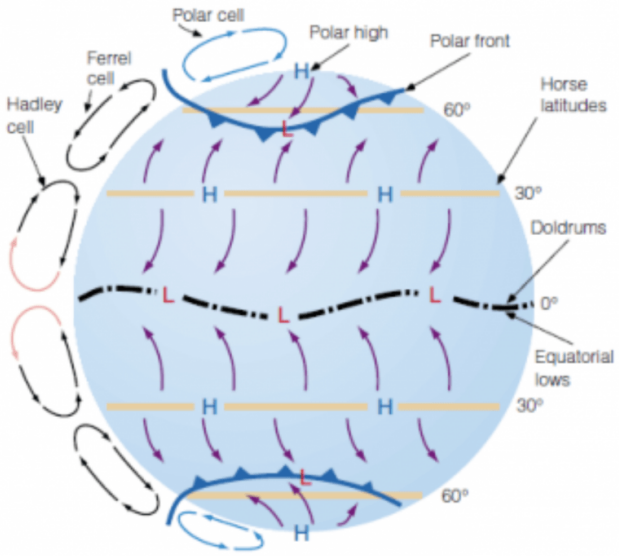
Both reasonings are plausible, its a matter of whether you focus on temperature or pressure. The temperature differences and the resulting pressure differences are intertwined and both important for the general circulation.

While this single-cell model can explain some phenomenon and works in some ways (and on some planetary bodies), it is not the reality on Earth. Earth is a rotating planet, so we need to consider the Coriolis force in addition to the pressure gradient force. In the single-cell model, as upper level air flows from the equator toward the poles, it would be deflected by the Coriolis force. In the northern hemisphere, for example, this deflection would be toward the right resulting in a wind from west to east at upper levels. In this way, the air moving from the equator to the poles would never make it there because of the rotation of Earth. A different model is needed.

This page titled 11.1.3: Single-Cell Model is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by LibreTexts.

11.1.4: Three-Cell Model

If we allow for the effects of a rotating planet, the simple single-cell model above breaks down into multiple cells in each hemisphere as shown in the figure below. It may look more complex and unrelated to the single-cell model, but there are many similarities from above. There is still excess heating in equatorial regions and excess cooling in polar regions. Instead of heat being redistributed by one massive Hadley cell from the equator to the poles, there are now three convective cells. The first of these is still the same thermally direct Hadley cell from before, but now it extends only from the equator to about 30° latitude. The poles still have a large high pressure system, while the equator has a large belt of low pressure along it. Let's take a closer look at what happens to the rising air just above the equator.



Three-cell model of the rotating Earth and the resulting wind circulations (CC BY-SA 4.0).

At the equator, the air near the surface is warm, winds are light, and the pressure gradient is weak. This region of monotonous weather is known as the **doldrums**. The warm air here rises, condensing into massive cumulonimbus clouds and thunderstorms, which release large amounts of latent heat as they form. The additional heat makes the air even more likely to rise, and provides the energy that drives the rising branch of the Hadley cell. This rising air reaches the stable tropopause, which blocks it from rising further, causing the air to diverge at upper levels and move poleward. Due to the Coriolis force, this upper level poleward flow is deflected to the right in the Northern Hemisphere and to the left in the Southern Hemisphere, providing westerlies aloft (near the tropopause) in both hemispheres in the Hadley cell.

As air moves poleward from equatorial regions, it is constantly experiencing radiational cooling as it emits infrared radiation. Simultaneously, this air begins to converge and pile up as it approaches the mid-latitudes (around 30° latitude in both hemispheres). This convergence of air far above the surface increases the mass of air aloft, increasing the pressure at the surface. This increase in surface pressure results in a belt of high pressure centers called **subtropical highs** around 30°N and 30°S. These latitudes are commonly known as the **horse latitudes**.

As this converging air above the subtropical highs slowly descends, it warms adiabatically by compression. This sinking air, dries the atmosphere creating generally clear skies and little rain. Over the oceans, weak pressure gradients in the high centers produce weak winds. Some of these lighter surface winds begin to move back toward the equator, and are deflected by the Coriolis force. This causes northeasterly winds in the Northern Hemisphere and southeasterly winds in the Southern Hemisphere in tropical regions. These winds are known as the **trade winds**, and they have a strong influence over the daily wind patterns in Hawai'i. Near the equator, the northeasterly and southeasterly trade winds converge at the surface at what is known as the **intertropical convergence zone (ITCZ)**. Here, convergence further reinforces the rising branch of the Hadley cell.

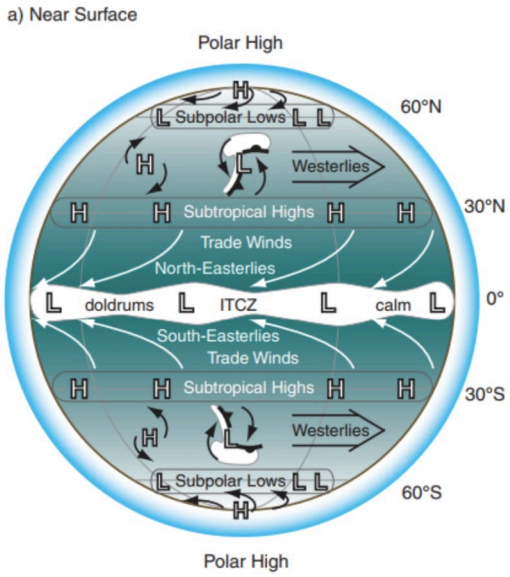
Back at 30° latitude, while some of the air sinking along the subtropical highs goes equatorward to complete the Hadley cell, some sinking air also moves poleward. This poleward moving surface air travels from from 30° to 60° and is again deflected by the Coriolis force. This results in the prevailing surface **westerlies** that impact the mid-latitudes in both hemispheres. It is for

this reason that weather moves west to east across the continental US. Often, this westerly flow is interrupted by high and low pressure systems that move with the mean surface flow. We'll learn more about this in the next two chapters. As the surface air travels poleward from 30° to 60°, it collides with cold polar air moving equatorward. These air masses do not mix easily, and are separated by a boundary known as the **polar front**. At the polar front, surface air converges and rises at the **subpolar low**, and storms and convection develop here. Some of this rising air goes all the way up to the tropopause where it moves back to 30° latitude and sinks at the subtropical high along with the descending branch of the Hadley cell. This circulation cell from 30° to 60° is known as the **Ferrel cell**, which is a *thermally indirect* circulation in which cool air rises and warm air sinks.

Behind the polar front in the Northern hemisphere, cold surface polar air moves from the poles toward 60°. As the air moves equatorward, it is again deflected by the Coriolis force. In the Arctic regions, air typically flows from the northeast while in the Antarctic, air flows from the southeast. These are known as the **polar easterlies**. Along the polar front where cold polar air collides with warm air from the Ferrel cell, some of the rising air moves back toward the poles, which gets deflected as a westerly wind aloft. Eventually this air reaches the poles, sinks back to the surface, and flows back toward the polar front, which gives us the **Polar cell**.

To summarize, looking back at the three-cell model picture: there are two major belts of high pressure and two major belts of low pressure in each hemisphere (if you include the equator in both). Areas of high pressure and sinking air exist near 30° latitude and at the poles. Regions of low pressure and rising air exist over the equator and near 60° latitude by the polar front. By knowing that winds travel counterclockwise (clockwise) around low pressure systems in the Northern Hemisphere (Southern Hemisphere), and clockwise (counterclockwise) around high pressure systems in the Northern Hemisphere (Southern Hemisphere), you can get a pretty general idea of how surface winds blow around the world on average. Trade winds blow from the subtropical highs at 30° to the equator, the westerlies blow from the subtropical highs to the polar front, and the polar easterlies blow from the poles to the polar front at the surface. Areas where these winds converge will have rising motion and low pressure at the surface, and regions where these winds diverge will have sinking motion and high pressure at the surface.

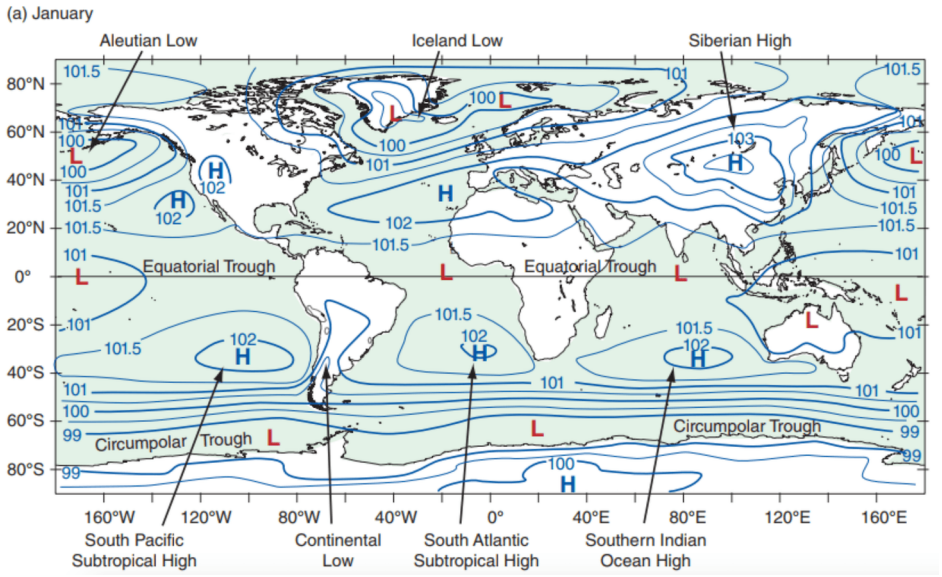
How does this three-cell model match with reality? While some minor discrepancies exist, for example in reality much of the upper-level winds in the mid-latitudes are westerly like the surface, while the Ferrel cell suggests there should be easterly winds aloft. However, this model is roughly accurate for surface winds and provides a really good first order pattern for general circulation.



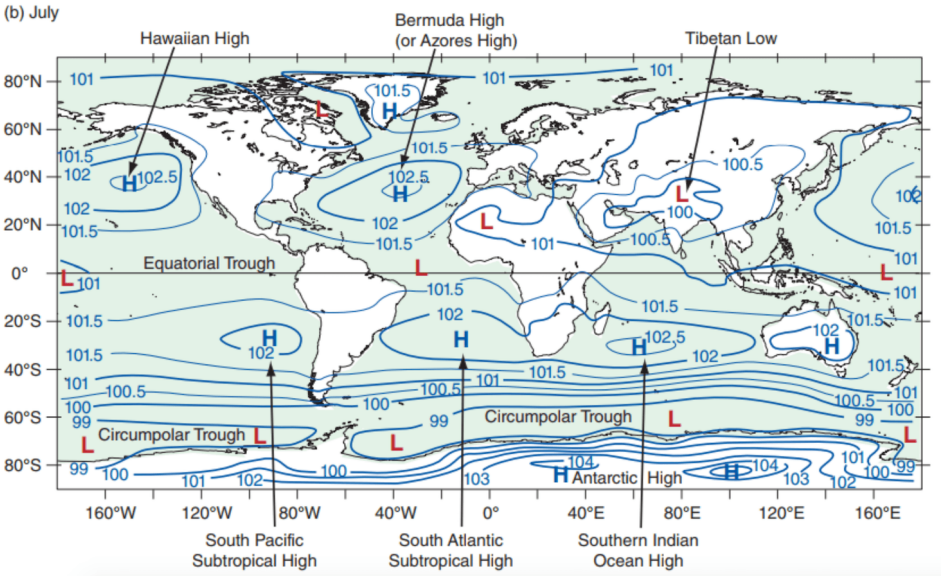
Average near surface winds (CC BY-NC-SA 4.0).

How does the real world's average surface sea level pressure field compare with the above picture? When we add in the continents, ice masses, oceans, mountains, and forest, we get an average that looks something like the below two figures. The following maps show the mean sea-level pressure field for January and July, averaged from 1981 to 2010.

Looking at the two maps below, you may notice that there are some areas where low and high pressure systems seem to persist throughout the year – these are known as **semipermanent highs** and **semipermanent lows**. These include the **Bermuda-Azores High**, the **Pacific High**, the **Icelandic Low**, and the **Aleutian Low**.



Map of semipermanent highs and lows during the month of January (CC BY-NC-SA 4.0).

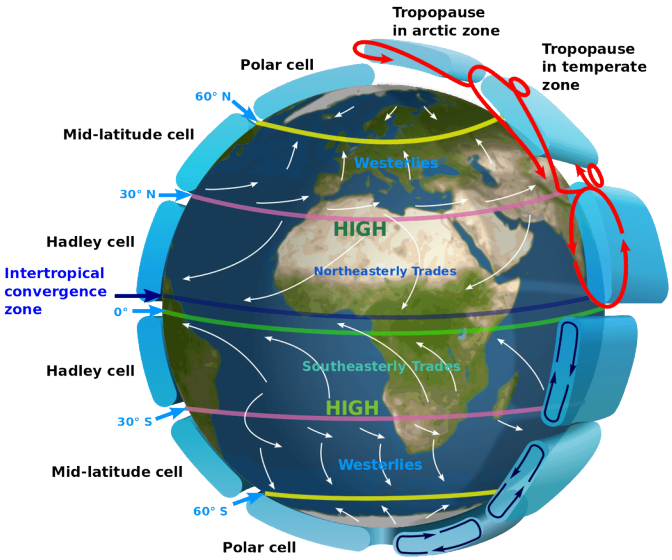


Map of semipermanent highs and lows during the month of July (CC BY-NC-SA 4.0).

This page titled 11.1.4: Three-Cell Model is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by LibreTexts.

11.1.5: Global Surface Winds

The following graphic nicely illustrates all of the above phenomenon for global surface winds. The polar easterlies, mid-latitude westerlies, and tropical trade winds are all visible. Now with the continents and land masses added, we are able to see where on Earth these surface winds are observed.

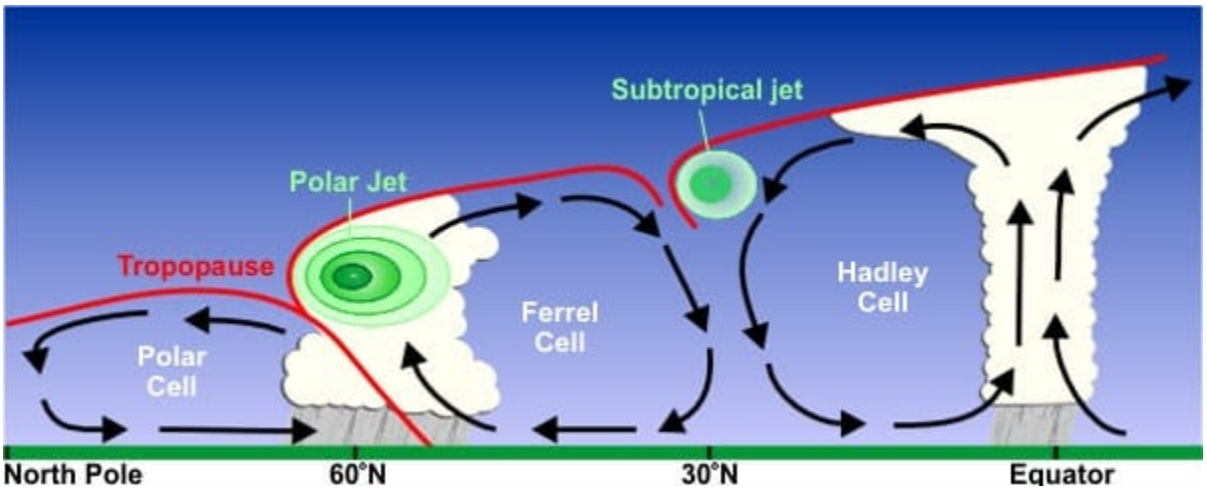


Circulation of Earth's atmosphere (CC BY-SA 3.0).

This page titled 11.1.5: Global Surface Winds is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by LibreTexts.

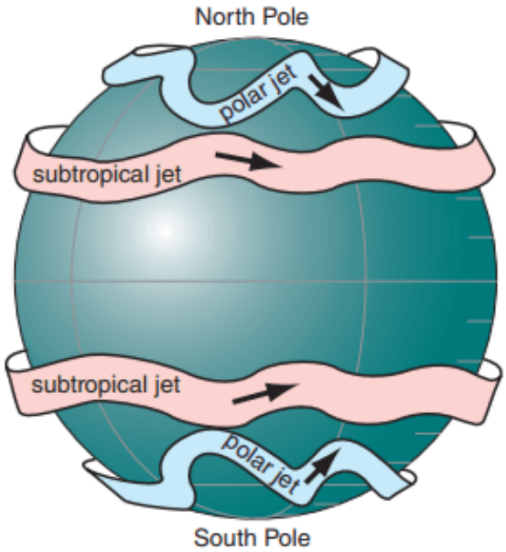
11.1.6: Jet Streams

There is one final important piece of general circulation that deserves a discussion, and that is jet streams. In the below image you can see two jet streams: the subtropical jet and the polar jet. The figure shows the average position of the jet streams in the Northern Hemisphere in the winter, as well as their relation to the tropopause. The figure shows jet streams flowing from west to east. We can see that there are two jets located right under the tropopause. The **subtropical jet stream** is located near 30° latitude about 13 km up, above the tropical high. The **polar jet stream** is located near the polar front about 10 km up, near 50° to 60° latitude. The difference in height of these jets is due to their location at the tropopause, and the fact that the tropopause is found higher in tropical regions than in polar regions due to average layer temperature differences of the troposphere underneath. The troposphere is thinner in polar regions than in tropical regions due to colder, denser air at the poles.



Cross-section of the northern hemisphere circulation, and the positions of the polar and subtropical jet streams (Public Domain).

If the general circulation of the atmosphere is like a giant meandering river of air around the globe, then **jet streams** are swiftly flowing currents within that river. Jet streams are thousands of kilometers in length, and hundreds of kilometers in width. In the core of a jet stream (called a **jet streak**), wind speeds are often higher than 100 knots and are occasionally higher than 200 knots. The polar jet can sometimes merge with the subtropical jet if it sweeps southward enough, and it occasionally splits into two jet streams.



The position of the polar jet stream and the subtropical jet stream (CC BY-NC-SA 4.0).

The existence of jet streams is ultimately due to the energy imbalance between tropical and polar regions. How exactly do they form? As mentioned before, the polar front is a boundary between colder polar air and warmer subtropical air. Because of this, the strongest temperature gradient occurs along the polar frontal zone. This rapid change of temperature with distance also causes a rapid pressure change, due to the thermal wind effect (a vertical shear in the geostrophic wind caused by a horizontal temperature gradient). This strong pressure gradient across the polar front causes intense wind speeds that become the jet stream. The temperature contrast between north and south along the polar front is more intense during the winter than during the summer, so the polar jet is also stronger during the winter. During winter, the leading edge of the cold polar air pushes further south into subtropical areas. During the summer, the polar front retreats into higher latitudes and is weakened.

The subtropical jet stream tends to form just above the descending branch of the Hadley cell, at about 12 km altitude. Here, a boundary exists between warmer equatorial air and cooler air that has been cycled up and around the Ferrel cell from the polar front. This is sometimes referred to as the *subtropical front*, but it does not extend all the way to the surface. Here, the temperature gradient is strongest aloft near the tropopause, which induces a sharp pressure gradient and strong winds aloft as well.

One final thought for Chapter 11: isn't it fascinating that all of these global winds are caused by differential heating and the rotation of Earth? Everything is connected in one way or another and fits together like a puzzle.

Chapter 11: Questions to Consider

Query 11.1.6.1

1. *This interactive element is not available in this format. Visit the link or scan the QR code to access it online.*

   View this Interactive Content online



2. What assumptions are made for a single-cell model of Earth's atmosphere?
3. Drag the terms to their correct position:

Query 11.1.6.2

This interactive element is not available in this format. Visit the link or scan the QR code to access it online.

View this Interactive Content online



4. Where are jet streams located? Why do they differ in height?

Selected Practice Question Answers:

Query 11.1.6.3

This interactive element is not available in this format. Visit the link or scan the QR code to access it online.

View this Interactive Content online



This page titled 11.1.6: Jet Streams is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by LibreTexts.

11.2: Air Masses and Fronts

This section examines large bodies of air and the boundaries that form where different air masses meet. Air masses acquire characteristic temperature and moisture properties from their source regions and can transport those conditions over large distances. When contrasting air masses interact, fronts develop, producing changes in temperature, wind, clouds, and precipitation that are important for understanding everyday weather and larger weather systems.

This page titled 11.2: Air Masses and Fronts is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Robert Wagner.

11.2.1: Air Masses

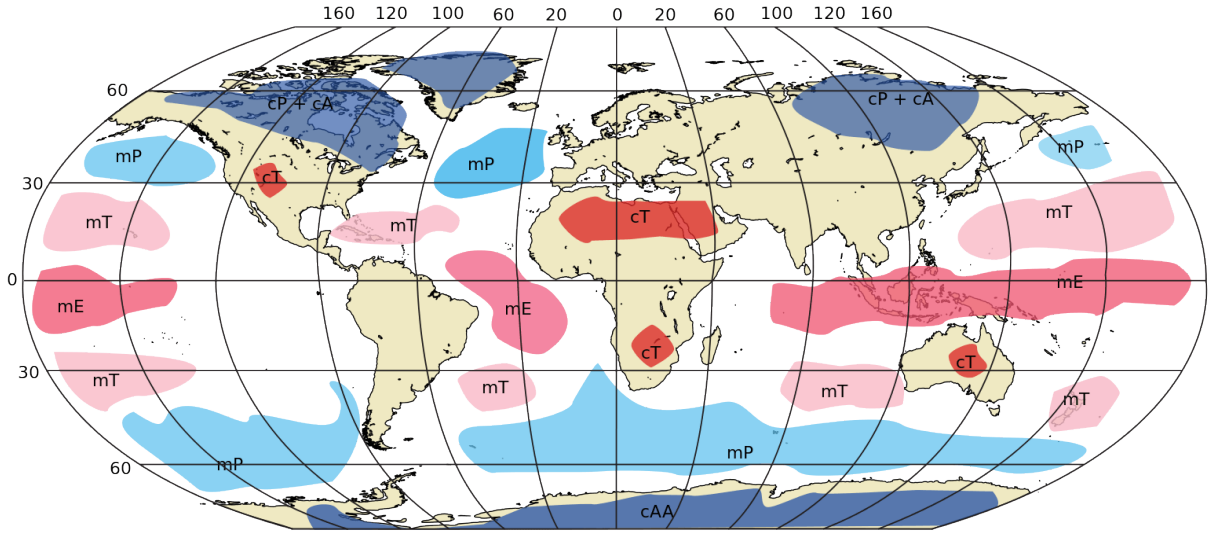
An air mass is an extensive body of air featuring generally similar temperature and moisture characteristics. They can extend thousands of square kilometers. Air mass are identified based on their temperature and humidity characteristics as well as their geographical region of origin. For example, if an air mass is dry and warm and originated from the tropical region over a continent, it would be called a continental Tropical (cT) air mass. If an air mass is humid and cold and originated over the ocean in the high latitudes, it would be called a maritime Polar (mP) air mass. The air mass name will always begin with a lower-case letter signifying either continental (c) or maritime (m) and a second capital letter signifying Equatorial (E), Tropical (T), Polar (P), Arctic (A), or Antarctic (AA).

Maritime air masses are humid air masses originating from oceans or large bodies of water. Continental air masses are dry air masses originating from land. Equatorial air masses are warm moist air masses originating from the equatorial region. Tropical air masses are warm air masses originating from the lower latitudes. Polar air masses are cold air masses originating from the upper latitudes. Arctic air masses are composed of extremely cold air that originated from the poles. Arctic air masses are even colder and drier than Polar air masses, and Antarctic air masses are even more extreme than Arctic air masses.

The common types of air masses are maritime Tropical (mT), maritime Polar (mP), continental Polar (cP), continental Tropical (cT), and continental Arctic (cA).

- Maritime Tropical (mT) air masses are warm, humid air masses originating from the oceans in the tropics.
- Maritime Polar (mP) air masses are cold, humid air masses originating from the oceans in the polar latitudes.
- Continental Polar (cP) air masses are cold, dry air masses originating from land regions in the polar latitudes.
- Continental Tropical (cT) air masses are hot, dry air masses originating from land in the tropics.
- Continental Arctic (cA) air masses are cold, dry air masses originating from the North Pole.
- Continental Antarctic (cAA) air masses are extremely cold and dry air masses originating from land at the South Pole.

See the below image to visualize where these different types of air masses typically originate.



Air masses and their location of origin (Public Domain).

Creation of Air Masses

Air masses develop when air is present over a surface for an *extended period of time*. This typically occurs in a high pressure system with light wind. The areas where air masses develop are called source regions. Air masses over warm surfaces usually develop faster than those over colder surfaces because there is weaker turbulence in the stable air over the cold surface. When air masses shift from their source regions, they change over time due to the surfaces and terrain over which the air masses flow.

Movement of Air Masses

Air masses do not remain over their source regions permanently. Slight changes in weather patterns may shift the air mass to a new location. As air masses move, two things can occur. First, as the air shifts over the different surface characteristics, the air mass begins changing. This process is called **air mass modification**. For example, an mP airmass that moves from the Pacific ocean over the mountains in the western continental US will typically dry as it crosses the mountains, rains out its moisture, and warms over the land surface until it becomes a cT airmass. The second thing that happens when air masses move is that they may collide with other air masses. When a collision occurs, the two air masses develop a boundary called a **front**.

This page titled 11.2.1: Air Masses is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by LibreTexts.

11.2.2: Surface Fronts

Surface fronts are the boundaries or transition zones between air masses at the Earth's surface. Changes in temperature, humidity, wind, pressure, visibility, as well as particular cloud and precipitation patterns are often observed at fronts. There are four main types of fronts.

1. Cold fronts
2. Warm fronts
3. Occluded fronts
4. Stationary fronts

Fronts are named based on the characteristics of the air mass that is replacing the prior air mass. For example, if a cold air mass is moving toward a warm air mass, the boundary between them will be called a cold front because the cold air is effectively replacing the warm air from the perspective of a stationary point on Earth's surface.

Fronts are usually associated with low pressure systems. Frontal boundaries on a map are labeled in the location where the temperature gradient of the front meets the Earth's surface.

Cold Fronts

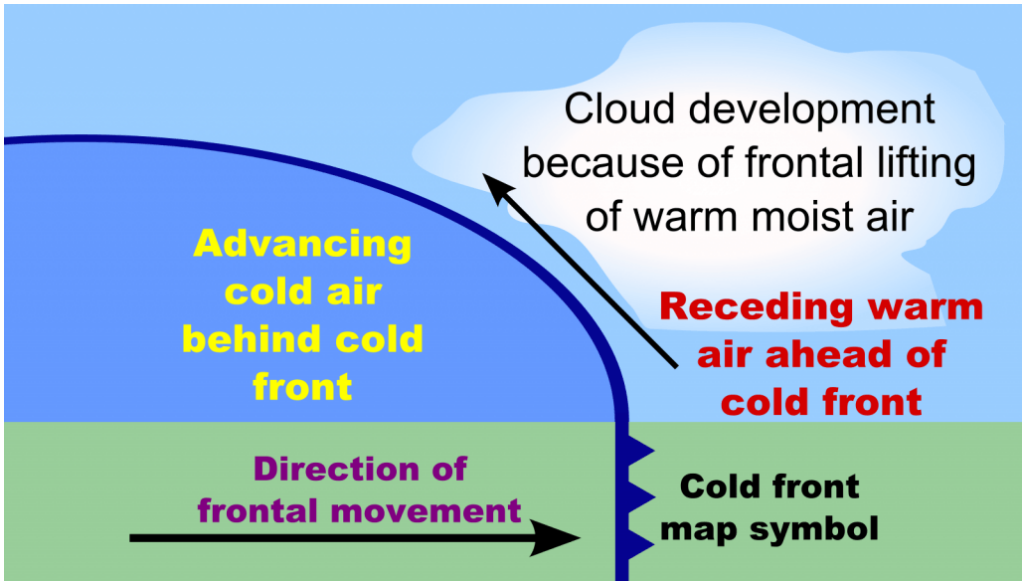


Diagram showing a vertical cross section through a cold front (CC BY-SA 4.0).

Cold fronts are a transition zone in which the advancing cold air mass replaces a retreating warmer air mass. On weather maps, cold fronts are drawn as blue lines with blue triangles pointing toward the warmer air mass, in the direction of frontal movement, as seen in the above image.

Because warm air is less dense than colder air, the cold air stays on the bottom and warm air is forced to rise above the advancing cold air. This forced lifting results in typical cloud patterns ahead of a cold front, which include cirrus and cirrostratus clouds. The number of cumulus clouds in the warm air mass increases as the frontal boundary approaches. Because the warm air mass is forced to rise, atmospheric instability occurs along the cold front and results in towering cumulus and cumulonimbus clouds, which may produce heavy rain and thunderstorms along the frontal boundary. During the passage of a cold front, the wind direction generally shifts from south or southwest (in the warm sector) to west or northwest (in the cold sector) in the northern hemisphere. After a cold front's passage, fair weather returns with the appearance of cumulus and stratocumulus clouds.

Warm Fronts

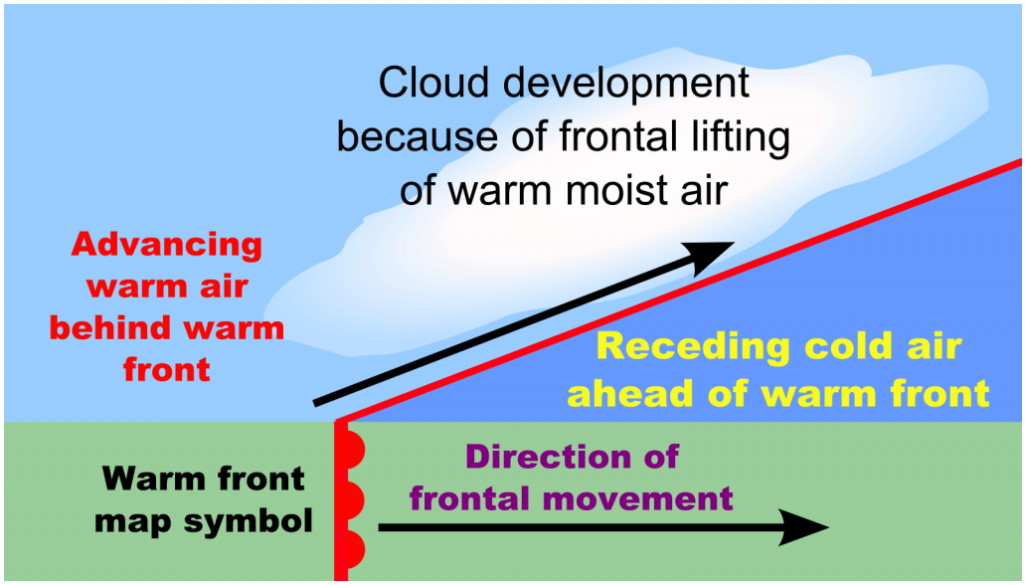


Diagram showing a vertical cross section through a warm front (CC BY-SA 3.0).

Warm fronts are a transition zone in which the advancing warm air mass replaces a retreating colder air mass. On weather maps, warm fronts are drawn as red lines with red semicircles pointing toward the colder air mass in the direction of the frontal movement.

Advancing warm air is forced to rise above the retreating cold dense air. Again, because of this forced lifting, typical cloud patterns are common ahead of a warm front. These include upper-level clouds such as cirrus and cirrostratus clouds before clouds thicken and lower-level clouds like altostratus, nimbostratus, and fog near the frontal boundary. Because the air mass is rising along the warm front, clouds form and steady precipitation may occur. During the passage of a warm front, wind direction shifts from east or southeast (in the cold sector) to southwest (in the warm sector) in the northern hemisphere. After a warm front's passage, cloud cover and precipitation decreases with only scattered cumulus clouds remaining.

Occluded Fronts

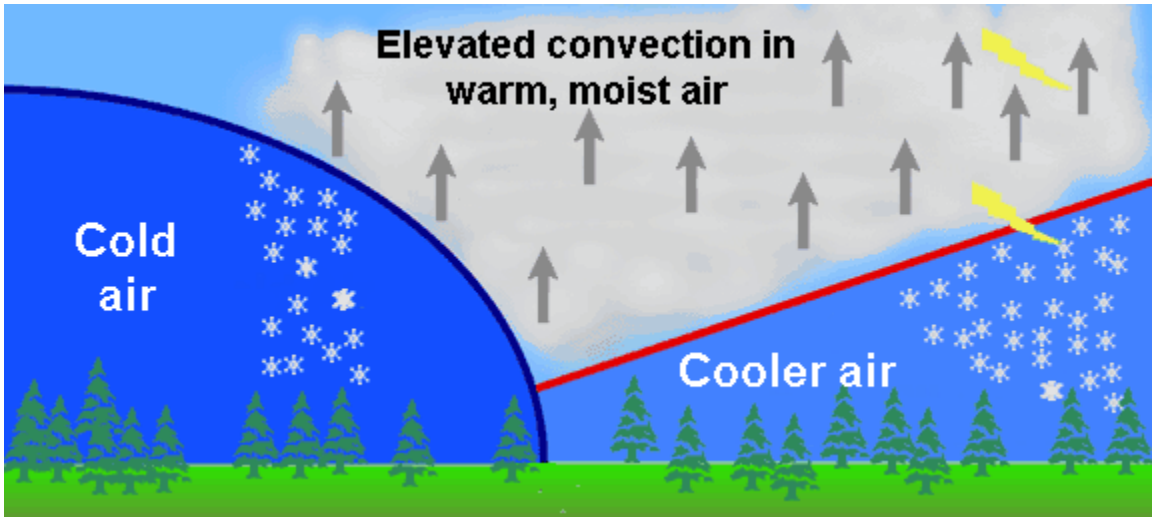


Diagram showing a vertical cross section through an occluded front (Public Domain).

Occluded fronts are a frontal boundary that forms when a cold front catches up to a warm front. Cold fronts move faster than warm fronts, so cold fronts can sometimes catch up to warm fronts, but not the other way around.

There are two types of occlusions: cold occlusions and warm occlusions. Cold occlusions (shown above) occur when the advancing

air mass is colder than the retreating air mass. Warm occlusions occur when the advancing air mass is warmer than the retreating air mass. On weather maps, occluded fronts are drawn as purple lines with purple triangles and purple semi circles. You'll notice that this symbol is a combination of the cold front and warm front symbols, which isn't surprising because an occluded front is effectively a combination of the two.



Frontal symbol for an occluded frontal boundary (Public Domain).

Occluded fronts are the final stage of a frontal boundary because warm air above cool air is a stable scenario.

Stationary Fronts

Stationary fronts are a type of frontal system that are almost stationary with the winds flowing nearly parallel and from the opposite paths in each side separated by the front. On weather maps, stationary fronts are drawn as alternating blue and red lines with blue triangles pointing toward the warmer air mass and red semicircles pointing toward the colder air mass. This is the only scenario where the direction of the symbols does not indicate a direction of movement.



Frontal symbol for a stationary frontal boundary (Public Domain).

Other Frontal-like Features

Not all weather features have frontal characteristics, such as a trough and a dryline. While troughs feature a change in cloud and precipitation pattern, it lacks the sharp change in temperature and moisture as observed in fronts. Drylines are a boundary between warm, moist air and warm, dry air. Because both air masses are warm, a dryline cannot be classified as either a warm front or a cold front. Drylines commonly occur during spring and summer in southwestern United States, particularly in Texas. Warm, humid air from the Gulf of Mexico meets with the warm, dry air from the desert plateau. During the afternoon, convective clouds and even thunderstorms can develop in drylines as moist air rises over the denser dry air.

Chapter 12: Questions to Consider

1. What is an air mass? What are some different types of air masses and where do they originate?
2. Label the cold front cross-section:

Query 11.2.2.1

This interactive element is not available in this format. Visit the link or scan the QR code to access it online.

View this Interactive Content online

3. Label the warm front cross-section:

Query 11.2.2.2

This interactive element is not available in this format. Visit the link or scan the QR code to access it online.

View this Interactive Content online

4. Match the front name with the correct symbology:

Query 11.2.2.3

This interactive element is not available in this format. Visit the link or scan the QR code to access it online.

View this Interactive Content online

5. What cloud and precipitation patterns are typically associated with warm and cold frontal passage?

Selected Practice Question Answers:

Query 11.2.2.4

This interactive element is not available in this format. Visit the link or scan the QR code to access it online.

View this Interactive Content online

This page titled 11.2.2: Surface Fronts is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by LibreTexts.

12: WEATHER SYSTEMS AND SEVERE WEATHER

12.1: Mid-latitude Frontal Cyclones

12.2: Tropical Cyclones

12.2.1: Tropical Cyclone Structure

12.2.2: Intensity and Geographic Distribution

12.2.3: Evolution

12.2.4: Hazards

12.3: Thunderstorms and Severe Weather

12.3.1: Thunderstorms

12.3.2: Atmospheric Instability and Thunderstorms

12.3.3: Precipitation and Hail

12.3.4: Lightning

12.3.5: Tornadoes

12.3.6: Keeping Yourself Safe

Chapter 12: Weather Systems and Severe Weather — Introduction

Weather is the result of many atmospheric processes acting together. Differences in temperature and pressure create winds, moisture provides the water needed for clouds and precipitation, and atmospheric instability allows air to rise and storms to develop. In this chapter, we bring these ideas together to examine some of the major weather systems that affect people around the world.

We begin with **mid-latitude cyclones**, the large weather systems responsible for many of the changing conditions experienced across the middle latitudes. These systems develop where contrasting air masses interact and are closely associated with fronts, changes in atmospheric pressure, and patterns in the upper atmosphere. We then turn to **tropical cyclones**, including hurricanes, and examine how they form, develop, intensify, and create hazards.

Finally, we examine **thunderstorms and severe weather**. Thunderstorms provide a particularly clear example of how moisture, instability, and atmospheric motion interact. Some storms remain relatively ordinary, while others produce hail, damaging winds, lightning, or tornadoes. Understanding the atmospheric conditions that produce these events also helps us understand why some weather systems become hazardous.

This chapter therefore serves as a culmination of the meteorology portion of the course. Rather than considering temperature, moisture, pressure, winds, clouds, and air masses separately, we can now see how they interact to produce the weather systems we observe and experience.

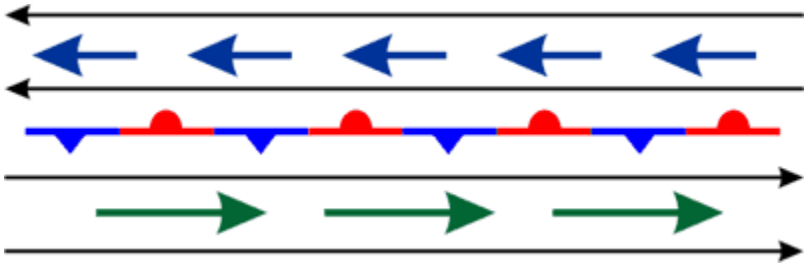
This page titled 12: Weather Systems and Severe Weather is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Robert Wagner.

12.1: Mid-latitude Frontal Cyclones

Following the Norwegian model, the development of a mid-latitude cyclone begins along the polar front. Recall from Chapter 11 that the polar front separates cold polar air from warmer subtropical air at around 60° latitude. It is a semi-continuous boundary and mid-latitude cyclones form and move along it as a series of waves. For this reason, a developing storm is sometimes referred to as a wave cyclone.

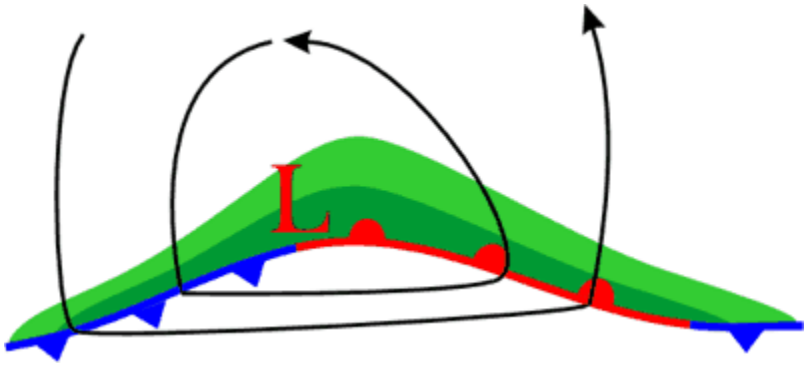
Cyclogenesis and Life Cycle

This figure below shows a portion of the polar front as a stationary front, with cold air to the north and warmer air to the south flowing parallel to the front in opposite directions. These winds moving in opposite directions set up rotation, similar to how a pen will turn if you place it between your hands and move them in opposite directions.



Part 1 of cyclogenesis: a stationary front with opposite moving winds on either side (Public Domain).

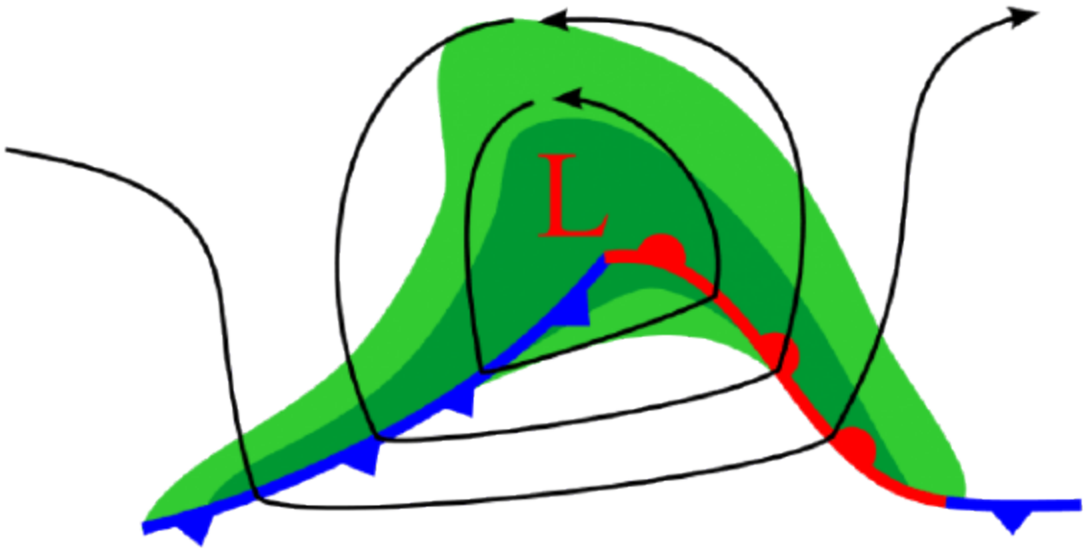
Under the right conditions, a **frontal wave** will begin to form along the front, with a cold front pushing southward and a warm front moving northward. The lowest pressure lies at the junction of the two fronts. The southward-moving cold front pushes warmer, less dense air upward, while the warm front overruns and moves over the colder air ahead of it. This creates rising motion in the column, and a narrow band of precipitation forms. The surface low pressure system is steered by winds aloft, typically moving eastward or northeastward, and it gradually becomes a fully-developed mature cyclone 12 to 24 hours after its incipient stage.



Part 2 of cyclogenesis: the formation of a frontal wave (Public Domain).

The central pressure lowers and the pressure gradient increases, causing a stronger cyclonic (counterclockwise) flow inward toward the low's center. The precipitation band widens ahead of the warm front, and narrows ahead of the cold front. The region of warmer air between the cold and warm fronts here is called the **warm sector**. Here, the weather is generally partly cloudy, with scattered showers possible if the air is conditionally unstable.

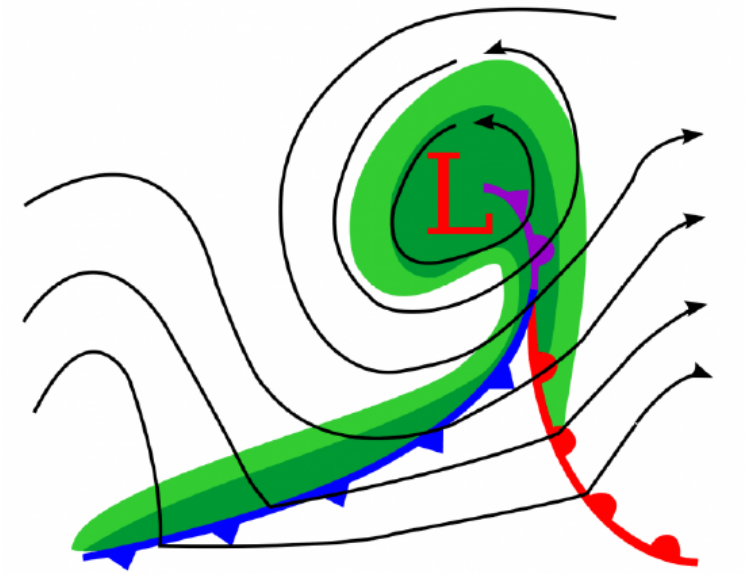
Where does the mid-latitude cyclone get its energy from? As the warmer and colder air masses attempt to regain equilibrium, warm air rises over the colder air, which transforms potential energy into kinetic (motion) energy. As warmer air rises, it condenses into clouds, which release latent heat energy into the system. Near the surface, winds converge inward toward the low's center. Wind speeds may increase as a result of a stronger pressure gradient force near the center, increasing the kinetic energy in the system.



Part 3 of cyclogenesis: a newly developed cyclone (Public Domain).

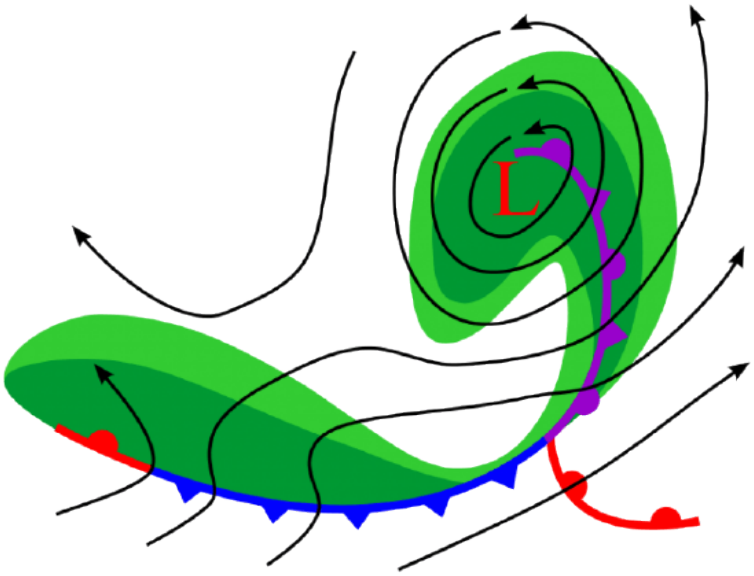
As the cyclone moves eastward, the central pressure continues to decrease and winds increase during its mature stage. The faster-moving cold front closes in on the warm front, decreasing the size of the warm sector. In the Norwegian cyclone model, the cold front overtakes the warm front and the cyclone becomes occluded. Cloud cover and precipitation cover a wide area and the storm is usually most intense at this stage.

The point where the cold front, warm front, and occluded front intersect is called the *triple-point*. Occasionally, a secondary low may form at this triple point, move eastward, and intensify into another cyclone.



Part 4 of cyclogenesis: formation of an occluded front and triple point (Public Domain).

Eventually, as occlusion advances, the low pressure center will begin to dissipate, because cold air exists on both sides of the occluded front. The sector of warm, rising air is removed from the center of the storm, so the storm gets cut off from its primary energy supply. Eventually the old storm dies out and gradually disappears.



Part 5 of cyclogenesis: continued formation of an occluded front and dying out of the cyclone (Public Domain).

This sequence of a developing mid-latitude cyclone is similar to a whirling, spinning eddy in a river that forms behind a stick or log, moves along with the river, and quickly disappears further downstream. This entire life cycle can last from several days to a little more than a week.

Cyclones in various stages of development can be seen all at once along the polar front—this succession of storms is known as a cyclone "family". As mentioned before, some cyclones form from dying previous cyclones and become a part of the succession.

This polar front model of development for a mid-latitude cyclone is rather simplified and, in fact, very few storms follow this model exactly. However, it is a good foundation for understanding storm structure.

Storm Tracks

The development of a mid-latitude cyclone is a process called **cyclogenesis**. Certain regions in North America are more favorable for cyclogenesis, including the eastern slopes of mountain ranges like the Rockies and Sierra Nevada, the Atlantic Ocean off the Carolina Coast, and the Gulf of Mexico. When air flows westward across a north-south extending mountain range, the air on the leeward (downwind) side tends to have cyclonic curvature, which adds to the development of a cyclone. This is called **lee cyclogenesis**, and cyclones that are a result of this are often called **lee-side lows/cyclones**.

Cyclones may also develop near Cape Hatteras, North Carolina, where warm moist air from the Gulf Stream can increase the north-south air mass temperature/moisture contrast to the point where cyclogenesis might occur. These cyclones are called **northeasters** (or nor'easters) and normally move northeast along the Atlantic Coast. These storms can bring heavy rain or snow and high winds to areas along the East Coast.

Typical cyclone storm tracks are named after the region in which they form, like the *Hatteras low*, *Alberta Clipper*, or *Colorado low*. Alberta clippers and Colorado lows form or re-develop on the lee-side of the Rockies. Mid-latitude cyclones always move toward the east due to the prevailing westerlies.

What influences the strength of a mid-latitude cyclone, and determines how long it will persist? There are some surface conditions that influence cyclogenesis, but the real key to mid-latitude cyclone development lies in the winds aloft. How are mid-latitude cyclones influenced by upper-level flow?

Mid-latitude Cyclone in Three Dimensions

Developing surface lows are usually more intense with height and appear on upper-level charts as a trough or a closed low. However, the low in the upper-levels usually exists to the west of the surface low (again, in the Northern Hemisphere). This is a necessary condition for a low pressure system to continue to develop and intensify. If the upper-level low were directly over

the surface low, the surface low would quickly dissipate. This is because winds converge inward toward the low, but only at the surface. This convergence at the surface causes the air mass to “pile up” and air density to increase just above the surface low. The increase in air mass causes surface pressures to rise, and the low fills in and dissipates. How then do cyclones intensify and develop? The air that piles up at the surface must have an “exit path” out of the column so that the surface air pressure can continue to decrease and the cyclone can strengthen. The vertical structure of the atmosphere must allow for air to rise out of a surface low pressure.

The following figure shows an idealized model of the vertical structure of a cyclone and anticyclone in the Northern Hemisphere. A surface low and a surface high are accompanied by an upper level trough and ridge respectively.

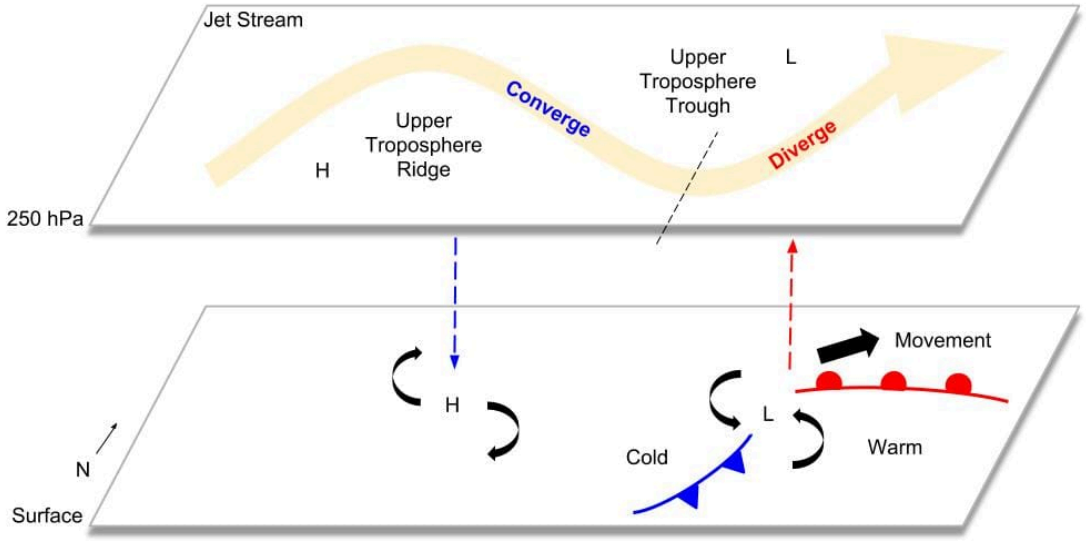


Diagram of the structure of a cyclone and anticyclone accompanied by a trough and ridge (Image Created by Britt Seifert).

On the right hand side is a Northern Hemisphere frontal cyclone with a warm and cold front. The cold air behind the cold front at the surface also extends upward aloft. Recall that the layer between two pressure surfaces is thinner when the air temperature of the layer is cold (more dense), and thicker when the layer is warm (less dense). When pressure levels are packed closer together, pressure decreases more rapidly with height in a column of cold air. Because of this, low pressure is found aloft a body of cold air, just as you find behind a cold front because the constant pressure surfaces are squeezed closer to the Earth’s surface. Thus an upper low is often found in the cold air aloft to the west of, or behind, the surface low. The surface low tilts toward the northwest moving up from the surface. Directly above the surface low, the airflow spreads out and diverges. This allows the converging surface air to rise and flow out of the air column at the tropopause, reinforcing vertical motion. When the divergence in the upper levels is stronger than convergence at the surface, surface pressures will lower further, and the low will intensify and deepen. Put another way: when air exits the column more rapidly aloft than it enters the column at the surface, then the amount of air in the column will reduce, the surface pressure will lower, and the cyclone will intensify.

Notice that there is convergence directly aloft of the high pressure system. Here, the air mass increases aloft and piles up, while air flows clockwise (in the Northern Hemisphere) and out of the anticyclone at the surface. If air were able to flow freely out of the anticyclone, the air pressure would rapidly drop and the anticyclone would dissipate. Therefore, to maintain or strengthen the high pressure system, air has to continually be added to the anticyclone. This happens when there is convergence above a surface high. The air that piles up aloft sinks in the column increasing surface pressure. If the convergence aloft is stronger than the divergence at the surface (more air is added than is removed), then the surface pressure will increase.

Winds at the 500-mb pressure level tend to steer surface low and high pressure systems. Generally speaking, surface storm systems tend to travel at about 16 knots in summer, and roughly 27 knots in winter. This is due to stronger jets and upper-level flow in the winter, a result of stronger north-south temperature differences.

Redistribution of Heat

Mid-latitude frontal cyclones are both a vital part of global circulation and a result of global circulation. They’re also an important pattern in the climatology of regions in the mid-latitudes.

The temperature gradients that cause frontal cyclones form as a result of the colliding surface air from the polar and Ferrel cells. The strong temperature gradient with cold air from the polar region and warm air from the tropics is the energy source that drives the frontal storms. On the flip side of the same token, frontal cyclones are a huge contributor to the redistribution of heat globally. Warm air moving poleward in a warm front and cold air moving equatorward in a cold front are later stabilized after the temperature gradient balances itself by forcing the cold air aloft (occluded front). Frontal cyclones are thus both a result of and a contributor to global circulation and the redistribution of heat from the equator to the poles.

The beginning of the chapter showed images of these storm systems extending over scales the size of continents and land-masses. It also described how these storm systems last from days to over a week. Instabilities along the polar front are always growing and dying, and passing over fixed points on Earth. The point is that if you live somewhere along the storm track in the Northern or Southern hemisphere, in the wintertime, these storm systems dictate your weather. For example, if you live in Boston, Massachusetts, your winter weather may look something like this: a few days of warm clear weather, a quick change (passing cold front) resulting in a large drop in temperatures and some heavy rain, then cold dry weather for a few days. This cold weather then transitions slowly to warm by some light rain and warming temperatures (warm front). While variable, this pattern repeats itself week after week. Depending on the stage of the frontal storm as it passes over you, it may be more or less severe, and you may receive more or less rain, snow, or other wintery weather. The precipitation and temperature variations resulting from frontal cyclones are an important part of the climatology of mid-latitude weather.

Chapter 13: Questions to Consider

1. Which front do mid-latitude cyclones form and move along?
2. How long does it take for a cyclone to fully develop?
3. What is the point where the cold front, warm front, and occluded front intersect called?
4. True or False:

Query 12.1.1

This interactive element is not available in this format. Visit the link or scan the QR code to access it online.

View this Interactive Content online

5. Put the steps of cyclogenesis in the correct order from 1 to 5:

Query 12.1.2

This interactive element is not available in this format. Visit the link or scan the QR code to access it online.

View this Interactive Content online

Selected Practice Question Answers:

Query 12.1.3

This interactive element is not available in this format. Visit the link or scan the QR code to access it online.

View this Interactive Content online

This page titled 12.1: Mid-latitude Frontal Cyclones is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by LibreTexts.

12.2: Tropical Cyclones

Intense synoptic-scale cyclones in the tropics are called **tropical cyclones**. As for all cyclones, tropical cyclones have low pressure in the cyclone center near sea level. Low-altitude winds also rotate cyclonically (counterclockwise in the N. Hemisphere) around these storms and spiral in towards their centers.

Tropical cyclones are called **hurricanes** over the Atlantic and eastern Pacific Oceans, the Caribbean Sea, and the Gulf of Mexico (Fig. 16.1). They are called **typhoons** over the western Pacific. Over the Indian Ocean and near Australia they are called **cyclones**. In this chapter we use “tropical cyclone” to refer to such storms anywhere in the world.

Comparing tropical and extratropical cyclones, tropical cyclones do not have fronts while mid-latitude cyclones do. Also, tropical cyclones have warm cores while mid-latitude cyclones have cold cores. Tropical cyclones can persist two to three times longer than typical mid-latitude cyclones. To help explain this behavior, we start by describing tropical cyclone structure.

Screen Shot 2020-03-24 at 11.42.32 PM.png

Figure 16.1 Visible-spectrum satellite picture of Hurricane Katrina over the Gulf of Mexico, taken 28 Aug 2005 at 1545 UTC. (GOES image courtesy of US DOC/NOAA.)

Thumbnail: This low-pressure system over Iceland spins counterclockwise due to balance between the Coriolis force and the pressure gradient force. (Public Domain; NASA’s Aqua/MODIS satellite)

This page titled 12.2: Tropical Cyclones is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Roland Stull via source content that was edited to the style and standards of the LibreTexts platform.

- 16: Tropical Cyclones by Roland Stull is licensed CC BY-NC-SA 4.0. Original source: https://www.eoas.ubc.ca/books/Practical_Meteorology.

12.2.1: Tropical Cyclone Structure

Tropical cyclones are made of thunderstorms. Near the center (**core**) of the tropical cyclone is a ring or circle of thunderstorms called the **eyewall**. This is the most violent part of the storm with the heaviest rain and the greatest radar reflectivity (Fig. 16.2). The tops of these thunderstorms can be in the lower stratosphere: 15 to 18 km high. Thunderstorm bases are very low: in the boundary layer. Thus, Tropical cyclones span the tropical-troposphere depth.

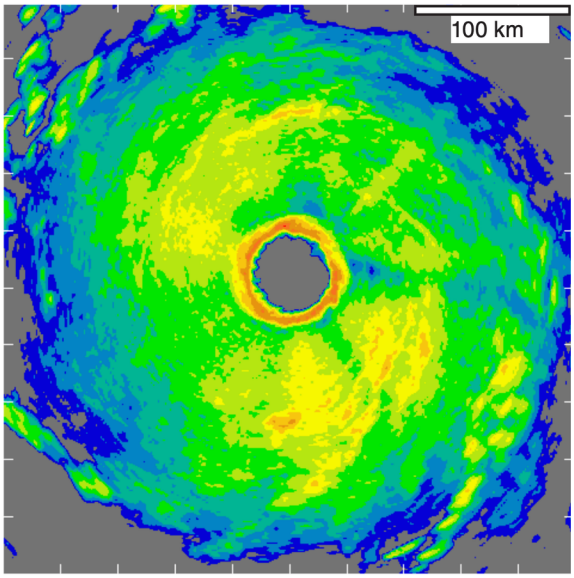


Figure 16.2 Radar reflectivity PPI image of Hurricane Katrina, taken from a research aircraft at altitude 2325 m on 28 Aug 2005 at 1747 UTC while it was over the Gulf of Mexico. Dark orange regions in the eyewall correspond to 35 to 40 dBZ radar reflectivity. (Image courtesy of US DOC/NOAA/AOML/ Hurricane Research Division.)

The anvils from each of the thunderstorms in the eyewall merge into one large roughly-circular **cloud shield** that is visible by satellite (Fig. 16.1). These anvils spread outward 75 to 150 km away from the eye wall. Hence, tropical cyclone diameters are roughly 10 to 20 times their depth (Fig. 16.3), although the high-altitude outflow from the top of some asymmetric tropical cyclones can reach 1000s of km (Fig. 16.4).

In the middle of the eyewall is a calmer region called the **eye** with warm temperatures, subsiding (sinking) air, and fewer or no clouds. Eye diameter at sea level is 20 to 50 km. The eye is conical, with the larger diameter at the storm top (Fig. 16.5).

Spiraling out from the eye wall can be zero or more bands of thunderstorms called **spiral bands**.

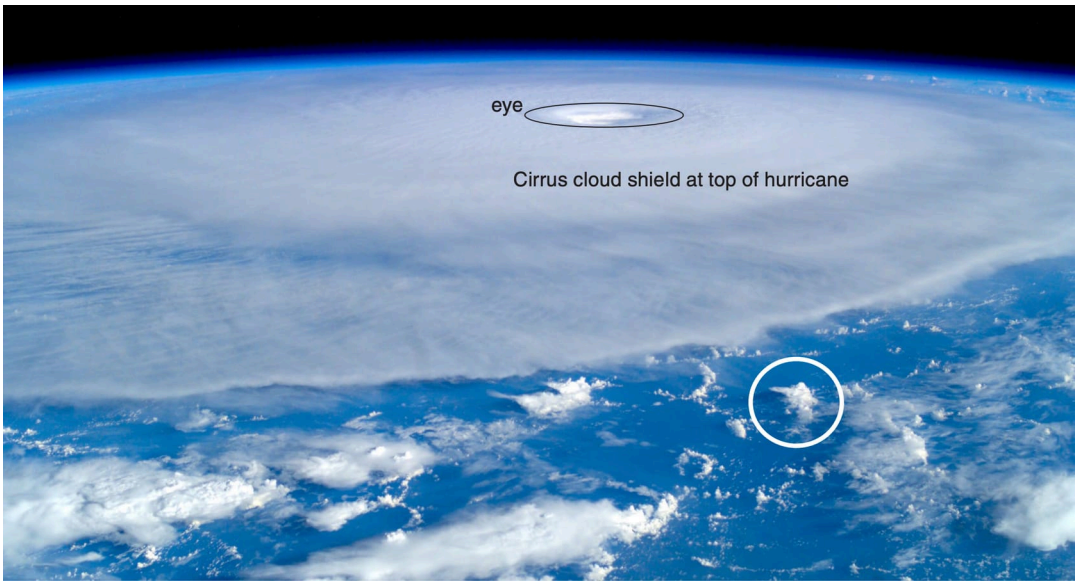


Figure 16.3 Hurricane Isabel as viewed from the International Space Station on 13 Sep 2003. The hurricane eye is outlined with the thin black oval, while the outflow from the eyewall thunderstorms fills most of this image. For scale comparison, the white circle outlines a nearby thunderstorm of depth and diameter of roughly 15 km. [Image courtesy of NASA - Johnson Spaceflight Center.]

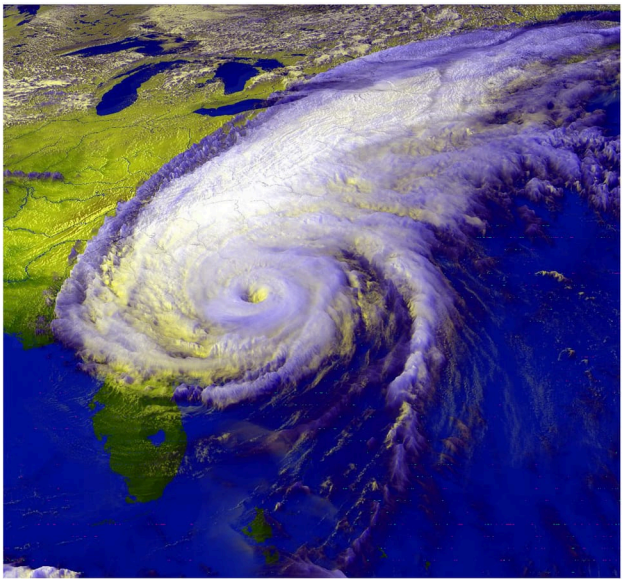


Figure 16.4 NOAA GOES satellite-derived 3-D rendering of hurricane Floyd at 2015 UTC on 15 Sep 1999, showing asymmetric outflow of this Category 4 storm extending 1000s of km north, past the Great Lakes into Canada. The eye is in the NW Atlantic just to the east of the Georgia-Florida border, USA. [Image by Hal Pierce is courtesy of the Laboratory for Atmospheres, NASA Goddard Space Flight Center.]

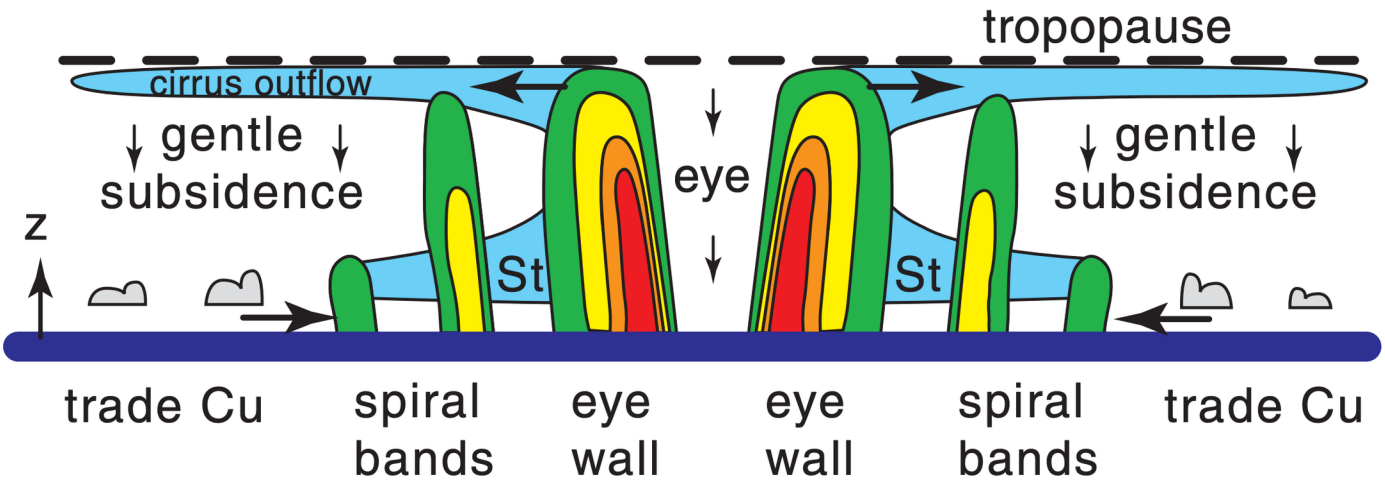


Figure 16.5 Vertical slice through a tropical cyclone. Green, yellow, orange, and red colors suggest the moderate to heavy rainfall as seen by radar. Light blue represents stratiform (St) clouds. The lighter rain that falls from the lower stratiform cloud deck of real tropical cyclones is not shown here. Tropopause altitude ≈ 15 km. Tropical-cyclone width ≈ 1500 km. Fair-weather trade-wind cumulus (Cu) clouds are sketched with grey shading. Arrows show radial and vertical wind directions.

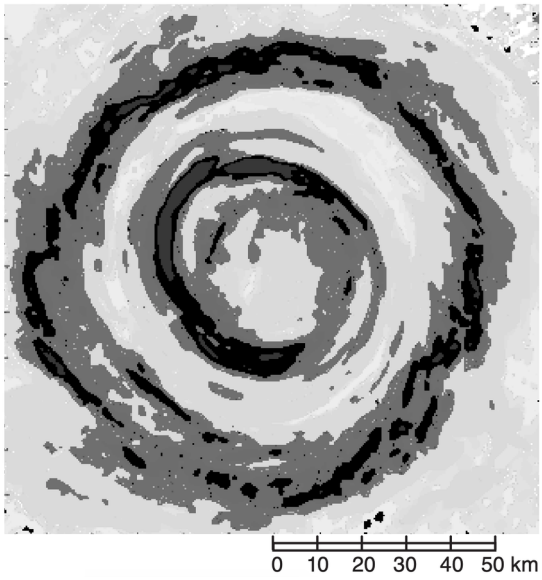


Figure 16.6 Zoomed view of double eyewall in Hurricane Rita, as viewed by airborne radar at 1 km altitude on 22 Sep 2005 at 1801 UTC. Darker greys show two concentric rings of heavier rain. Darkest grey corresponds to roughly 40 dBZ. [Modified from original image by Michael Bell and Wen-Chau Lee.]

Sometimes these spiral **rain bands** will merge to form a temporary second eyewall of thunderstorms around the original eye wall (Fig. 16.6). The lighter rain between the two eyewalls is called the **moat**.

During an **eyewall replacement cycle**, in very strong tropical cyclones, the inner eyewall dissipates and is replaced by the outer eyewall. When this happens, tropical cyclone intensity sometimes diminishes temporarily, and then strengthens again when the new outer eyewall diameter shrinks to that of the original eyewall.

This page titled 12.2.1: Tropical Cyclone Structure is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Roland Stull via source content that was edited to the style and standards of the LibreTexts platform.

- 16.1: Tropical Cyclone Structure by Roland Stull is licensed CC BY-NC-SA 4.0. Original source: https://www.eoas.ubc.ca/books/Practical_Meteorology.

12.2.2: Intensity and Geographic Distribution

Lower sea-level pressures in the eye and faster winds in the lower troposphere indicate stronger tropical cyclones. Several different tropical cyclone scales have been devised to classify tropical cyclone strength, as summarized below.

16.2.1. Saffir-Simpson Hurricane Wind Scale

Herbert Saffir and Robert Simpson created a scale for Atlantic hurricane strength that came to be known as the **Saffir-Simpson Hurricane Scale**. This scale has been used since the 1970s to inform disaster-response officials about hurricane strength.

Table 16-1. Saffir-Simpson Hurricane Wind Scale — Definition. Based on max 1-minute sustained wind speed measured at standard anemometer height of 10 m. Applies in the Atlantic Ocean, and in the eastern Pacific from the coast of the Americas to 180°W.

Category	Wind Speed			
	$m\ s^{-1}$	$km\ h^{-1}$	knots	$miles\ h^{-1}$
1	33 - 42	119 - 153	64 - 82	74 - 95
2	43 - 49	154 - 177	83 - 95	96 - 110
3	50 - 58	178 - 209	96 - 112	111 - 129
4	59 - 69	209 - 251	113 - 136	130 - 156
5	≥ 70	≥ 252	≥ 137	≥ 157

Table 16-2. Saffir-Simpson Hurricane Wind Scale — Description of Expected Damage — Examples.

Category	Concise Statement
	Damage Expected
	Examples
1	Very dangerous winds will produce some damage.
	People, livestock, and pets struck by flying or falling debris could be injured or killed. Older (mainly pre-1994 construction) mobile homes could be destroyed, especially if they are not anchored properly as they tend to shift or roll off their foundations. Newer mobile homes that are anchored properly can sustain damage involving the removal of shingle or metal roof coverings, and loss of vinyl siding, as well as damage to carports, sunrooms, or lanais. Some poorly constructed frame homes can experience major damage, involving loss of the roof covering and damage to gable ends as well as the removal of porch coverings and awnings. Unprotected windows may break if struck by flying debris. Masonry chimneys can be toppled. Well-constructed frame homes could have damage to roof shingles, vinyl siding, soffit panels, and gutters. Failure of aluminum, screened-in swimming pool enclosures can occur. Some apartment building and shopping center roof coverings could be partially removed. Industrial buildings can lose roofing and siding especially from windward corners, rakes, and eaves. Failures to overhead doors and unprotected windows will be common. Windows in high-rise buildings can be broken by flying debris. Falling and broken glass will pose a significant danger even after the storm. There will be occasional damage to commercial signage, fences, and canopies. Large branches of trees will snap and shallow rooted trees can be toppled. Extensive damage to power lines and poles will likely result in power outages that could last a few to several days.
	Hurricane Dolly (2008) brought Category 1 winds and impacts to South Padre Island, Texas
2	Extremely dangerous winds will cause extensive damage.

	There is a substantial risk of injury or death to people, livestock, and pets due to flying and falling debris. Older (mainly pre-1994 construction) mobile homes have a very high chance of being destroyed and the flying debris generated can shred nearby mobile homes. Newer mobile homes can also be destroyed. Poorly constructed frame homes have a high chance of having their roof structures removed especially if they are not anchored properly. Unprotected windows will have a high probability of being broken by flying debris. Well-constructed frame homes could sustain major roof and siding damage. Failure of aluminum, screened-in swimming pool enclosures will be common. There will be a substantial percentage of roof and siding damage to apartment buildings and industrial buildings. Unreinforced masonry walls can collapse. Windows in high-rise buildings can be broken by flying debris. Falling and broken glass will pose a significant danger even after the storm. Commercial signage, fences, and canopies will be damaged and often destroyed. Many shallowly rooted trees will be snapped or uprooted and block numerous roads. Near-total power loss is expected with outages that could last from several days to weeks. Potable water could become scarce as filtration systems begin to fail.
	Hurricane Frances (2004) brought Category 2 winds and impacts to coastal portions of Port St. Lucie, Florida.
3	Devastating damage will occur.
	There is a high risk of injury or death to people, livestock, and pets due to flying and falling debris. Nearly all older (pre-1994) mobile homes will be destroyed. Most newer mobile homes will sustain severe damage with potential for complete roof failure and wall collapse. Poorly constructed frame homes can be destroyed by the removal of the roof and exterior walls. Unprotected windows will be broken by flying debris. Well-built frame homes can experience major damage involving the removal of roof decking and gable ends. There will be a high percentage of roof covering and siding damage to apartment buildings and industrial buildings. Isolated structural damage to wood or steel framing can occur. Complete failure of older metal buildings is possible, and older unreinforced masonry buildings can collapse. Numerous windows will be blown out of high-rise buildings resulting in falling glass, which will pose a threat for days to weeks after the storm. Most commercial signage, fences, and canopies will be destroyed. Many trees will be snapped or uprooted, blocking numerous roads. Electricity and potable water will be unavailable for several days to a few weeks after the storm passes.
	Hurricane Ivan (2004) brought Category 3 winds and impacts to coastal portions of Gulf Shores, Alabama.
4	Catastrophic damage will occur.
	There is a very high risk of injury or death to people, livestock, and pets due to flying and falling debris. Nearly all older (pre-1994) mobile homes will be destroyed. A high percentage of newer mobile homes also will be destroyed. Poorly constructed homes can sustain complete collapse of all walls as well as the loss of the roof structure. Well-built homes also can sustain severe damage with loss of most of the roof structure and/or some exterior walls. Extensive damage to roof coverings, windows, and doors will occur. Large amounts of windborne debris will be lofted into the air. Windborne debris damage will break most unprotected windows and penetrate some protected windows. There will be a high percentage of structural damage to the top floors of apartment buildings. Steel frames in older industrial buildings can collapse. There will be a high percentage of collapse to older unreinforced masonry buildings. Most windows will be blown out of highrise buildings resulting in falling glass, which will pose a threat for days to weeks after the storm. Nearly all commercial signage, fences, and canopies will be destroyed. Most trees will be snapped or uprooted and power poles downed. Fallen trees and power poles will isolate residential areas. Power outages will last for weeks to possibly months. Long-term potable-water shortages will increase human suffering. Most of the area will be uninhabitable for weeks or months.
	Hurricane Charley (2004) brought Category 4 winds and impacts to coastal portions of Punta Gorda, Florida.
5	Catastrophic damage will occur.

	People, livestock, and pets are at very high risk of injury or death from flying or falling debris, even if indoors in mobile homes or framed homes. Almost complete destruction of all mobile homes will occur, regardless of age or construction. A high percentage of frame homes will be destroyed, with total roof failure and wall collapse. Extensive damage to roof covers, windows, and doors will occur. Large amounts of windborne debris will be lofted into the air. Windborne debris damage will occur to nearly all unprotected windows and many protected windows. Significant damage to wood roof commercial buildings will occur due to loss of roof sheathing. Complete collapse of many older metal buildings can occur. Most unreinforced masonry walls will fail, which can lead to collapse of buildings. A high percentage of industrial buildings and low-rise apartment buildings will be destroyed. Nearly all windows will be blown out of high-rise buildings resulting in falling glass, which will pose a threat for days to weeks after the storm. Nearly all commercial signage, fences, and canopies will be destroyed. Nearly all trees will be snapped or uprooted and power poles downed. Fallen trees and power poles will isolate residential areas. Power outages will last for weeks to possibly months. Long-term potable-water shortages will increase human suffering. Most of the area will be uninhabitable for weeks or months.
	Hurricane Andrew (1992) brought Category 5 winds and impacts to coastal portions of Cutler Ridge, Florida.

From an online report "The Saffir-Simpson Hurricane Wind Scale", National Hurricane Center, National Weather Service, NOAA. 2010.

In 2010, the US National Hurricane Center updated this scale (now called the **Saffir-Simpson Hurricane Wind Scale**), and defined hurricane intensity on wind speed only (Table 16-1). The scale ranges from category 1 for a weak hurricane to category 5 for a strong hurricane. Table 16-2 has a description of the expected damage for each category.

16.2.2. Typhoon Intensity Scales

For typhoons in the western North Pacific Ocean, storm intensity has been classified three different ways by three different organizations: the **Japan Meteorological Agency** (**JMA**), the **Hong Kong Observatory** (**HKO**), and the **US Joint Typhoon Warning Center** (**JTWC**). See Table 16-3.

16.2.3. Other Tropical-Cyclone Scales

Additional tropical-cyclone intensity scales (with different max wind-speed M_{max} definitions and category names) have been defined by agencies in:

- Australia (**Australian Bureau of Meteorology**, for S. Hemisphere east of 90°E). Cat. 1 for $M_{max} < 125$; Cat. 2 for $125 \le M_{max} \le 164$; Cat. 3 for $165 \le M_{max} \le 224$; Cat. 4 for $225 \le M_{max} \le 279$; and Cat. 5 for $280 < M_{max}$, where all winds are in km h^{-1}.
- India (**Regional Specialized Meteorological Center in New Delhi**, for N. Hemisphere Indian Ocean between 45°E and 100°E).
- France (**Météo-France**, for S. Hemisphere west of 90°E).

Even more confusing, some agencies use 1-minute average (sustained) winds, some use 10-minute average (sustained) winds, and some use gusts.

16.2.4. Geographic Distribution and Movement

Fig. 16.7 outlines the regions of greatest frequency of tropical cyclones and shows typical storm tracks.

Tropical cyclones are steered mostly by the large-scale global circulation. Most tropical cyclones form between 10° and 30° latitude, which is the trade-wind region. Hence, most tropical cyclones are steered from east to west initially

Later, these storms tend to turn poleward under the influence of monsoon circulations. For example, the Bermuda High (also known as the Azores High) over the North Atlantic Ocean has a clockwise circulation around it (see Fig. 11.31 in the General Circulation chapter), which turns tropical cyclones northward in the Western North Atlantic Ocean. The tracks of tropical cyclones and storms that have weakened from tropical cyclones to tropical storms continue to turn toward the northeast as they proceed further into the mid-latitudes and encounter west winds in the global circulation.

A striking observation in Fig. 16.7 is that no tropical cyclones form at the equator. Also, none cross the equator. The next section explains why.

INFO • Cyclone Damage Potential (CDP)

The total amount of destruction depends on the area of hurricane force winds, max wind speed, and the storm duration over any location. For this reason, the insurance industry utilizes a **cyclone damage potential index** (CDP) defined as follows:

$$CDP = \left(\frac{M_{to}}{M_t}\right) \cdot \left[\left(\frac{M_{\max}}{M_{\max o}}\right)^3 + 5\left(\frac{R_h}{R_{ho}}\right)\right] \qquad (12.2.2.1)$$

where M_t = translation speed (km h^{-1}) of the eye, M_{max} is max surface wind speed (km h^{-1}), R_h = average radius (km) of hurricane force winds, and the corresponding parameters are M_{to} = 7.4 km h^{-1}, Mmaxo = 120 km h^{-1}, R_{ho} = 96 km. If M_t < 9.25 km h^{-1}, then set M_t = 9.25 km h^{-1}. CDP varies between 0 and 10 (dimensionless), with 10 indicating the worst destruction.

For example, when hurricane Katrina was a category 5 storm before landfall, it had M_t = 20 km h^{-1}, M_{max} =280 km h^{-1}, & R_h = 100 km, yielding **CDP ≈ 6.7** .

Table 16-3. Typhoon (tropical cyclone) intensity scale, based on max sustained winds during 10-minute periods. Defined by JMA, unless otherwise noted. Applies between 100°E to 180°E.

Category	Wind Speed		
	m s^{-1}	km h^{-1}	knots
Tropical Depression	< 17	< 61	< 33
Tropical Storm	17 - 24	62 - 88	34 - 47
Severe Tropical Storm	25 - 32	89 - 117	48 - 63
Typhoon	33 - 41	118- 56	64 - 74
Severe (or Very Strong) Typhoon	42 - 51 (HKO) 44 - 54	150 - 184 (HKO) 157 - 193	80 - 99 (HKO) 85 - 104
Super Typhoon	> 51 (HKO) > 59 (JTWC)	>185 (HKO) > 211 (JTWC) > 194	>100 (HKO) > 114 (JTWC) > 105

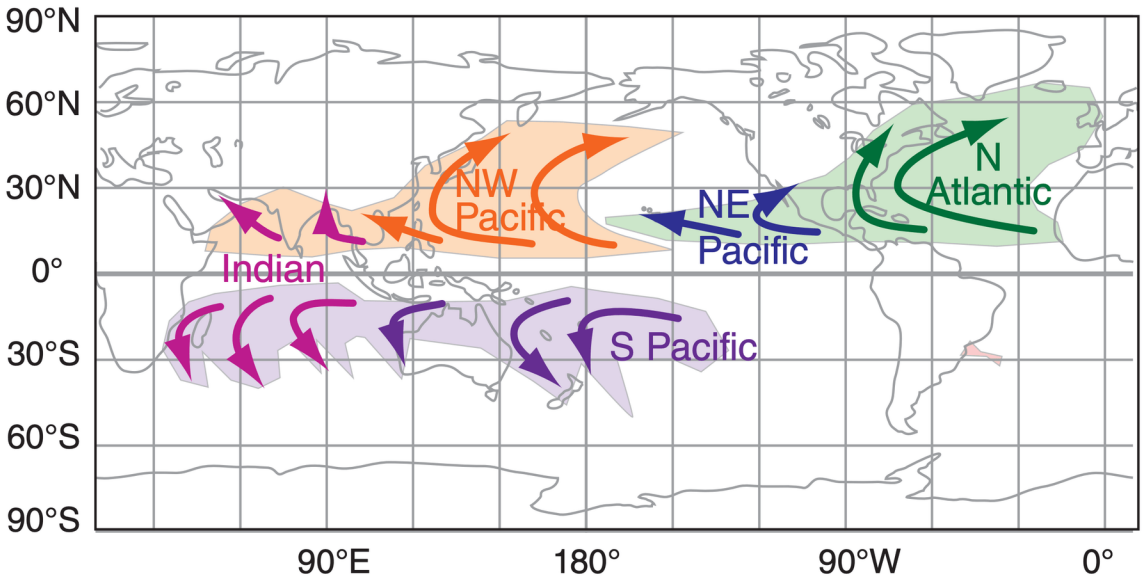


Figure 16.7 Map of tropical-cyclone locations (shading), with a general sense of the cyclone tracks (arrows).

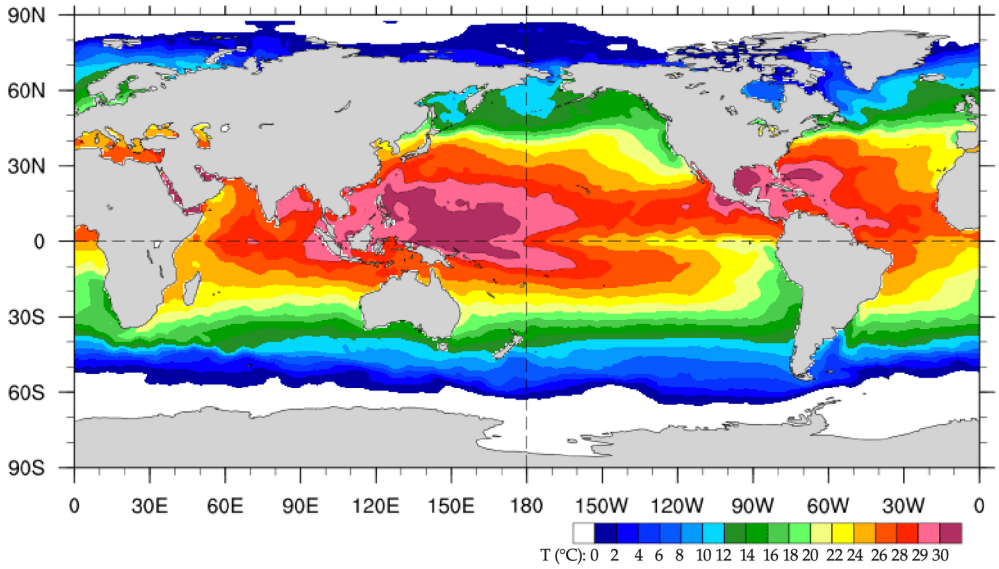


Figure 16.8 Sea surface temperatures (°C) as one-week average during 4 - 10 Sep 2016. [Courtesy of the NOAA/Earth System Research Lab/ Physical Sciences Division.]

This page titled 12.2.2: Intensity and Geographic Distribution is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Roland Stull via source content that was edited to the style and standards of the LibreTexts platform.

- 16.2: Intensity and Geographic Distribution by Roland Stull is licensed CC BY-NC-SA 4.0. Original source: https://www.eoas.ubc.ca/books/Practical_Meteorology.

12.2.3: Evolution

16.3.1. Requirements for Cyclogenesis

Seven conditions are necessary for tropical cyclones to form: a warm sea surface, non-zero Coriolis force, nonlocal conditional instability, high humidity in the mid troposphere, weak ambient wind shear, enhanced synoptic-scale vorticity, and a trigger.

16.3.1.1. Warm Sea Surface

The **sea surface temperature** (**SST**) must be approximately 26.5°C or warmer (Figs. 16.8 and 16.9 on this page, and 16.51 at the end of this chapter), and the warm surface waters must be at least 50 m deep. This warm temperature is needed to enable strong evaporation and heat transfer from the sea surface into the boundary layer. The warmer, more-humid boundary-layer air serves as the fuel for thunderstorms in the tropical cyclone.

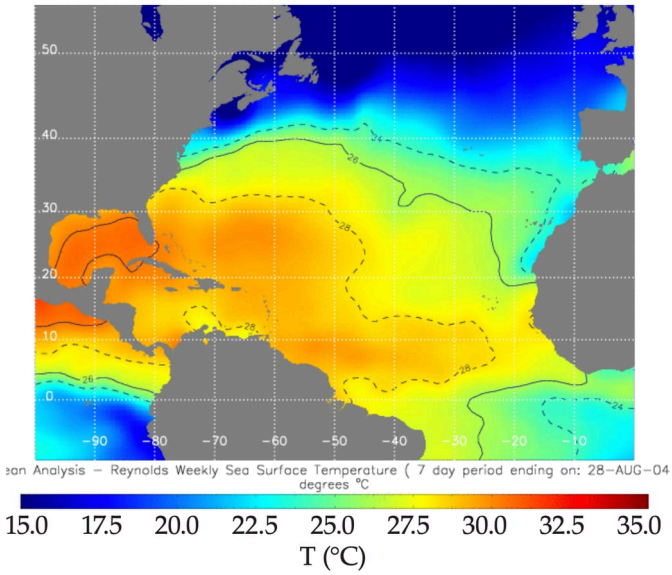


Figure 16.9 Analysis of Atlantic-ocean sea-surface temperatures (°C) averaged over 7 days ending on 28 Aug 2004. Hurricane Francis occurred during this time period . [Adapted from a USA National Hurricane Center image using NCEP/NOAA data.]

The fast winds in tropical cyclones create large waves that stir the top part of the ocean. If the warm waters are too shallow, then this turbulent mixing will stir colder deeper water up to the surface (see INFO Box two pages later). When this happens, the sea-surface temperature decreases below the 26.5°C threshold, and the tropical cyclone kills itself.

16.3.1.2. Coriolis Force

Tropical cyclones cannot exist within about 500 km of the equator (i.e., ≤ 5° latitude), because Coriolis force is near zero there (and is exactly zero right at the equator). Rarely, very small-diameter tropical cyclones have been observed closer to the equator, but none are observed right at the equator. Tropical cyclones cannot form at the equator, and existing cyclones cannot cross the equator (Fig. 16.7).

Without Coriolis force, boundary-layer air would be sucked directly into the eye by the low pressure there. Thus, air molecules would accumulate in the eye, causing pressure to increase towards ambient values. The result is that the low pressure would disappear, winds would die, and the tropical cyclone would cease to exist in less than a day. This is what happens to tropical cyclones that approach the equator (Fig. 16.10).

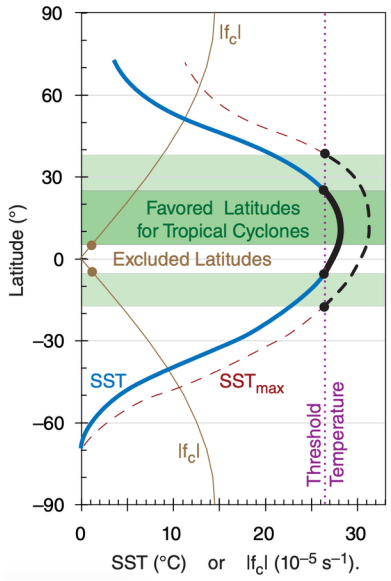


Figure 16.10 Approximate zonal average sea-surface temperature SST (°C, thick solid blue line) averaged around the globe on 3 Aug 2009. Also shown is the max SST (°C, thin dashed red line) on that day at each latitude. The magnitude of the Coriolis parameter f_C (10^{-5} s^{-1}) is indicated by the thin solid brown line. Dotted purple line indicates the SST threshold of roughly 26.5°C for tropical cyclones. Those latitudes that exceed the temperature threshold and have Coriolis parameter that is not near zero are favored locations for tropical cyclones (shaded medium green). Tropical cyclones can also occur in limited regions (shaded light green) at other latitudes where the max SST exceeds the threshold. "Excluded latitudes" are those for which Coriolis force is too small. As the seasons progress, the location of peak SST shifts, causing the favored latitudes to differ from this particular August example.

But with Coriolis force, winds in the bottom half of the troposphere are forced around the eye at gradient- or cyclostrophic-wind speeds. Namely, most of the air is flowing tangentially around the eye rather than flowing radially into it. Only close to the ground does drag change the winds into boundarylayer gradient winds, resulting in a small amount of convergence toward the eye. Thus, the tropical cyclone can persist for many days.

16.3.1.3. Nonlocal Conditional Instability

Because tropical cyclones are made of thunderstorms, the tropical environment must have sufficient nonlocal conditional instability to support deep thunderstorm convection. Namely, there must be a stable layer (i.e., a cap) above a warm humid boundary layer, and the mid-troposphere must be relatively cool compared to the boundary layer.

The warm humid boundary layer is achieved via strong heat and moisture fluxes from the warm sea surface into the air. The cap is the trade-wind inversion that was discussed in the General Circulation chapter. These conditions, combined with a cool mid troposphere, lead to large values of convective available potential energy (CAPE), as was thoroughly explained in the Thunderstorm chapter. Hence, large values of CAPE imply sufficient nonlocal conditional instability for tropical cyclones.

16.3.1.4. High Humidity in Mid-troposphere

In a deep layer of air centered at roughly 5 km above sea level, humidity must be high. Otherwise, the incipient thunderstorms cannot continue to grow and organize into tropical cyclones.

Note that this differs from mid-latitude thunderstorms, where a drier mid-troposphere allows more violent thunderstorms. When dry environmental air is entrained into the sides of mid-latitude thunderstorms, some of the storm's hydrometeors evaporate, causing the strong downdrafts that define supercell storms and which create downbursts, gust fronts, and can help trigger tornadoes.

In the tropics, such downbursts from any one thunderstorm can disrupt neighboring thunderstorms, reducing the chance that neighboring thunderstorms can work together in an incipient eyewall. Also, the cold downburst air accumulates at the bottom of the troposphere, thereby increasing static stability and reducing deep convection.

16.3.1.5. Weak Ambient Wind Shear

Wind shear within four degrees of latitude of the incipient storm must be weak ($\Delta M < 10$ m s^{-1} between pressure levels 80 and 25 kPa) to enable thunderstorm clusters to form. These clusters are the precursors to tropical cyclones.

If the shear is too strong, the updrafts in the thunderstorms become tilted, and latent heating due to water-vapor condensation is spread over a broader area. This results in less-concentrated warming, and a reduced ability to create a low-pressure center at sea level around which the thunderstorms can become organized into a tropical cyclone.

This requirement differs from that for mid-latitude thunderstorms. At mid-latitudes, strong shear in the ambient environment is good for storms because it encourages the creation of mesocyclones and supercell thunderstorms. Apparently, in the tropics such rotation of individual thunderstorms is bad because it interferes with the collaboration of many thunderstorms to create an eyewall.

16.3.1.6. Synoptic-scale Vorticity

A relative maximum of relative vorticity in the bottom half of the troposphere can help organize the thunderstorms into an incipient tropical cyclone. Otherwise, any thunderstorms that form would act somewhat independently of each other.

INFO • Trop. Cyclone-induced Currents

Near-surface winds cause net ocean transport 90° to the right of the wind vector in the N. Hemisphere, as explained later in the Storm Surge section. Given tangential winds that circle a tropical cyclone, we expect the net Ekman transport (average ocean currents induced by the tropical cyclone) to be outward (Fig. 16.a)

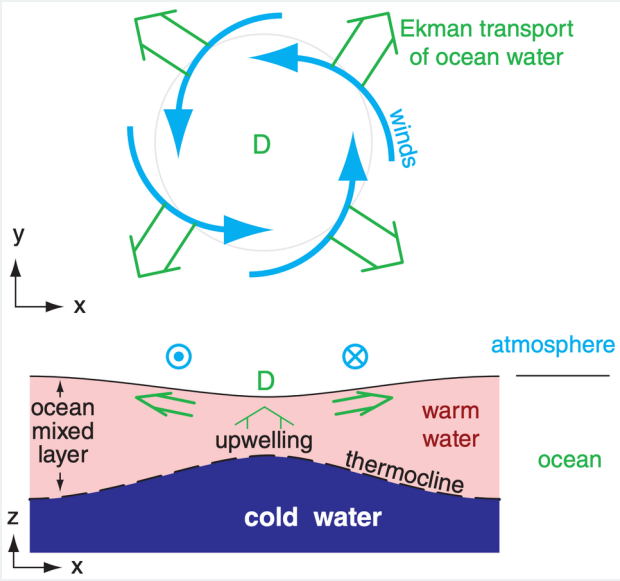


Fig. 16.a. Ocean currents (green double-shaft arrows) induced by tropical-cyclone winds (cyan arrows).

Such horizontal divergence (D) of the ocean-surface water does three things: (1) it lowers the sea surface by removing sea water; (2) it causes upwelling of water toward the surface; and (3) it causes the colder deep water to be brought closer to the surface (Fig. 16.a-bottom). The **thermocline** is the interface between cold deep water and warmer mixed-layer above.

The tropical cyclone brings cold water closer to the surface, where ocean mixed-layer turbulence caused by breaking waves can further mix the warm and cold waters. Thus, tropical cyclones will kill themselves by cooling the sea-surface temperature unless the prestorm warm ocean mixed-layer is ≥ 50 m deep.

16.3.2. Tropical Cyclone Triggers

Even if all six previous conditions are satisfied, a method to trigger the tropical cyclone is also needed. A trigger is anything that creates synoptic-scale horizontal convergence in the atmospheric boundary layer. Such horizontal convergence forces upward motion out of the boundary-layer top as required by mass conservation (see the Atmospheric Forces & Winds chapter). Synoptic-scale upward motion can initiate and support an organized cluster of thunderstorms — incipient tropical cyclones.

Some triggers are: the ITCZ, easterly waves, Monsoon troughs, mid-latitude fronts that reach the tropics, and Tropical Upper Tropospheric Troughs.

16.3.2.1. ITCZ

The **Intertropical Convergence Zone** (**ITCZ,** see General Circulation chapter) is the region where the northeasterly trade winds from the Northern Hemisphere meet the southeasterly trade winds from the Southern Hemisphere. This convergence zone shifts between about 10°N during August and September (end of N. Hemisphere summer, Fig. 16.11), and 10°S during February and March (end of S. Hemisphere summer). During these months of maximum asymmetry in global heating, the ITCZ is far enough from the equator for sufficient Coriolis force to enable tropical cyclones.

16.3.2.2. African Easterly Wave

Although equatorial Africa is hot, the Sahara Desert (roughly spanning 15° to 30°N) and Arabian Peninsula are even hotter. This temperature excess is within an atmospheric boundary layer that can reach 5 km depth by late afternoon, although average depth over 24 h is about 3 km. Thus, along the southern edge of the Sahara and the Arabian Peninsula is a strong temperature gradient (Fig. 16.12).

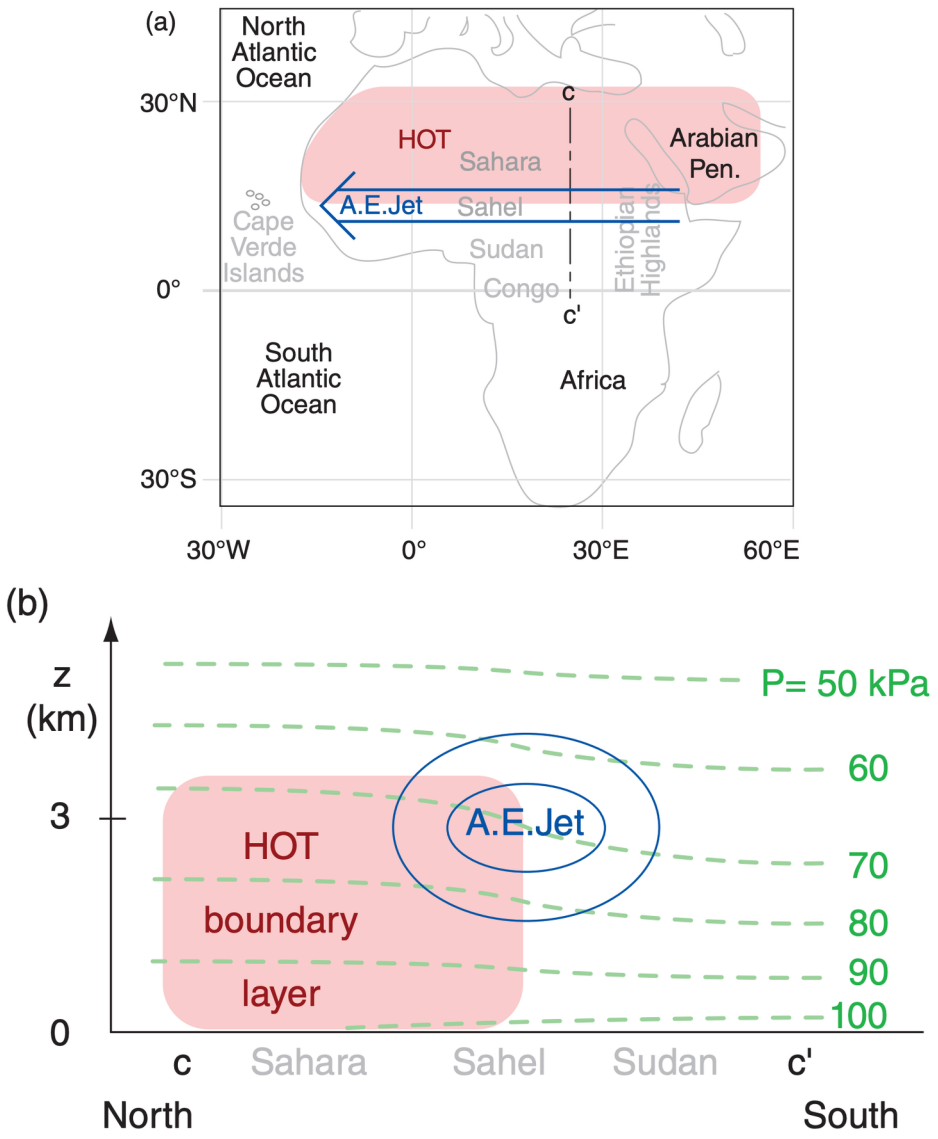


Figure 16.12 African Easterly (A.E.) Jet. (a) Map of Africa. (b) Vertical cross section along c - c' in (a), where thin blue ovals represent isotachs around the core of the AEJ. Pink area indicates hot air. Note the low altitude of this A.E. jet.

As explained in the thermal-wind section of the General Circulation chapter, a north-south temperature gradient creates a north-south pressure gradient that increases with altitude (up to the top of the boundary layer for this scenario). For Africa, higher pressure aloft is over the Sahara Desert and lower pressure aloft is further south over the Sudan and Congo.

This pressure gradient drives a geostrophic wind from the east over the African Sahel (at about 15°N). These winds reach maximum speeds of 10 to 25 m s^{-1} at an altitude of about 3 km in late summer. This is called the **African Easterly Jet** (**AEJ**).

Like the subtropical and mid-latitude jets, the AEJ is barotropically and baroclinically unstable. This means that the jet tends to meander north and south while blowing from east to west. The mountains in eastern Africa (Ethiopia Highlands and Darfur Mountains) can trigger such oscillations.

The oscillations in this low-altitude jet are called waves, and are known as **Easterly Waves** because the wave troughs and crests move from the east at speeds of 7 to 8 m s^{-1}. These waves have wavelength of about 2,000 to 2,500 km, and have a wave period of 3 to 4 days. Although these waves are created over Africa, they continue to propagate west across the tropical North Atlantic, the Caribbean, central America, and finally into the tropical Northeast Pacific (Fig. 16.13).

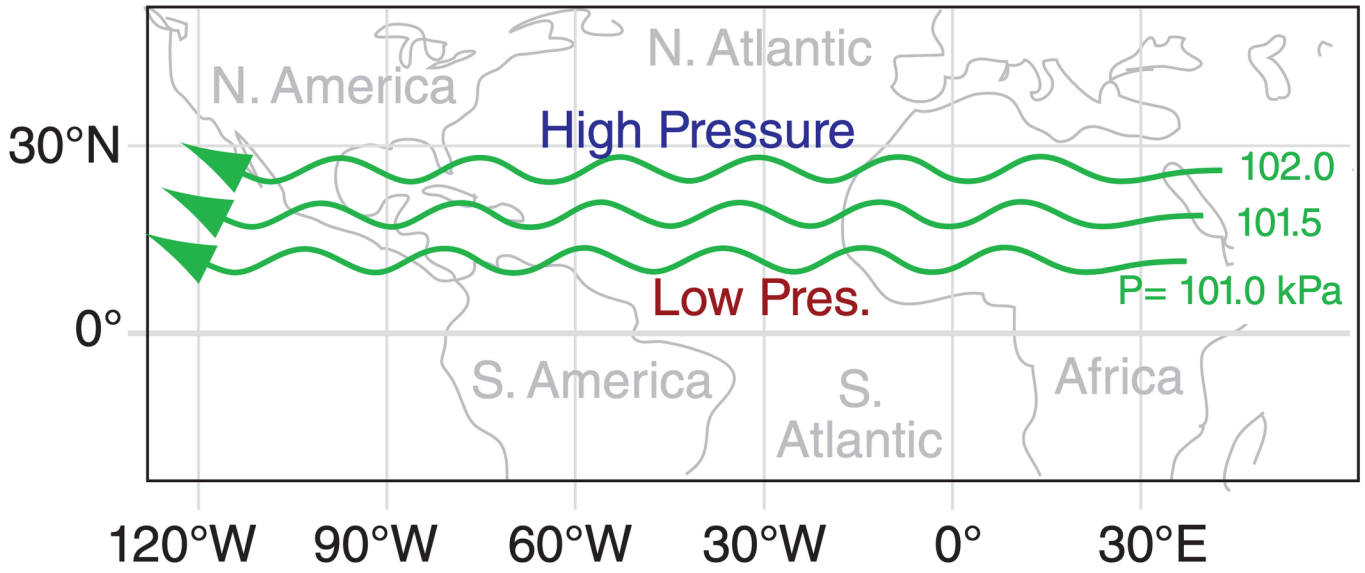


Figure 16.13 Easterly waves near the bottom of the troposphere. Wavelength is roughly 20° to 25° of longitude (≈ 2000 to 2500 km). Isobars of sea-level pressure are solid green lines with arrows. Wave troughs propagate toward the west. Streamlines are roughly parallel to the isobars.

Easterly waves can trigger tropical cyclones because of the low-altitude convergence that occurs east of the wave troughs. These waves are found predominantly in a latitude band between 10°N and 30°N, where there is lower pressure to the south and the subtropical high-pressure belt to the north. Thus, a trough of low pressure in an easterly wave corresponds to a region where the isobars meander to the north (Fig. 16.14) in the N. Hemisphere.

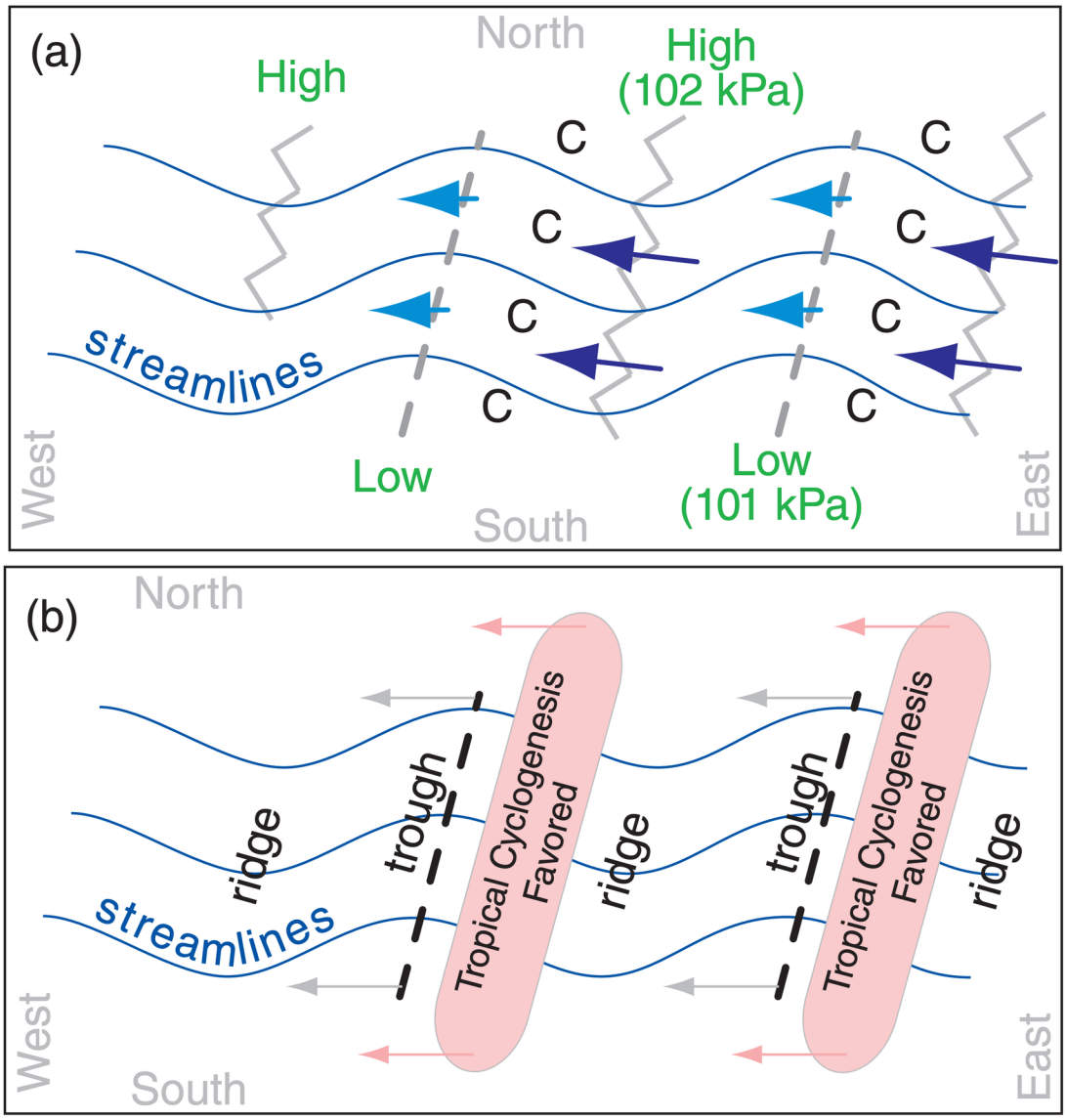


Figure 16.14 Characteristics of easterly waves in lower tropospheric flow over the tropical North Atlantic. (a) Blue arrows indicate winds; longer arrows denote faster winds; C = convergence. (b) Grey and pink arrows show how the easterly waves and the tropical cyclogenesis regions (pink shading) propagate toward the west.

Because the pressure gradient is weak and sometimes difficult to analyze from sparsely spaced weather stations in this region, tropical meteorologists will often look at the streamlines instead of isobars. Streamlines are weather-map lines that are everywhere parallel to the flow at any instant in time (see the Regional Winds chapter). They more-or-less follow the isobars. Thus, regions where streamlines for low-altitude winds meander to the north can also be used to locate troughs in the easterly waves.

Recall that mid-latitude cyclones are favored in the updrafts under divergence regions in the polar jet (centered at the top of the troposphere). However, for tropical cyclogenesis we also look for updrafts above the convergence regions of the jet near the bottom of the troposphere (Fig. 16.15a).

Where are these convergence regions in easterly waves? According to the Atmospheric Forces & Winds chapter, winds are slower-than-geostrophic around troughs, and faster around ridges. This rule continues to hold in the tropics and subtropics. Thus, the convergence zone with fast inflow and slow outflow is east of the trough axis (Fig. 16.14a).

Also, since the trough axis often tilts eastward with increasing altitude, it means that a mid-tropospheric trough is often over a lower-tropospheric convergence region. This trough has the cyclonic vorticity that encourages organization of thunderstorms into incipient tropical cyclones.

It is these regions east of troughs in easterly waves where the thunderstorms of incipient tropical cyclones can be triggered (Fig. 16.14b). As the easterly waves propagate from east to west, so move the convergence regions favoring tropical cyclone formation. This convergence mechanism is important for tropical cyclone triggering in both the North Atlantic and Northeast Pacific. Roughly 85% of intense Atlantic hurricanes are triggered by easterly waves.

16.3.2.3. Monsoon Trough

A **monsoon trough** (MT) is where the ITCZ merges into a monsoon circulation. For example, Fig. 16.16 shows a monsoon low pressure (L) region over southeast Asia (left side of Fig.), with winds rotating counterclockwise around it. Near the dashed line labeled MT in the western Pacific, monsoon winds from the southwest converge with easterly trade winds from the Hawaiian (Pacific) High.

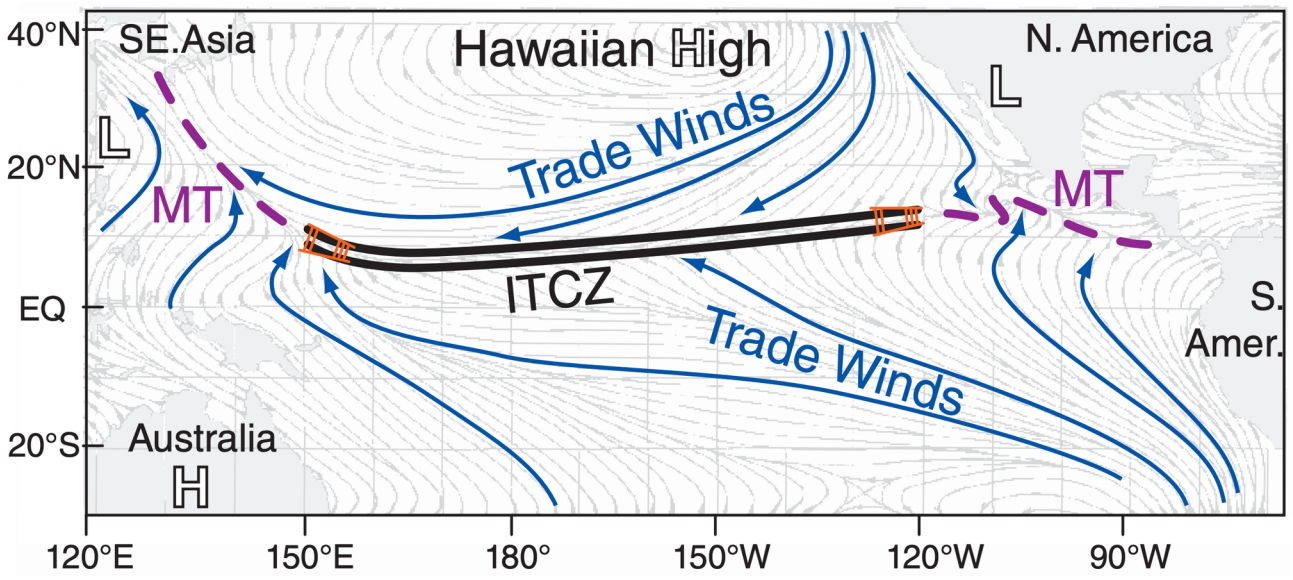


Figure 16.16 Streamline analysis (blue lines with arrows) of near-surface winds, averaged during August. MT denotes Monsoon Trough, shown by the thick dashed lines. L and H are monsoon lows and highs. [modified from Naval Research Lab. image]

Thus, the monsoon trough in the western Pacific is a convergence region that can trigger thunderstorms for incipient typhoons. Similar monsoon troughs have been observed in the Indian Ocean and the Eastern Pacific. In Fig. 16.16, the region where the monsoon trough in the western Pacific meets the ITCZ has enhanced convergence, and is a more effective trigger for tropical cyclones.

16.3.2.4. Mid-latitude Frontal Boundary

Sometimes cold fronts from midlatitudes can move sufficiently far equatorward to reach the subtropics. Examples include cold fronts reaching southern Florida (Fig. 16.17). Although these fronts are often weak and slow moving by the time the reach the subtropics, they nonetheless have convergence across them that can trigger tropical cyclone thunderstorms in the Caribbean and the subtropical western North Atlantic.

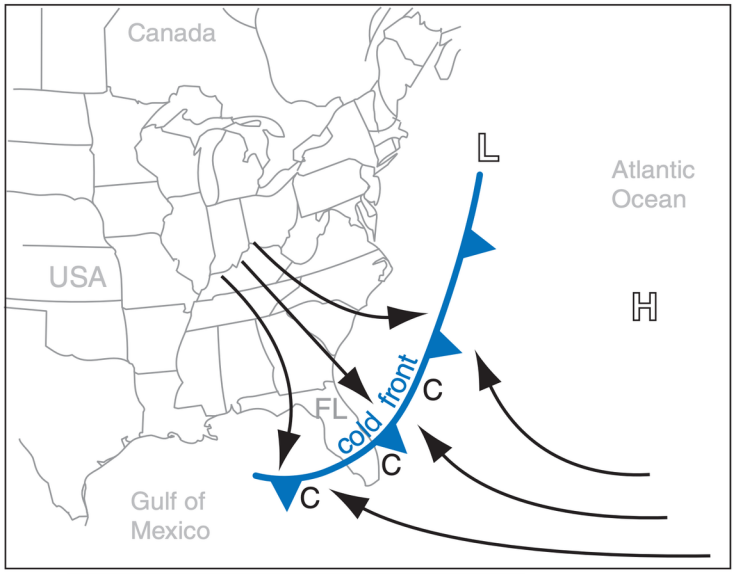


Figure 16.17 Lower tropospheric convergence (C) along a cold front over Florida (FL), USA.

16.3.2.5. TUTT

TUTT stands for **Tropical Upper Tropospheric Trough**. This is a high-altitude cold-core low-pressure system that can form in the subtropics. East of the TUTT axis is a region of horizontal divergence aloft, which creates upward motion below it in order to satisfy air mass continuity (Fig. 16.15b). This upward motion can help trigger thunderstorms as precursors to tropical cyclones.

16.3.3. Life Cycle

At locations where all of the necessary conditions are met (including any one trigger), the incipient tropical cyclones usually progress through the following intensification stages: tropical disturbance, tropical depression, tropical storm, tropical cyclone.

16.3.3.1. Tropical Disturbance

The US National Hurricane Center defines a **tropical disturbance** as “a discrete tropical weather system of apparently organized convection (generally 200 to 600 km in diameter) originating in the tropics or subtropics, having a nonfrontal migratory character, and maintaining its identity for 24 hours or more.” Namely, it is a cluster of thunderstorms that stay together as they move across the ocean.

This thunderstorm cluster is visible by satellite as consisting of distinct thunderstorms with their own anvils and separate precipitation regions. There is no eye, and little or no rotation visible. Some tropical disturbances form out of **Mesoscale Convective Systems** (**MCSs**), particularly those that develop a **Mesoscale Convective Vortex** (**MCV**, see the Thunderstorm chapter). Some are identified as the thunderstorm clusters that move with easterly waves.

Most tropical disturbances do not evolve into tropical cyclones. Hence, tropical disturbances are not usually named or numbered. However, tropical meteorologists watch them carefully as they evolve, as potential future tropical cyclones.

As condensation and precipitation continue in this thunderstorm cluster, more and more latent heat is released in those storms. Namely, the mid-tropospheric air near this cluster becomes warmer than the surrounding ambient air. According to the hypsometric relationship, pressure decreases more slowly with increasing altitude in this warm region. Thus, synoptic-scale high pressure starts to form near the top of the troposphere in the region of this cluster.

16.3.3.2. Tropical Depression (TD)

The high pressure aloft begins to create a thermal circulation (see General Circulation chapter), where air aloft is driven horizontally outward — down the pressure gradient toward the lower pressure outside of the cluster. This diverging air aloft also

starts to rotate anticyclonically.

The outward moving air aloft removes air molecules from the thunderstorm cluster region of the troposphere, thereby lowering the surface pressure under the cluster. The “depressed” pressure at sea level is why this stage gets its name: **tropical depression**. Tropical depressions are given an identification number, starting with number 1 each year in each region.

As this weak surface low forms, it creates a pressure gradient that starts to suck winds horizontally toward the center of the low. This inflowing air begins rotating cyclonically due to Coriolis force. The surface low pressure is usually too weak to measure directly. However, this tropical depression stage is defined by measurable near-surface winds of 17 m s–1 or less (Table 16-4) turning cyclonically as a closed circulation around the cluster. There is still no eye at this stage, however the thunderstorms are becoming aligned in spiral rain bands (Fig. 16.18). Most tropical depressions do not strengthen further.

Table 16-4. Stages leading to tropical cyclones. M_{max} = near-surface wind-speed maximum. n/a = not applicable.

Stage	M_{max} ($m\ s^{-1}$)	($km\ h^{-1}$)	(knots)
Tropical Disturbance	n/a	n/a	n/a
Tropical Depression	< 17	< 61	≤ 33
Tropical Storm	17 - 32	61 - 118	33 - 63
Tropical Cycl.(Hurricane)	≥ 33	≥ 119	≥ 64

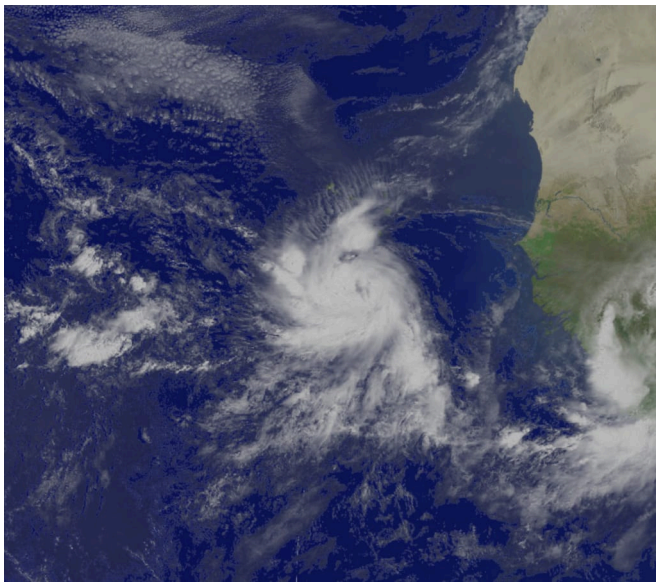


Figure 16.18 Tropical Depression 10 (in center of photo), over the tropical Atlantic Ocean south-southeast of the Cape Verde Islands. Africa is visible in the right quarter of the photo. (Visible satellite image taken 11:45 UTC on 2 Sep 2008, courtesy of NOAA.)

However, for the few storms that do continue to strengthen, the near-surface winds are boundarylayer gradient winds that spiral in towards the center of the cluster. This radial inflow draws in more warm humid boundary-layer air, thereby refueling the thunderstorms and allowing them to persist and strengthen. Condensation and precipitation increase, causing a corresponding increase in latentheat release.

This warm core of the storm further strengthens the high pressure aloft via the hypsometric relationship, which drives more outflow aloft and removes more air molecules from the troposphere near the cluster. Thus, the surface low pressure can continue to deepen in spite of the boundary-layer inflow. Namely, the high pressure aloft creates an exhaust system, while the boundary layer is the intake system to the incipient cyclone.

16.3.3.3. Tropical Storm (TS)

When the surface low is deep enough to drive winds faster than 17 m s^{-1} (but less than 33 m s^{-1}) in a closed cyclone circulation, then the system is classified as a **tropical storm**. On weather maps, it is indicated with symbol 

At this point, the thunderstorm rain bands have nearly completely wrapped into a circle — the future eyewall (Fig. 16.19d). The anvils from the thunderstorms have usually merged in to a somewhat **circular central dense overcast** (**CDO**), which is clearly visible as a large-diameter cold high cloud. There is no eye at this stage. Tropical storms are organized sufficiently to be able to modify their local environment to allow them to persist, without relying so much on a favorable ambient environment.

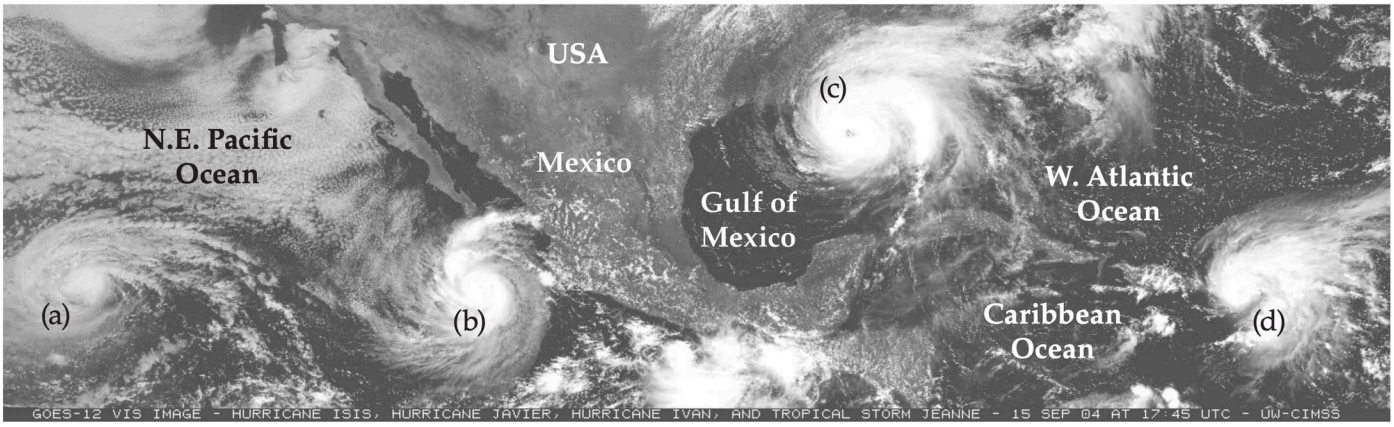


Figure 16.19 Visible satellite image at 17:45 UTC on 15 Sep 2004, showing four tropical systems. From left to right: (a) Hurricane Isis, (b) Hurricane Javier, (c) Hurricane Ivan, (d) Tropical Storm Jeanne. This image covers the eastern N. Pacific Ocean, Central America, Gulf of Mexico, Caribbean Ocean, and western N. Atlantic. (Image courtesy of University of Wisconsin–Madison CIMSS/ SSEC.)

At this stage the storm is given a name for identification (and its former tropical-depression number is dropped). The same name is used if the storm further strengthens into a tropical cyclone, and for the rest of its life cycle.

Each ocean region has its own naming convention for these tropical storms. Hurricanes (Atlantic and Eastern Pacific) and Cyclones (Indian Ocean and Southwestern Pacific near Australia) are usually given names of men or women. These are assigned in alphabetical order according to lists that have been set in advance by the **World Meteorological Organization** (**WMO**). Typhoons are given names of flowers, animals, birds, trees, foods (not in alphabetical order but in an order set by the WMO's Typhoon Committee).

Tropical-storm and tropical-cyclone names are re-used in a six-year cycle. However the names of the strongest, most destructive tropical cyclones are "retired" and never used again (they are replaced with other names starting with the same letter).

16.3.3.4. Tropical Cyclone (Hurricane, Typhoon)

Roughly half of the tropical storms in the Atlantic continue to strengthen into tropical cyclones. By this stage there is a well defined eye surrounded by an eyewall of thunderstorms. Max wind speeds are 33 m s^{-1} or greater in a closed circulation around

the eye. The central dense overcast usually has a cloudfree eye. The weather map symbol for a tropical cyclone is
.

There is heavy precipitation from the eyewall thunderstorms, and the core of the storm is significantly warmer than the environment outside the storm. Sea-level pressure decrease in the eye is measurable and significant, and a rise of sea level into a storm surge is possible. The storm organization allows it to persist for days to weeks as it moves across ocean basins (Fig. 16.19a-c).

16.3.4. Movement/Track

16.3.4.1. Steering by the General Circulation

Tropical cyclones are moved by the larger-scale winds near them. These winds are variable, causing a wide variety of hurricane tracks and translation speeds. However, by focusing on the climatological average trade winds and the monsoon circulations (Figs. 16.7 & 16.20), we can then anticipate average movement of tropical cyclones.

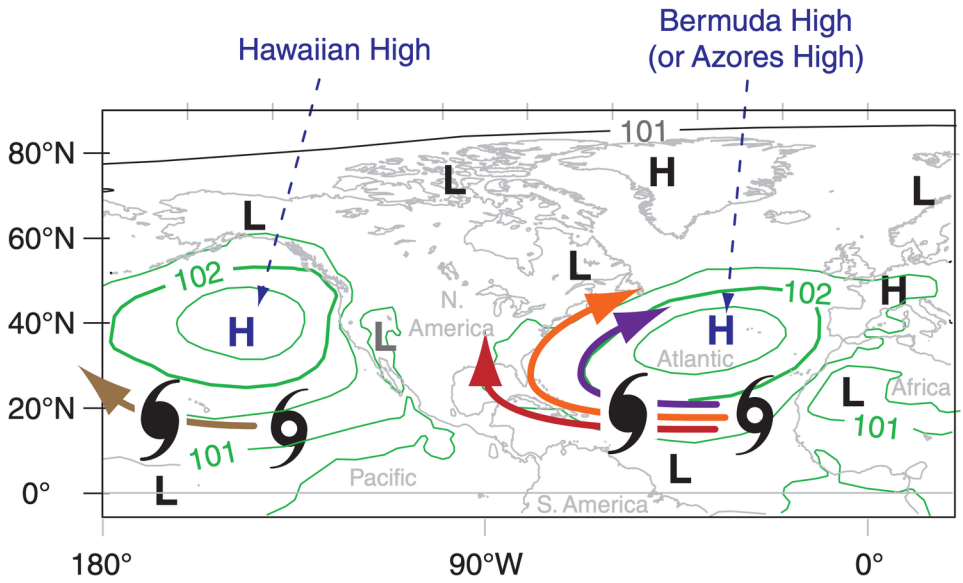


Figure 16.20 Monsoon sea-level pressure (kPa) pattern for July (copied from the General Circulation chapter), with isobars as thin green lines. Winds and average tropical cyclone tracks (dark colored arrows) tend to follow the isobars around the monsoon highs, although actual tracks are quite variable.

During late Summer and early Fall, monsoon high-pressure regions form over the oceans, as was discussed in the General Circulation chapter. Winds rotate clockwise (counterclockwise) around these highs in the Northern (Southern) Hemisphere. Equatorward of the high center the monsoon winds are in the same direction as the trade winds — from east to west. Thus, average tropical cyclone tracks in the eastern and central longitudes of an ocean basin are usually zonal — from east to west.

However, near the west sides of the ocean basins, the monsoon winds turn poleward (Fig. 16.20), and thus steer the tropical cyclones away from the equator. Many of these tropical cyclones travel along the east coast of continents in a poleward direction. The tracks vary from year to year, and vary within any one year (Fig. 16.20). Some of the tracks in the Atlantic might never hit N. America. Other tracks can be along the East Coast of North America, causing significant damage to coastal cities. Yet other

tracks continue further westward across the Caribbean Sea and into the Gulf of Mexico before reaching the mainland.

16.3.4.2. Extratropical Transition

As the monsoon circulations move the tropical cyclones poleward, the cyclones leave the tropics and subtropics and enter the mid-latitudes. Here, prevailing winds are from the west in the global circulation. Also, winds on the poleward side of the monsoon highs are from the west. Thus, many tropical cyclone tracks eventually turn towards the east at midlatitudes.

Extratropical transition is the name given to the movement of tropical cyclones out of the tropics and into the midlatitudes. At these latitudes the storms are generally moving over colder water, causing them to rapidly die. Nonetheless, these former tropical cyclones can change into extratropical cyclones and still cause damaging floods, wind, and waves if over coastal areas. Even dying tropical cyclones well away from a coastline can influence midlatitude weather downwind by pumping copious amounts of moisture into the air. This humid air can later serve as the fuel for mid-latitude cyclones thousands of kilometers downwind, creating heavy rain and possible flooding.

16.3.5. Tropical Cyclolysis

Tropical cyclones can exist for weeks because they have the ability to create their own fuel supply of warm humid boundary-layer air. They create this fuel by tapping the heat and water stored in the upper ocean. Tropical cyclones die (**tropical cyclolysis**) when they move to a location where they cannot create their own fuel, or if they are destroyed by larger-scale weather systems, as described next.

16.3.5.1. Movement over Land

If the global circulation steers a tropical cyclone over land, it begins to die rapidly because it is unable to tap the warm ocean for more fuel (Fig. 16.21). If the land is a small island or narrow isthmus, then the weakened tropical cyclone can possibly re-intensify when it again moves over warm ocean water.

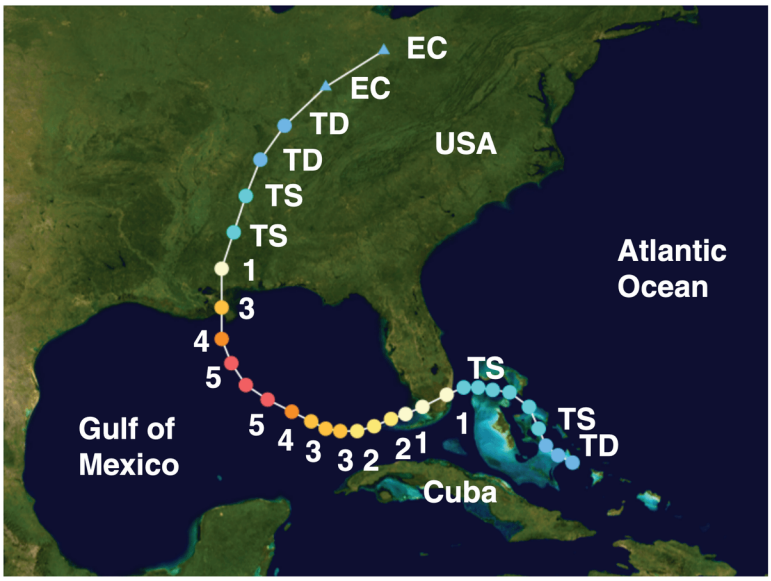


Figure 16.21 Track of hurricane Katrina, showing its intensity during 23 - 31 Aug 2005. 1 to 5 is Saffir-Simpson category. TS = tropical storm. TD = tropical depression. EC = extratropical cyclone.

16.3.5.2. Movement over Cold Water

If the global circulation steers a tropical cyclone over water colder than 26.5°C, then it is unable to create new warm humid air as fast the eyewall thunderstorms consume this fuel. As a result, tropical cyclones quickly weaken over cold water.

This often happens when tropical cyclones are steered poleward toward mid-latitudes, where sea surface temperatures (**SST**)

are colder (Fig. 16.8). In the Atlantic, the Gulf Stream of warm ocean water moving north along the east coast of the USA allows tropical cyclones there to survive. However, further north, the very cold Labrador Current (Fig. 16.22) moving south from the Arctic along the east coast of Canada causes most tropical cyclones to diminish below tropical cyclone force before making landfall (see INFO Box on next page).

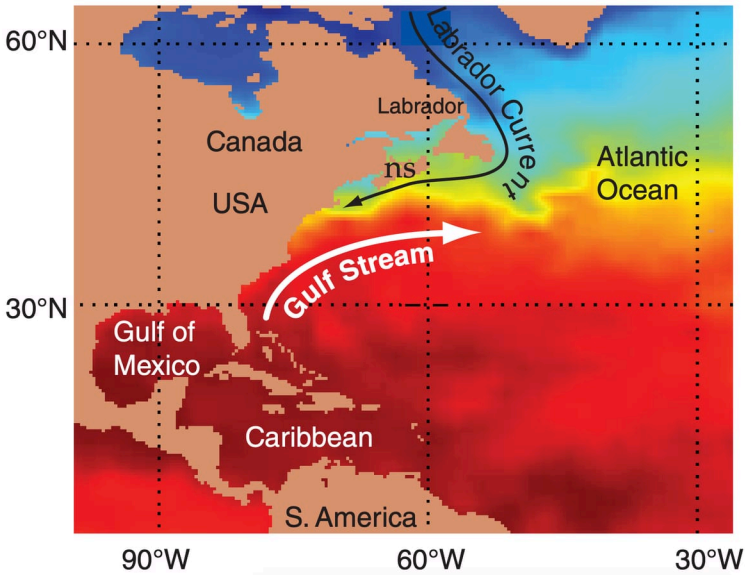


Figure 16.22 Enlargement of part of Fig. 16.8, showing sea-surface temperatures (warm = red fill; cold = blue fill). The Labrador Current (black arrow) and Gulf Stream (white arrow) are shown. ns = Nova Scotia.

Tropical cyclolysis can also occur if the layer of warm ocean water is too shallow, such that tropical cyclone-induced ocean waves stir colder deep water to the surface. Also, if a tropical cyclone crosses the path of a previous tropical cyclone, the ocean surface along this previous path is often colder due to the wave stirring.

16.3.5.3. Interaction with Mid-latitude Lows

When tropical cyclones move into mid-latitudes, they can encounter the larger, more powerful mid-latitude cyclones (Fig. 16.23) that were discussed in a previous chapter. Strong mid-latitude cold fronts can inject cold air into the tropical cyclone, causing the tropical cyclone to die. The fronts and associated jet stream aloft often have strong wind shear, and can rip apart the tropical cyclones. Sometimes tropical cyclones will move into the path of a mid-latitude cyclone, allowing the mid-latitude cyclone to capture some of the tropical cyclone's energy and consume the tropical cyclone as the two systems merge.

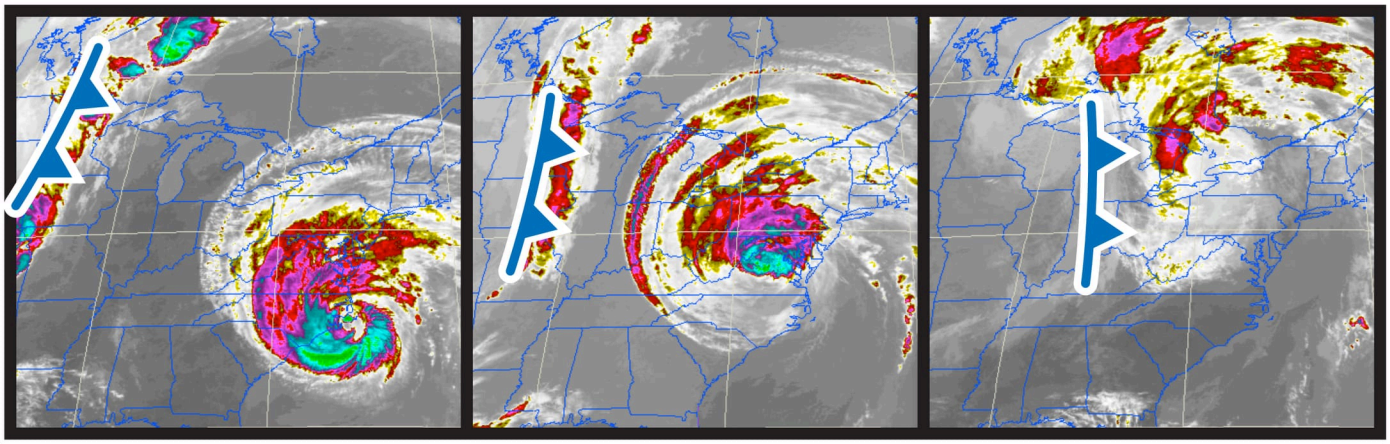


Figure 16.23 Collision of an extratropical cold front with former Hurricane Isabel in Sep 2003. (a) A cold front is over Minnesota, USA, while the hurricane eye makes landfall on the N. Carolina coast. (b) Later. (c) The front and hurricane collide over the Great Lakes. (IR GOES satellite images courtesy of NOAA.)

This page titled 12.2.3: Evolution is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Roland Stull via source content that was edited to the style and standards of the LibreTexts platform.

- 16.3: Evolution by Roland Stull is licensed CC BY-NC-SA 4.0. Original source: https://www.eoas.ubc.ca/books/Practical_Meteorology.

12.2.4: Hazards

16.8.1. Human Population and Zoning

The most important factor causing increased deaths and destruction from tropical cyclones is the increase in **global population**. With population growth, more people live in coastal areas, which are seen as desirable in spite of the threat of tropical cyclones. Even if tropical-cyclone activity were to remain relatively constant, the impacts on humans would increase as population increases.

As more structures are built in vulnerable areas, so increase the property losses caused by tropical cyclone destruction. It also becomes more difficult to evacuate people along inadequate transportation networks. As more people move from farms to cities, there are increased fatalities due to urban flooding caused by heavy tropical-cyclone rainfall.

In highly developed countries, an easy solution would be proper land-use **zoning**. Namely, governments would not let people live and work in threatened areas. Instead, these areas would be used for parks, floodable farmland, wildlife refuges, etc. However, zoning is a political activity that sometimes results in poor decisions in the face of pressure from real-estate developers who want to build more waterfront properties. An example of a questionable decision is the reconstruction of New Orleans, Louisiana, USA, at its original location after being destroyed by hurricane Katrina in 2005.

Also, people are unfortunately encouraged to live in threatened areas because of the existence of hurricane **insurance** and hurricane **disaster relief**. Namely, some individuals choose to make these poor decisions on where to live because they do not have to bear the full costs of reconstruction — instead the cost is borne by all taxpayers.

In less developed, highly populated, low-lying countries such as Bangladesh, an additional problem is an inadequate **warning** system. Even when tropical cyclone tracks are successfully predicted, sometimes the warning does not reach rural people, and often there is inadequate transportation to enable their evacuation. Bangladesh has suffered terribly from tropical cyclones: 500,000 deaths in Nov 1970, over 11,000 deaths in May 1985 , and 150,000 deaths in Apr 1991.

These aspects of tropical-storm hazards are therefore social (cultural, political, religious, etc.). It cannot be assumed that all problems can be ameliorated by technical solutions (more dams; higher levies). Instead, some tough decisions need to be made on zoning, transportation, and population growth.

16.8.2. Storm Surge

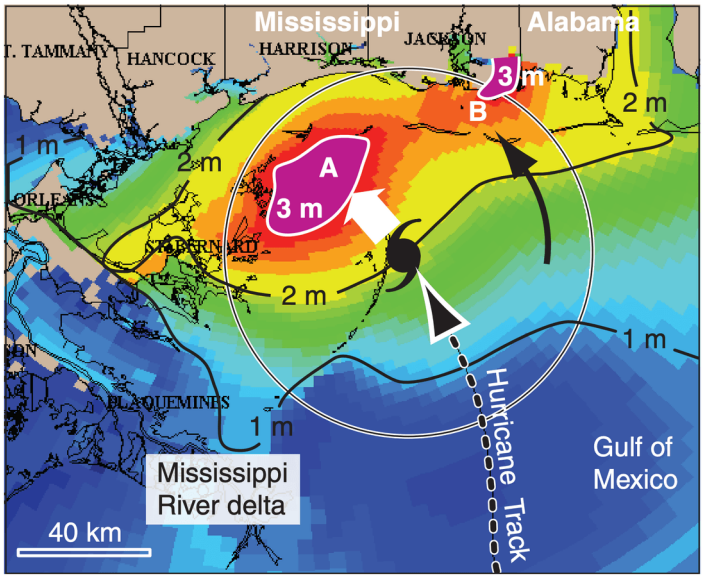


Figure 16.41 Simulated storm surge associated with a hypothetical landfalling hurricane. Color shading and contours indicate rise (m) in sea level. (Modified from a NOAA/HRD figure).

Much of the tropical cyclone-caused damage results from inundation of coastal areas by high seas (Fig. 16.41). The rise in sea level (i.e., the **storm surge**) is caused by the reduced atmospheric pressure in the eye, and by wind blowing the water against the coast to form a large propagating surge called a Kelvin wave. Table 16-7 gives typical stormsurge heights. High tides and high surface-waves can exacerbate the damage.

16.8.2.1. Atmospheric Pressure Head

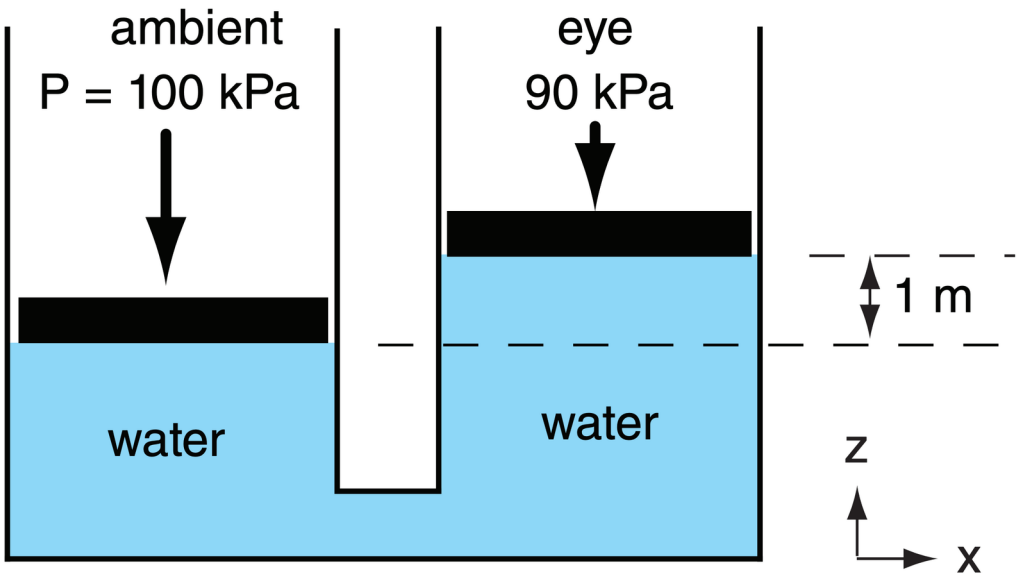


Figure 16.42 Reduced atmospheric pressure in the eye allows sea-level to rise.

In the eye of the tropical cyclone, atmospheric surface pressure is lower than ambient. Hence the force per unit area pushing on the top of the water is less. This allows the water to rise in the eye until the additional head (weight of fluid above) of water compensates for the reduced head of air (Fig. 16.42).

The amount of rise Δz of water in the eye is

$$\Delta z = \frac{\Delta P_{\max}}{\rho_{\text{liq}} \cdot |g|} \tag{16.19}$$

where |g| is gravitational acceleration magnitude (9.8 m s^{-2}), ρ_{liq} = 1025 kg m^{-3} is the density of sea water, and ΔP_{max} is the atmospheric surface pressure difference between the eye and the undisturbed environment.

To good approximation, this is

$$\Delta z \approx a \cdot \Delta P_{\max} \tag{16.20}$$

where a = 0.1 m kPa^{-1}. Thus, in a strong tropical cyclone with eye pressure of 90 kPa (causing ΔP_{max} ≈ 10 kPa), the sea level would rise 1 m.

16.8.2.2. Ekman Transport

Recall from the General Circulation chapter that ocean currents are generated by wind drag on the sea surface. The Ekman spiral describes how the current direction and speed varies with depth. The net **Ekman transport**, accumulated over all depths, is exactly perpendicular to the surface wind direction.

In the Northern Hemisphere, this net transport of water is to the right of the wind, and has magnitude:

$$\frac{Vol}{\Delta t \cdot \Delta y} = \frac{\rho_{\text{air}}}{\rho_{\text{water}}} \cdot \frac{C_D \cdot V^2}{f_c} \tag{16.21}$$

where Vol is the volume of water transported during time interval Δt, Δy is a unit of length of coastline parallel to the mean wind,

V is the wind speed near the surface (actually at 10 m above the surface), C_D is the drag coefficient, f_c is the Coriolis parameter, ρ_{air} = 1.225 kg·m^{-3} is the air density at sea level, and ρ_{water} = 1025 kg·m^{-3} is sea-water density.

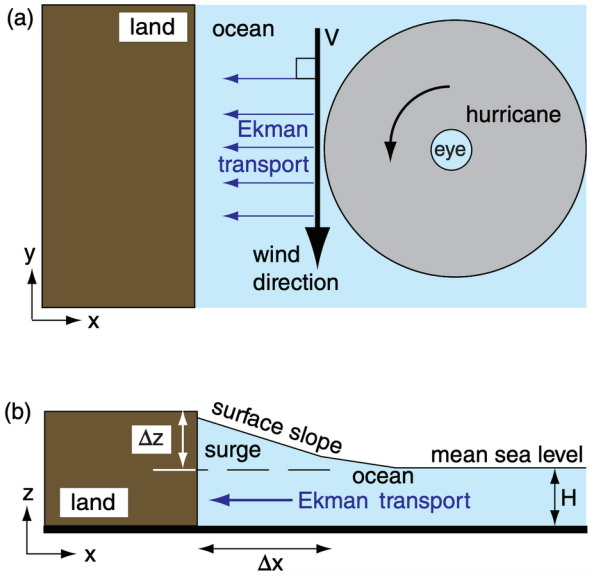


Figure 16.43 Onshore Ekman transport prior to tropical cyclone landfall, in the Northern Hemisphere. (a) Top view. (b) View from south.

As a tropical cyclone approaches the eastern coast of continents, the winds along the front edge of the tropical cyclone are parallel to the coast, from north to south in the Northern Hemisphere. Hence, there is net Ekman transport of water directly toward the shore, where it begins to pile up, creating a storm surge (Fig. 16.43).

If the tropical cyclone were to hover just offshore for sufficient time to allow a steady-state condition to develop, then the Ekman transport toward the shore would be balanced by downslope sloshing of the surge. The surge slope Δz/Δx for that hypothetical equilibrium is:

$$\frac{\Delta z}{\Delta x} \approx \frac{\rho_{air}}{\rho_{water}} \cdot \frac{C_D \cdot V^2}{|g| \cdot H} \tag{16.22}$$

where H is the unperturbed ocean depth (e.g., 50 m) near the coast, and the other variables are the same as for eq. (16.21).

Table 16-7. Storm surge height S and sea-level pressure P_{eye}, from the old Saffir-Simpson classification. CAUTION: Actual storm-surge heights can vary significantly from these typical values. This is one of the reasons why the new Saffir-Simpson scale doesn't use S.

Category	P_{eye} (kPa)	S (m)
1	≥ 98.0	1.2 - 1.6
2	97.9 - 96.5	1.7 - 2.5
3	96.4 - 94.5	2.6 - 3.9
4	94.4 - 92.0	4.0 - 5.5
5	< 92.0	> 5.5

Sample Application

For tropical cyclone-force winds of 40 m s^{-1}, over a continental shelf portion of ocean of depth 50 m, find the volume transport rate and equilibrium surge slope. Assume C_D = 0.01 .

Find the Answer

Given: ρair = 1.225 kg·m^{-3}, ρ_{water} = 1025 kg·m^{-3}, M = 40 m s^{-1}, H = 50 m, C_D = 0.01

Find: Vol/(Δt·Δy) = ? m^2 s^{-1}, and Δz/Δx = ?

Assume: f_c = 0.00005 s^{-1}

Use eq. (16.21):

$$dfracVol\Delta t \cdot \Delta y = \frac{(1.225\mathrm{kg/m^3})}{(1025\mathrm{kg/m^3})} \cdot \frac{(0.01)\cdot(40\mathrm{m/s})^2}{(0.00005\mathrm{s^{-1}})} = \underline{\mathbf{382}\mathrm{m^2 s^{-1}}} \tag{12.2.4.1}$$

Use eq. (16.22):

$$dfrac\Delta z\Delta x \approx \frac{(1.225\mathrm{kg/m^3})}{(1025\mathrm{kg/m^3})} \cdot \frac{(0.01)\cdot(40\mathrm{m/s})^2}{(9.8\mathrm{m/s^2})\cdot(50\mathrm{m})} = \underline{\mathbf{3.9\times10^{-5}}} \tag{12.2.4.2}$$

Check: Units OK. Physics OK.

Exposition: The flow of water is tremendous. As it starts to pile up, the gradient of sea-level begins to drive water away from the surge, so it does not continue growing. The slope corresponds to 4 cm rise per km distance toward the shore. Over 10s to 100s km, the rise along the coast can be significant.

16.8.2.3. Kelvin Wave

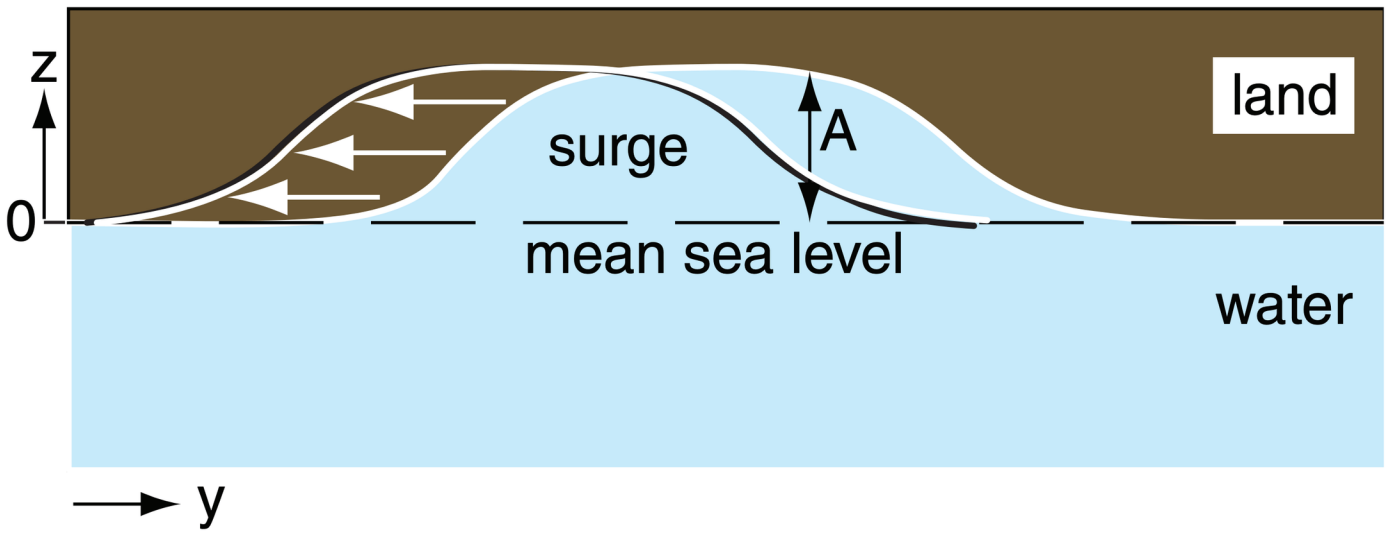


Figure 16.44 The surge, viewed from the east towards vertical cliffs of land, is a Kelvin wave that propagates south. For low-lying coastal areas with no cliffs, the Kelvin-wave storm surge would spread inland and inundate (flood) the coastal areas.

Because the tropical cyclone has finite size, Ekman transport is localized to the region immediately in front of the tropical cyclone. Thus, water is piled higher between the tropical cyclone and the coast than it is further north or south along the coast. When viewed from the East, the surge appears as a long-wavelength wave, called a **Kelvin wave** (Fig. 16.44).

The propagation speed of the wave, called the **phase speed** c, is

$$c = \sqrt{|g|\cdot H} \tag{16.23}$$

which is also known as the **shallow-water wave speed**, where |g| is gravitational acceleration magnitude and H is average water depth. Typical phase speeds are 15 to 30 m s^{-1}. These waves always travel with the coast to their right in the Northern Hemisphere,

so they propagate southward along the east coast of continents, and northward along west coasts.

As the wave propagates south along the East Coast, it will hug the coast and inundate the shore immediately south of the tropical cyclone. Meanwhile, Ekman transport continues to build the surge in the original location. The net result is a continuous surge of high water along the shore that is closest to, and south of, the tropical cyclone. Typical surge depths can be 2 to 10 m at the coast, with extreme values of 13 to 20 m.

Should the tropical cyclone move southward at a speed nearly equal to the Kelvin wave speed, then Ekman pumping would continue to reinforce and build the surge, causing the amplitude of the wave A to grow according to:

$$\frac{\Delta A}{\Delta t} \approx \frac{\rho_{\text{air}}}{\rho_{\text{watter}}} \cdot \frac{C_D \cdot V^2}{\sqrt{|g| \cdot H}} \tag{16.24}$$

where the amplitude A is measured as maximum height of the surge above mean sea level, and the other variables are the same as for eq. (16.21).

Sample Application

Using info from the previous Sample Application, find the Kelvin wave phase speed, and the growth rate if the tropical cyclone follows the wave southward.

Find the Answer

Given: (same as previous example)

Find: c = ? m s^{-1}, ΔA/Δt = ? m s^{-1}

Use eq. (16.23):

$$c = \sqrt{(9.8\text{m/s}^2) \cdot (50\text{m})} = \underline{\mathbf{22.1}\text{ms}^{-1}} \tag{12.2.4.3}$$

Use eq. (16.24):

Misplaced &

Check: Units OK. Physics OK.

Exposition: Tropical cyclones usually turn northward. If tropical cyclones were to translate southward with speed 22.1 m s^{-1}, matching the Kelvin wave speed, then an exceptionally dangerous situation would develop with amplification of the surge by over 3 m h^{-1}.

16.8.3. Surface Wind-waves

Waves are generated on the sea surface by action of the winds. Greater winds acting over longer distances (called **fetch**) for greater time durations can excite higher waves. High waves caused by tropical cyclone-force winds are not only a hazard to shipping, but can batter structures and homes along the coast.

Four coastal hazards of a tropical cyclone are:

- wave scour of the beach under structures,
- wave battering of structures,
- surge flooding, and
- wind damage.

The first two hazards exist only right on the coast, in the beach area. But the inundation (flooding) from the storm surge can reach

10 to 15 km inland from the coast, depending on the slope of the land.

For wind speeds M up to tropical cyclone force, the maximum-possible wave height (for unlimited fetch and duration) can be estimated from:

$$h = h_2 \cdot \left(\frac{M}{M_2}\right)^{3/2} \tag{16.25}$$

where h_2 = 4 m and M_2 = 10 m s^{-1}. Wave heights are plotted in Fig. 16.45.

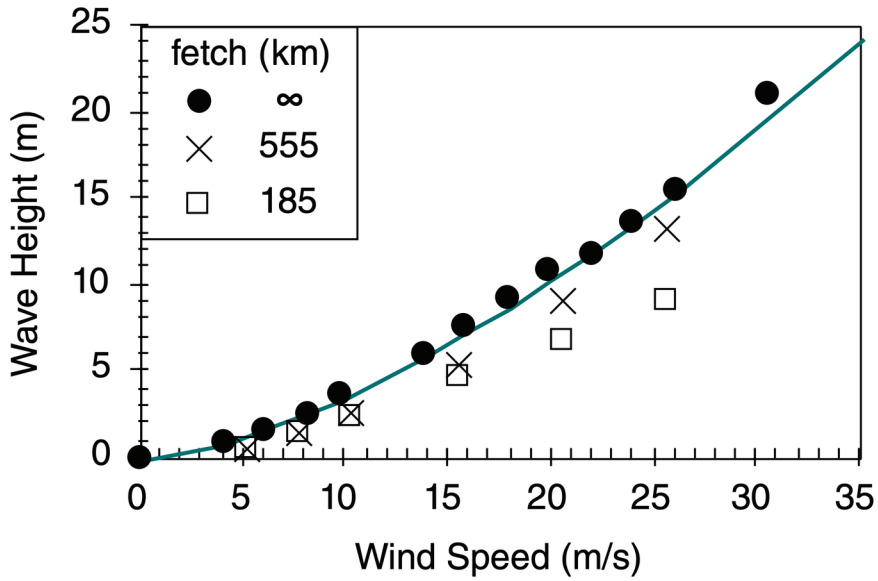


Figure 16.45 Wave height of wind-generated surface waves. Solid line is eq. (16.25) for unlimited fetch and duration. Data points are ocean observations with different fetch.

As winds increase beyond tropical cyclone force, the wave tops become partially chopped off by the winds. Thus, wave height does not continue to increase according to eq. (16.25). For extreme winds of 70 m s^{-1}, the sea surface is somewhat flat, but poorly defined because of the mixture of spray, foam, and chaotic seas that appear greenish white during daytime.

Wavelengths of the wind-waves also increase with wind speed. Average wavelengths λ can be approximated by:

$$\lambda = \lambda_2 \left(\frac{M}{M_2}\right)^{1.8} \tag{16.26}$$

where λ_2 = 35 m, and M_2 = 10 m s^{-1}. Wavelengths are plotted in Fig. 16.46.

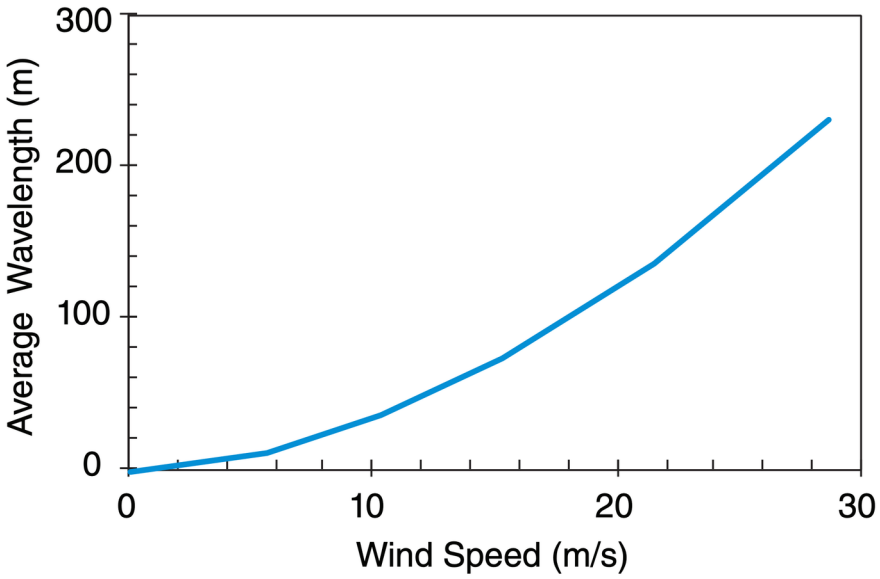


Figure 16.46 Wavelength of wind waves. Solid line is eq. (16.26).

INFO • Beaufort Wind-Scale History

In 1805, Admiral Beaufort of the British Navy devised a system to estimate and report wind speeds based on the amount of canvas sail that a full-rigged frigate could carry. It was updated in 1874, as listed in Table 16-8 for historical interest. Modern descriptors for the Beaufort wind scale are in Table 16-9.

Table 16-8. Legend: B = Beaufort number; D = modern classification; M = wind speed in knots (2 knots ≈ 1 m s^{-1}), S1 = speed through smooth water of a well-conditioned man-of-war carrying all sails un-reefed (full), sailing close to the wind; S2 = un-reefed (full) sails that a well-conditioned manof-war could just carry in chase, sailing close to the wind; S3 = sails that a well-conditioned man-of-war could scarcely bear. [A reefed sail exposes less than its full area to the wind.]

B	D	M (kt)	Deep Sea Criteria
0	Calm	0 - 1	S1 = Becalmed
1	Light Air	1 - 3	S1 = Just sufficient to give steerageway
2	Slight Breeze	4 - 6	S1 = 1 - 2 knots
3	Gentle Breeze	7 - 10	S1 = 3 - 4 knots
4	Moderate Breeze	11 - 16	S1 = 5 - 6 knots
5	Fresh Breeze	17 - 21	S2 = Royals, etc.
6	Strong Breeze	22 - 27	S2 = Topgallant sails
7	High Wind	28 - 33	S2 = Topsails, jib, etc.
8	Gale	34 - 40	S2 = Reefed upper topsails and courses
9	Strong Gale	41 - 48	S2 = Lower topsails and courses
10	Whole Gale	49 - 55	S3 = lower main topsail and reefed foresail
11	Storm	56 - 65	S3 = storm staysails
12	Hurricane	> 65	S3 = no canvas

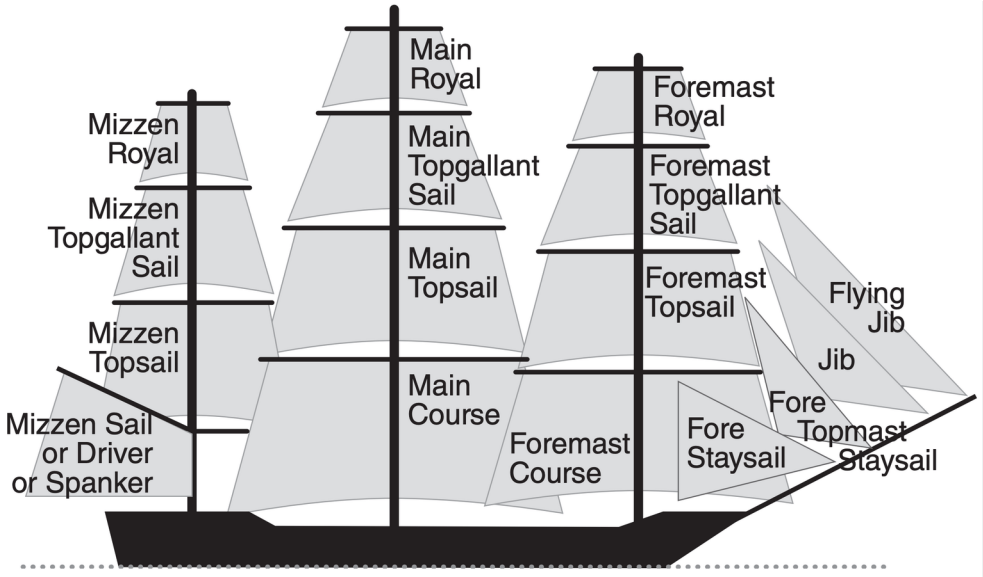


Figure 16.d. Frigate sails.

The longest wavelength waves are called swell, and can propagate large distances, such as across whole oceans. Hence, a tropical cyclone in the middle tropical Atlantic can cause large surf in Florida well before the storm reaches the coast.

Wind and waves as affect mariners are classified according to the Beaufort scale. The INFO Box with Table 16-8 gives a historical description of the Beaufort scale. A modern description is in Table 16-9.

Table 16-9. Beaufort wind scale. B = Beaufort Number. (See INFO Box for historical info.)

B	Description	Wind Speed (km/h)	Wind Speed (m/s)	Wave Height (m)	Sea Conditions (in deep ocean)	Land Conditions
0	Calm	< 1	< 0.3	0	Flat.	Calm. Smoke rises vertically.
1	Light Air	1 - 5	0.3 - 1.5	0 -00.2	Ripples without crests.	Wind motion visible in smoke.
2	Light Breeze	6 - 11	1.5 - 3.3	0.2 - 0.5	Small wavelets. Crests of glassy appearance, not breaking.	Wind felt on exposed skin. Leaves rustle.
3	Gentle Breeze	12 - 19	3.3 - 5.5	0.5 - 1	Large wavelets. Crests begin to break; scattered whitecaps.	Leaves and smaller twigs in constant motion.
4	Moderate Breeze	20 - 28	5.5 - 8.0	1 - 2	Small waves with breaking crests. Fairly frequent whitecaps.	Dust and loose paper raised. Small branches begin to move.
5	Fresh Breeze	29 - 38	8.0 - 11	2 - 3	Moderate waves of some length. Many whitecaps. Small amounts of spray.	Branches of a moderate size move. Small trees begin to sway.
6	Strong Breeze	39 - 49	11 - 14	3 - 4	Long waves begin to form. White foam crests are very frequent. Some airborne spray is present.	Large branches in motion. Whistling heard in overhead wires. Umbrella use becomes difficult. Empty plastic garbage cans tip over.
7	High Wind, Moderate Gale, Near Gale	50 - 61	14 - 17	4 - 5.5	Sea heaps up. Some foam from breaking waves is blown into streaks along wind direction. Moderate amounts of airborne spray.	Whole trees in motion. Effort needed to walk against the wind. Swaying of skyscrapers may be felt, especially by people on upper floors.

8	Gale, Fresh Gale	62 - 74	17 - 20	5.5 - 7.5	Moderately high waves with breaking crests forming spindrift. Well-marked streaks of foam are blown along wind direction. Considerable airborne spray.	Some twigs broken from trees. Cars veer on road. Progress on foot is seriously impeded.
9	Strong Gale	75 - 88	21 - 24	7 - 10	High waves whose crests sometimes roll over. Dense foam is blown along wind direction. Large amounts of airborne spray may begin to reduce visibility.	Some branches break off trees, and some small trees blow over. Construction/ temporary signs and barricades blow over. Damage to circus tents and canopies.
10	Storm, Whole Gale	89 - 102	25 - 28	9 - 12.5	Very high waves with overhanging crests. Large patches of foam from wave crests give the sea a white appearance. Considerable tumbling of waves with heavy impact. Large amounts of airborne spray reduce visibility	Trees are broken off or uprooted, saplings bent and deformed. Poorly attached asphalt shingles and shingles in poor condition peel off roofs.
11	Violent Storm	103 - 117	29 - 32	11.5 - 16	Exceptionally high waves. Very large patches of foam, driven before the wind, cover much of the sea surface. Very large amounts of airborne spray severely reduce visibility.	Widespread vegetation damage. Many roofing surfaces are damaged; asphalt tiles that have curled up and/or fractured due to age may break away completely.

12	Hurricane	≥ 118	≥ 33	≥14	Huge waves. Sea is completely white with foam and spray. Air is filled with driving spray, greatly reducing visibility.	Very widespread damage to vegetation. Some windows may break; mobile homes and poorly constructed sheds and barns are damaged. Debris may be hurled about.

Tides, storm surges, and wind waves are additive (Fig. 16.47). Namely, if the storm surge and high waves happen to occur during high tide according to routine astronomic tide tables, then the coastal destruction is likely to be greatest.

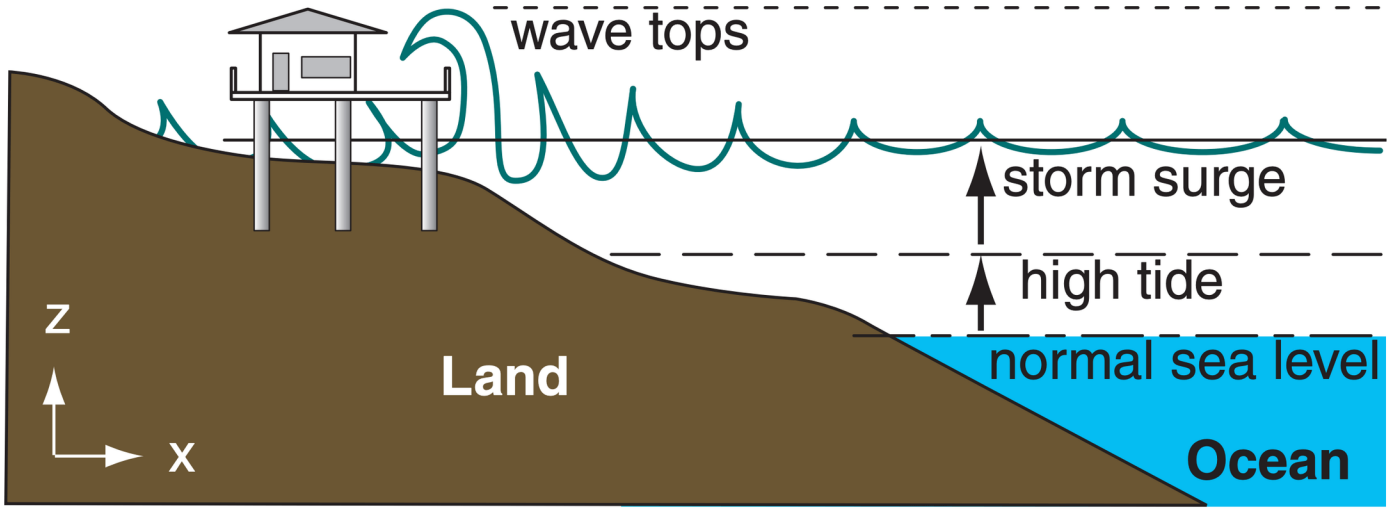


Figure 16.47 Tides, storm surge, and waves are additive.

For safety, houses at coasts threatened with storm surges and tsunami are usually built on top of concrete or steel piles. These piles are driven very deep into the land to survive wave scour and erosion of the land. While the floor deck is above expected storm-surge plus high tide levels in this figure, the waves can still batter and damage the structure.

Sample Application

Find the maximum possible wave height (assuming unlimited fetch) and wavelength for tropical cyclone force winds of 35 m s^{-1}.

Find the Answer

Given: M = 35 m s^{-1}

Find: h = ? m, λ = ? m

Use eq. (16.25):

$$h = (4m) \cdot \left(\frac{35m/s}{10m/s} \right)^{3/2} = \underline{\mathbf{26.2m}} \tag{12.2.4.4}$$

Use eq. (16.26):

$$lambda = (35\mathrm{m}) \cdot \left(\frac{35\mathrm{m/s}}{10\mathrm{m/s}} \right)^{1.8} = \underline{\mathbf{334m}} \tag{12.2.4.5}$$

Check: Units OK. Physics OK. Agrees with Figs.

Exposition: Wavelengths are much longer than wave heights. Thus, wave slopes are small — less that 1/10. Only when these waves reach shore does wave slope grow until the waves break as surf.

16.8.4. Inland Flooding

In developed countries such as the USA, stormsurge warning and evacuation systems are increasingly successful in saving lives of coastal dwellers. However, inland flooding due to heavy rain from decaying tropical storms is increasingly fatal — causing roughly 60% of the tropical cyclone-related deaths in the USA during the past 30 years.

Streams and storm drains overflow, trapping people in cars and on roof tops. For this reason, people should not be complacent about former tropical cyclones reaching their inland homes, because these dying tropical cyclones contain so much tropical moisture that they can cause record-setting rainfalls.

The inland flooding hazard can affect people hundreds of kilometers from the coast. Of the 56 people who died in Hurricane Floyd (1999) in the eastern USA, 50 drowned in inland floods caused by heavy rains. Tropical storm Alberto dropped 53 cm of rain over Americus, Georgia, where 33 people drowned in 1994. Over 200 people drowned in Pennsylvania, New York, and New England from Hurricane Diane in 1955.

More recently, many unnecessary drowning fatalities have been caused by people driving into water of unknown depth covering the road. For people comfortable in driving their usual roads day after day, many find it hard to believe that these roads can become impassable. They unknowingly drive into deep water, causing the engine to stall and the car to stop in the middle of the water. If the water continues to rise, the car and passengers can be carried away. An easy solution is to approach each flooded road with caution, and be prepared to interrupt your journey and wait until the flood waters subside.

16.8.5. Thunderstorms, Lightning & Tornado Outbreaks

Because tropical cyclones are made of thunderstorms, they contain all the same hazards as thunderstorms. These include lightning, downbursts, gust fronts, downpours of rain, and tornadoes (see the Thunderstorm chapter for hazard details). Lightning is somewhat infrequent in the eyewall (about 12 cloud-to-ground strikes per hour), compared to about 1000 strikes per hour for midlatitude MCSs.

Tropical cyclones can cause **tornado outbreaks**. For example, Hurricanes Cindy and Katrina each spawned 44 tornadoes in the USA in 2005. Hurricane Rita spawned 101 tornadoes in Sep 2005. Hurricanes Frances and Ivan spawned 103 and 127 tornadoes, respectively, in Sep 2004.

Most tornadoes are weak (≤ EF2), and occur in the right front quadrant of tropical cyclones in the N. Hemisphere. Although some form near the eyewall, most tornadoes form in the thunderstorms of outer rain bands 80 to 480 km from the cyclone center.

This page titled 12.2.4: Hazards is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Roland Stull via source content that was edited to the style and standards of the LibreTexts platform.

- 16.8: Hazards by Roland Stull is licensed CC BY-NC-SA 4.0. Original source: https://www.eoas.ubc.ca/books/Practical_Meteorology.

12.3: Thunderstorms and Severe Weather

This section examines the development of thunderstorms and the severe weather they can produce. It explains how atmospheric instability, moisture, and rising air contribute to thunderstorm formation and how these storms can generate heavy precipitation, hail, lightning, and tornadoes. The section also considers the hazards associated with severe thunderstorms and emphasizes practical steps for staying safe when severe weather occurs.

This page titled 12.3: Thunderstorms and Severe Weather is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by LibreTexts.

12.3.1: Thunderstorms

The primary energy that drives thunderstorms is the conversion of moist air into clouds and precipitation, which releases significant amounts of latent heat in the condensation process. Over land, thunderstorms occur most frequently in the late afternoon and early evening. Storms are most likely to occur soon after the warmest surface temperatures, which helps to give warm air parcels the initial buoyancy they need to begin to rise. Thunderstorms can last anywhere from less than an hour to more than 12 hours. Because of their rapid growth, relatively small (mesoscale) size, and sensitivity to environmental conditions, it is difficult to predict the exact time and location of a storm. Typically a forecast will say there is a chance of thunderstorms over an area—meaning that although conditions are conducive for a storm, the forecasters don't know exactly where or when it will occur.

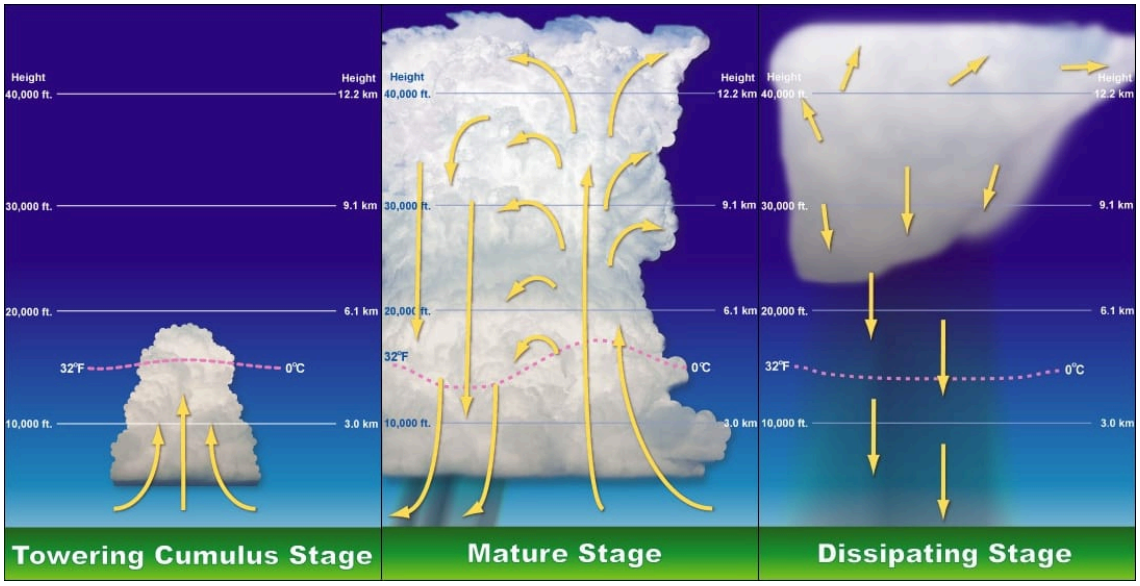
Thunderstorm Life Cycle

The three main stages of a thunderstorms are as follows:

1. Developing stage
2. Mature stage
3. Dissipating stage

In the **developing stage**, rising air called "**updrafts**" are dominant in this stage and towering cumulus clouds form. This type of cloud is also known as a cumulus congestus. Next is the **mature stage** where the towering cumulus becomes a cumulonimbus cloud featuring both updrafts and "**downdrafts**" or sinking air. Downdrafts develop as a result of water droplets falling to the ground as precipitation (e.g., snow, rain, hail). The falling drops drag air downward as they fall, creating a downward current of air in the cumulonimbus cloud. Finally, in the **dissipating stage** downdrafts are dominant and cut off the updrafts needed for a cumulonimbus cloud to form and sustain itself. As a result, the cloud dissipates.

If a cumulonimbus cloud has thunder and lightning, it is called a **thunderstorm**. Lightning is caused by a charge separation in clouds. For this charge separation to occur, a cloud must have ice crystals inside of it, which is why thunderstorms are always mixed-phase or ice-phase clouds.



Formation and dissipation of a thunderstorm (Public Domain).

Airmass Thunderstorm

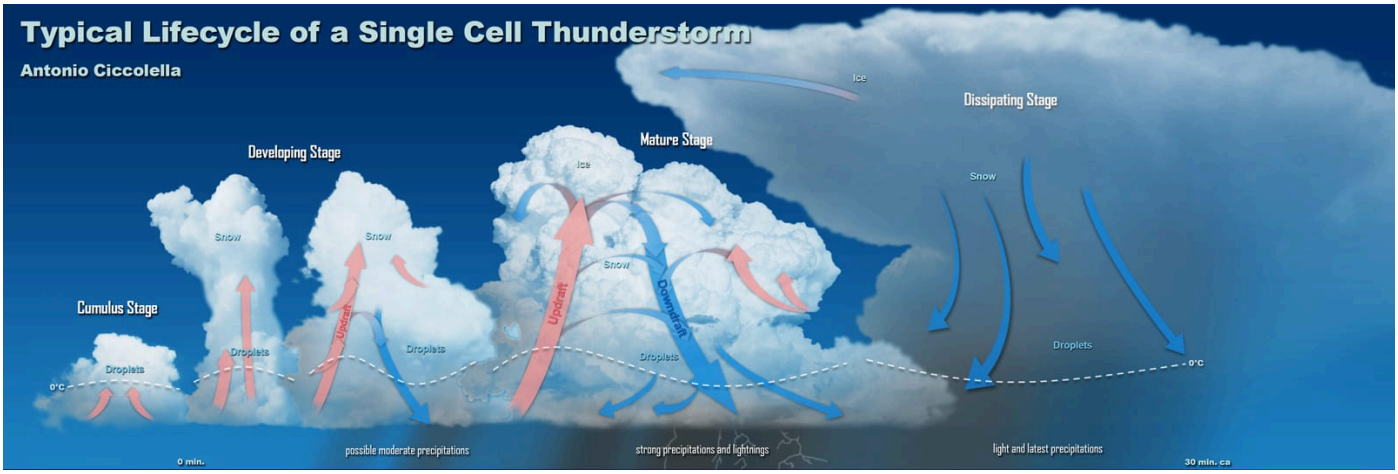
The thunderstorm described above is called an airmass thunderstorm. In general, the ingredients necessary for a thunderstorm to occur are **warm moist air**, some type of **instability**, and a **trigger** that can initiate the storm system. Airmass thunderstorms

typically occur in environments with similar properties and little wind shear. Airmass thunderstorms are the most benign types of thunderstorms and typically short-lived, lasting less than a few hours.

The reason airmass thunderstorms are so short-lived and benign is because the storm shuts off its ability to maintain itself. This is due to two primary reasons:

1. Below cloud base the falling rain evaporates, cooling the air in the boundary layer. The storm deprives itself of the heat and moisture in the boundary-layer air. Without the warm buoyant air to drive further convection, the storm loses its strength and dissipates.
2. Rain falls within the updraft of the storm, reducing its strength, again causing dissipation.

Another diagram of an airmass thunderstorm or single cell thunderstorm is shown below. The initial small cumulus cloud grows to a cumulus congestus, which has mostly updrafts. It then grows further into a mature cumulonimbus cloud with well defined updrafts and downdrafts. The final stage is the dissipating stage, which has mostly downdrafts.



Another diagram showing the typical cycle of a thunderstorm (CC BY-SA 4.0).

The trigger for a thunderstorm can be a warm humid air mass heated from the bottom by daytime solar heating or forced lifting by terrain (orographic thunderstorms). Collision of airmasses can be another trigger, as you might find along a warm front, cold front, or dryline.

Thunderstorm Features

The main features of a mature thunderstorm includes updrafts, downdrafts, overshooting tops, and an anvil. An **overshooting top** occurs at the top of the thunderstorm. Overshooting tops are dome-shaped as a result of strong updrafts in intense thunderstorms reaching the tropopause and pushing upward into the stable region. However, ultimately, because the tropopause is so stable, the air gets pushed back downward and spread laterally. This lateral spreading forms the cloud **anvil** at the top of the thunderstorm along the tropopause boundary.

Severe Thunderstorm

While basic airmass thunderstorms typically consist of one “cell”, severe storms often contain multiple cells, consisting of more than one rising thermal connected with the same storm system. In a multicell storm, each cell is typically at a different life cycle stage and they continue on one from another.

In order for a thunderstorm to become severe, one important additional ingredient is necessary. In addition to warm moist air, some sort of instability, and a trigger, severe thunderstorms need **wind shear**. Wind shear is defined as a change in wind speed or direction with altitude. The reason wind shear is important for severe storms is easy; wind shear helps to tilt a storm such that the updraft and downdraft are displaced from one another. Without the competition between upward moving air and downward moving air, severe thunderstorms can strengthen and last much much longer than airmass thunderstorms. Severe thunderstorms come in many different forms. A few will be discussed below.

Severe Thunderstorm Types

The three main types of thunderstorms include squall lines, mesoscale convective complexes (MCCs), and supercells.

Squall Line

A **squall line** is a linear propagating severe thunderstorm. Squall lines consist of a long, narrow line of merged thunderstorms, often triggered by a cold front, dry line, or gust front. Squall lines can be many hundreds of kilometers long, but their width is usually much shorter, between 15 and 100 km. They can sustain themselves from several hours to several days.

A squall line continues propagating with the help of a gust front. When rain falls into the drier environment below the cloud base, it evaporates, cooling the air. This cool air continues to move downward, sometimes with strong force and reinforcement from a continuous downdraft. When the cool downdraft hits the surface of Earth, it spreads laterally, literally the same as a density current. Because it is relatively cool, it acts like a mini cold front, pushing warmer air up over it as it moves along. This upward push can initiate convection, and continue fueling the larger storm system.

Mesoscale Convective Complex

A mesoscale convective complex (MCC) is a type of severe storm that has a cloud shield (anvil) with a diameter of at least 350 km, elliptical or circular shape, and lasts between 6 and 12 hours. MCCs are huge storms that occur multiple times per year especially in the central United States. They are truly impressive masses of convection.

Supercell

A **supercell** is a special type of severe storm that has rotation. Supercells form when the environment has **directional wind shear** causing updrafts within a cloud to gain vorticity as they rise. While all severe storms have a structure that separates the updraft from the downdraft, supercell storms have a very clearly defined structure, which allows them to sometimes form tornados. As shown in the below image, supercells have additional features like a **wall cloud** and a **flanking line**. The wall cloud exists beneath the cloud base and near the border between the downdraft of cold air and the lower-level inflow of warm humid air. The flanking line is the region of towering cumulus that indicates extensive updrafts ahead of intense thunderstorms.

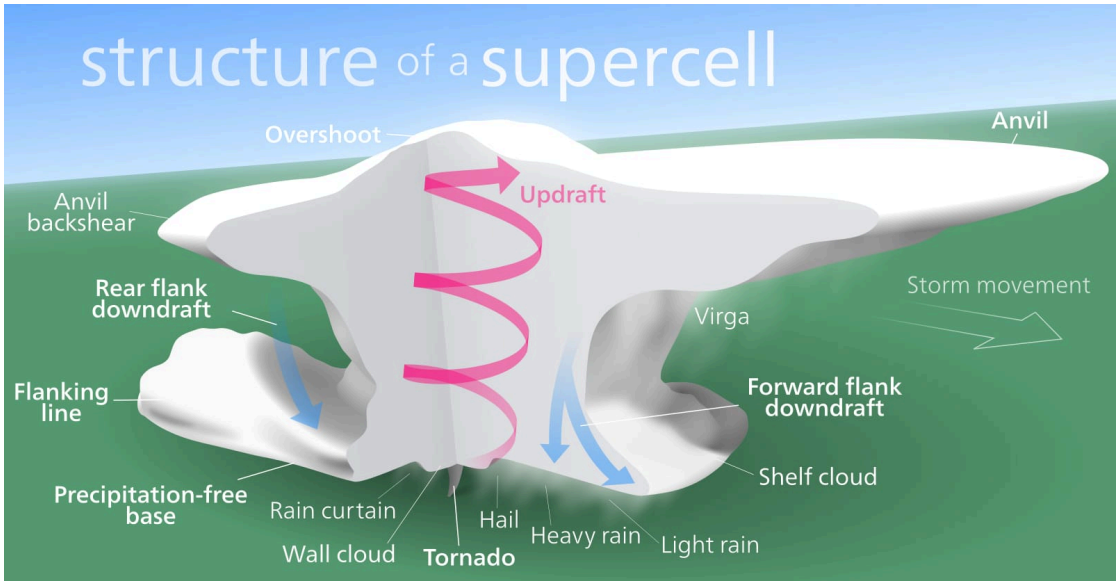
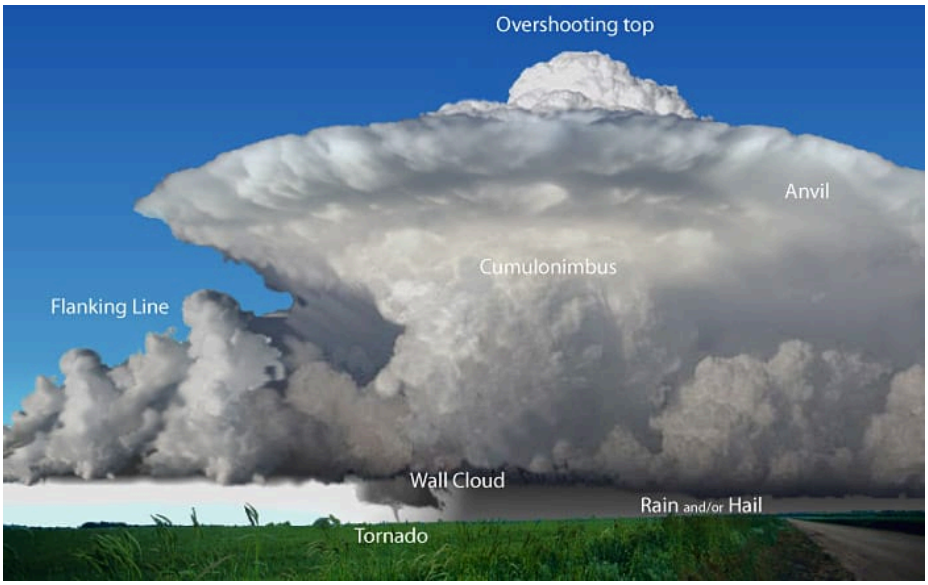


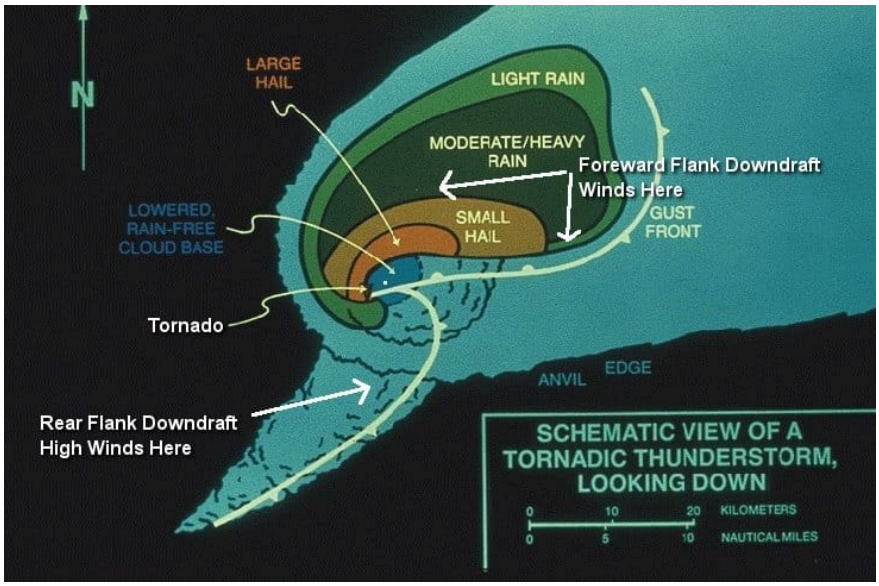
Diagram of a supercell thunderstorm with an anvil shape (CC BY-SA 3.0).

Here is a diagramed image of an actual supercell thunderstorm that created a tornado. Remember that supercells are rotating storm systems that can produce tornados, but they don't always. In fact, very few supercells produce tornados, but the ones that do get the most attention!



An image of a supercell cumulonimbus cloud (Public Domain).

Here is one last image showing the structure of a supercell from the top view. The image shows the location of the gust front, the location of various precipitation types, and the location of downdrafts within the storm system.



A top view diagram of a supercell thunderstorm (Public Domain).

Thunderstorm Formation

Let's review. The ingredients needed for thunderstorm formation include high humidity, conditional instability, and a trigger that initiates rising air. A symmetric short-lived storm is called an airmass thunderstorm. When we add wind shear to an airmass thunderstorm, a severe thunderstorm can result. We discussed three types of severe thunderstorms including squall lines, MCCs, and supercells. Supercells are a type of severe storm with rotation resulting from directional wind shear in the environment.

High humidity in the atmospheric boundary layer is required for thunderstorms to occur. When water vapor condenses, latent heat is released. Latent heat is the primary energy source for thunderstorms. The higher the humidity, the more latent heat is released and the stronger the thunderstorm becomes.

Instability is necessary for convection to occur. The best case scenario is conditional instability, which often occurs when cold air lies above warm moist air capped by a temperature inversion. Cold air in the upper troposphere gives greater buoyancy to the

updraft of warmer air from below and, therefore, a greater chance for strong thunderstorms to occur.

Wind shear is the change in wind speed and or wind direction with height. In an environment with wind but without wind shear, thunderstorms would last between 15 minutes and 1 hour as the thunderstorm and boundary-layer air would remain together. In this scenario, an airmass thunderstorm dissipates after it deprives itself of heat and moisture in the boundary-layer air. Strong wind shear pushes thunderstorms away from the depleted boundary-layer air and into areas with warm, humid boundary-layer air. Thus, strong wind shear helps thunderstorms sustain themselves for longer durations and to grow stronger.

A trigger refers to any process that forces an air parcel to rise through a cap of stable air and result in thunderstorm development. Triggers can be caused by frontal lifting, orographic effects, or surface heating.

This page titled 12.3.1: Thunderstorms is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by LibreTexts.

12.3.2: Atmospheric Instability and Thunderstorms

Because the potential for thunderstorms to develop depends on atmospheric stability and layering, atmospheric soundings (e.g., Skew-T log-P) are used by meteorologists to help forecast storms. The soundings receive their data from the rawinsonde balloon launches, aircraft observations, dropsondes, satellites, or other meteorological data sources.

In atmospheric soundings, there are several labels indicating atmospheric stability. The labels are lifted condensation level (LCL), level of free convection (LFC), equilibrium level (EL), convective available potential energy (CAPE), and convective inhibition (CIN).

Lifted Condensation Level (LCL) was discussed in previous chapters. It is the altitude where the temperature cools to the dew point temperature, resulting in saturation and condensation. The LCL is the location where the cloud base of a thunderstorm develops.

Level of Free Convection (LFC) is the height at which the environmental temperature rate decreases faster than the air parcel's moist adiabatic lapse rate. This leads to atmospheric instability. In an atmospheric sounding, the LFC can be found where the moist adiabatic lapse rate of the rising parcel returns to the buoyant or warmer side of the environmental temperature.

Equilibrium Level (EL) is the height in the atmosphere where the temperature of the rising air parcel is the same as the temperature of its surroundings. The EL caps the atmospheric instability. This is where the moist adiabatic lapse rate returns to the negatively buoyant colder side of the environmental temperature. This is also where the anvil top of the thunderstorm is typically located.

Convective Inhibition (CIN) is the amount of energy that prevents a rising air parcel from reaching the level of free convection. On a thermodynamic diagram (Skew-T Log-P), it is the negative area between the environmental lapse rate and the air parcel lapse rate. CIN must be overcome for CAPE to be realized.

Convective Available Potential Energy (CAPE) is the amount of energy an air parcel would have if lifted vertically through the atmosphere over a particular distance. It is an indicator of atmospheric instability. On a thermodynamic diagram (Skew-T Log-P), CAPE can be found by the positive area between the environmental lapse rate and the air parcel lapse rate. It is an integrated measure of the total amount of buoyancy available to a rising air parcel.

CAPE can be used to estimate the maximum updraft velocity in thunderstorms.

$$\text{Maximum Updraft Velocity } = \sqrt{2 \cdot CAPE}$$

While the above equation gives a good estimate, the updraft velocity equation typically gives an unrealistically high value as it ignores many important processes. For example, dry air entrainment, liquid-water loading, and frictional drag are ignored. Observational studies find that the typical updraft velocity is roughly half of the maximum updraft velocity value.

$$\text{Likely Updraft Velocity } = \frac{\text{Maximum Updraft Velocity}}{2}$$

Thunderstorms are important features of Earth's atmospheric system. It is safe to say that there is always a thunderstorm occurring somewhere on Earth at all times. In many places, thunderstorms provide needed rain and are an important piece of the hydrological cycle. However, thunderstorms can also be hazardous and impacts will be discussed in the following chapter.

Chapter 14: Questions to Consider

1. Label the typical lifecycle of a single cell thunderstorm:

Query 12.3.2.1

This interactive element is not available in this format. Visit the link or scan the QR code to access it online.

View this Interactive Content online

2. Describe the importance of updraft separation from the downdraft and precipitation portion of a thunderstorm.

Query 12.3.2.2

3. *This interactive element is not available in this format. Visit the link or scan the QR code to access it online.*

View this Interactive Content online

Query 12.3.2.3

4. *This interactive element is not available in this format. Visit the link or scan the QR code to access it online.*

View this Interactive Content online

5. With a CAPE of 1280 J/kg, calculate the likely updraft velocity.
6. Describe the importance of wind shear for thunderstorms.

Selected Practice Question Answers:

Query 12.3.2.4

This interactive element is not available in this format. Visit the link or scan the QR code to access it online.

View this Interactive Content online

This page titled 12.3.2: Atmospheric Instability and Thunderstorms is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by LibreTexts.

12.3.3: Precipitation and Hail

Thunderstorms consist of deep convective clouds that can produce large raindrops (2–8 mm diameter) within 5–10 km wide rain shafts that move horizontally across the ground with precipitation that last from 1–20 minutes over a fixed point. The rainfall rate may be heavy, as much as 10 to 1000 mm of rain per hour. The below photo depicts heavy rain shafts produced by a supercell thunderstorm in Montana.



A supercell thunderstorm in Montana (CC BY 2.0).

The cloud tops of thunderstorms extend far up in the troposphere. In the upper parts of the clouds, ice-phase processes occur. Throughout the cloud, different sizes of ice crystals and water droplets exist, and the heavier hydrometeors fall faster than the smaller ones, sometimes colliding and collecting each other. If these heavier ice particles fall through regions of supercooled droplets (liquid water droplets below freezing), they may grow through a process called **riming**. The liquid water droplets instantly freeze and adhere on contact to the outside surface of falling ice particles. This forms snow pellets called **graupel**, which are less than 5 mm in diameter. Fallen graupel is pictured below.



Graupel pellets on a car roof (CC BY 3.0).

Alternatively, smaller ice crystals that fall just below the 0°C level may partially melt and stick to other partially melted ice crystals through collision, causing them to grow through **aggregation**. These snow aggregates can grow as large as 1 cm in diameter. Snow aggregates and graupel can sometimes reach the ground still frozen or partially frozen, even in the summer, if they fall within cooler, saturated downdrafts. More frequently, these larger ice particles will melt completely into large raindrops before hitting the ground.

Hailstones are irregularly shaped balls of ice larger than 0.5 cm in diameter that are produced in a cumulonimbus cloud—particularly in intense, severe thunderstorms. Hail is formed when graupel or other particles act as embryos that grow through the accretion of supercooled liquid droplets in a cloud. One raindrop takes about one million cloud droplets to form, but a single hailstone may take as many as 10 billion cloud droplets to form. A golf ball-sized hailstone, as pictured below, must travel through the cloud for 5–10 minutes to be able to grow to this size. Strong, violent updrafts within the storm are able to carry smaller ice particles far above the freezing level where collisions with supercooled droplets allow them to grow into hailstones. Rotating updrafts, such as those found in a supercell, are capable of sweeping hail horizontally through the storm, which may allow them to grow much larger. As these growing ice particles pass through areas of the cloud with high liquid water content, extra coatings of ice form around them. Larger hailstones may ascend very slowly in a strong updraft, almost "floating" in place as many supercooled water droplets continue to collide into them. When the hailstone is carried away from the updraft, or if it gets too heavy, it will fall because it is no longer supported by the rising air.

Just below the cloud, the hailstones will begin to melt in the warmer air. Small hail might completely melt before hitting the ground, but in the strong updrafts of severe thunderstorms, hailstones can grow large enough to reach the surface before melting. Notice in the figure below that the hailstone has concentric layers inside, similar to tree rings. The concentric layers alternate clear and opaque ice. This is related to the path that the hailstone takes when it passes through the cloud, as well as the liquid water content of the different areas the hailstone passes through. In areas with low liquid water content (dry growth regime), supercooled water droplets freeze onto the hailstone instantly, which produces the milky opaque ice that contains many air bubbles. In areas with high liquid water content (wet growth regime), supercooled water droplets collect very rapidly onto the hailstone. The transition from liquid to ice releases latent heat, which keeps the surface temperature of the hailstone at 0°C. As a result, supercooled droplets will form a coat of water around the stone rather than freezing instantly. This allows porous areas to be filled in as the water coat freezes slowly and air bubbles escape from the surface. This leaves a layer of clear ice around the hailstone.



Golf ball sized hail cut in half so the inside concentric layers can be viewed (Public Domain).

Large hail can cause severe damage to crops, vegetation, aircraft, roofs, windows, and create a traffic hazard if hailstones accumulate on a roadway. Damage can be greater if the strong winds in a thunderstorm cause hailstones to move horizontally as they fall. Larger hailstones have higher terminal fall velocities and the potential to reach speeds higher than 50 $m \cdot s^{-1}$ (112 mph).

This page titled 12.3.3: Precipitation and Hail is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by LibreTexts.

12.3.4: Lightning

Lightning is caused by a discharge of electricity that typically occurs in mature thunderstorms. Lightning can strike within the same cloud, from one cloud to another, from a cloud to open air, or from cloud to ground. Only about 20% of lightning strikes are cloud-to-ground (CG) strikes, while the majority of strikes occur within the cloud (CC). As a lightning stroke travels through the atmosphere, it heats the surrounding air to a staggering 30,000°C, which is roughly five times hotter than the surface of the sun. This incredible heating causes a sudden, explosive expansion of the air, generating a shock wave in all directions from the flash. This booms outward in the form of a sound wave called **thunder**.

Due to the speed that light travels, we see lightning instantly as it happens. Thunder travels much more slowly at about 330 $m \cdot s^{-1}$ so it takes much longer to reach our ears. You can estimate how distant a lightning stroke is by counting the seconds after you see the flash until you hear the thunder. It takes sound about 5 seconds to travel one mile. If you hear thunder roughly 10 seconds after you see the lightning flash, then the strike occurred about 2 miles away. However, if the lightning is too far, you may not hear any thunder, because sound waves attenuate and bend as they travel through the atmosphere, especially through layers of differing air temperature.



Lightning during a desert storm (CC 0).

Why do we have lightning? During ordinary fair weather conditions, the earth's surface has negative charge, while the upper atmosphere is positively charged. Lightning requires differing regions of opposite charge inside a cumulonimbus cloud. There are several theories as to how this separation of charge may come about.

One theory is that clouds can become electrified when graupel and hailstones fall through areas of supercooled droplets, causing the droplets to freeze and release latent heat. This release of latent heat keeps the hailstone surface warmer than the surrounding ice particles. A net transfer of positive ions (molecules that carry a charge) occurs from the warmer hailstone to the colder surrounding ice crystals when they come in contact. Therefore, the larger, warmer hailstone becomes negatively charged and the smaller, cooler ice crystal becomes positively charged. This also happens when supercooled liquid droplets freeze when they come into contact with a warmer hailstone and little fragments of positively charged ice break off. These tiny positively-charged particles are easily swept to the upper regions of the cloud through updrafts. The larger, heavier, negatively-charged hailstones tend to remain suspended at the same level in the updraft or fall to the lower parts of the cloud. This causes the colder upper regions of the cloud to have a net positive charge, while the middle level of the cloud has negative charge. The lowest portion of the cloud generally has a mixture of negative and positive charges, with a positively charged region occasionally present in precipitation bands near the melting level.

In short, thunderstorms must have ice phase processes to produce thunder and lightning—lightning is produced from a discharge of electricity and thunder is simply produced as a result of the lightning.

This page titled 12.3.4: Lightning is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by LibreTexts.

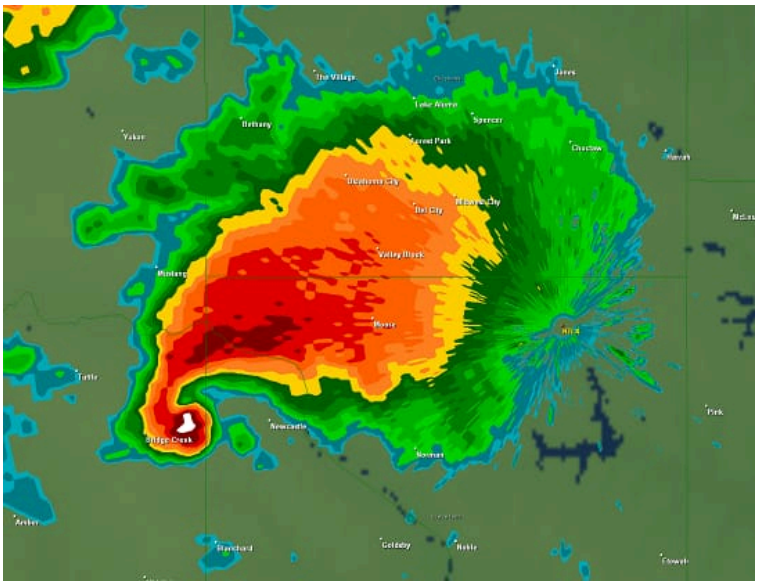
12.3.5: Tornadoes

Recall from the prior chapter that supercell storms can produce tornados. A supercell is a severe thunderstorm with rotation. Typically the rotation of a supercell thunderstorm comes from directional wind shear in the environment. Occasionally a supercell storm produces a tornado, a region of rapidly rotating air in cyclostrophic balance. Recall from Chapter 10 that cyclostrophic balance is a balance between the pressure gradient force and the centrifugal force. As the rotating air motion in a supercell is vertically stretched, it narrows and speeds up. To picture this, imagine a figure skater pulling their arms inward as they spin. Their diameter narrows and their rotation speed increases. The same physics—conservation of angular momentum—occurs as a tornado forms.



An EF 4 Tornado in Marquette, Kansas (Public Domain).

Tornados are difficult to forecast well in advance. Typically a Tornado Watch will be issued by a Storm Prediction Center when all of the ingredients for a tornado are present in the atmosphere. These ingredients are the same as for a severe thunderstorm: warm moist air; instability (quantified by large values of CAPE); potential for a trigger; and, of course, directional wind shear to provide rotation. A Tornado Warning is issued when a tornado is imminent or occurring. For example, one may be issued when a funnel cloud is spotted or when a "**hook echo**" is seen on radar. The issue of forecasting tornados is called "nowcasting". Although very little advance notice can be given to the public, forecasters do their best to provide as much notice as possible.



Hook echo in a radar image (Public Domain).

The above image shows a hook echo from radar. The key feature is the bottom left region of the storm where a region of rain can be seen rapping around a region of dry inflow air. Right at the cusp of the hook is where a tornado would form if it were occurring. The strong rotation in a tornado creates this radar signature.

Tornado strength is categorized after they occur. A scale called the **Enhanced-Fujita scale** is used to classify tornado strength based on the amount of damage that was caused. While the wind speed for each step in the EF scale is given, it is not measured. The wind speed is determined afterwards by the scale of damage and the winds required to cause the damage.

Enhanced Fujita Scale

Scale	Wind Speed (3 sec. gust, mph)	Damage
EF0	65-85	Minor
EF1	86-110	Moderate
EF2	111-135	Considerable
EF3	136-165	Severe
EF4	166-200	Extreme
EF5	>200	Massive/Incredible

The central US is the world's hot-spot for tornados, however, tornados and tornado-like storm features can occur elsewhere. A dust devil and a waterspout are tornado-like features that are associated with cumuliform and sometimes cumulonimbus clouds. However, dust devils and waterspouts typically are not associated with a supercell. Waterspouts look like tornados but form over water surfaces. Waterspouts are often narrow, relatively short lived, and formed by the same physics as a tornado—a rotating updraft with condensation in the funnel cloud. Below is an image of a waterspout observed in Hawaii off the coast of Oahu.



Waterspout spotted off O'ahu's southern shore (Taken on May 2, 2011, by Staff Sgt. Mike Meares).

This page titled 12.3.5: Tornadoes is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by LibreTexts.

12.3.6: Keeping Yourself Safe

From reading this storm hazards chapter, hopefully you've learned that severe storms and storm hazards should be taken seriously. Part of the issue with severe weather forecasting is that the impacts are often extremely localized. It is difficult to know if you will be in the exact location that hail will fall or a tornado will pass. Oftentimes, you'll simply be aware that the possibility is there for storm impacts. Therefore, it is important to take precautions, regardless of whether you are in the right place at the wrong time.

For example, here are a few simple steps you can follow:

- "Turn around, don't drown" is the saying from the National Weather Service. If you see flowing water over a roadway, don't drive over it!
- If there is a possibility for thunderstorms, don't go hiking or do outdoor activities near streams or rivers. Flash flooding is a possibility.
- Similarly, if there is a possibility for thunderstorms, don't put yourself in an exposed outdoor location. As soon as you hear thunder, a thunderstorm is close enough to be a hazard. Get yourself off of that open sports field, beach, or stadium, and into a shelter until it passes.
- If a tornado is approaching, the safest place is inside the interior room in your house (unless you have a tornado cellar). Get in your closet or your bathtub with a mattress or something to cover you.

Depending on where you live in the world, some hazards are more likely than others. Still, it is important to think in advance about what you would do if you were in a severe storm situation. Oftentimes there isn't a lot of time to react.

Chapter 15: Questions to Consider

1. Why does hail have concentric layers that alternate clear and opaque ice?

Query 12.3.6.1

2. *This interactive element is not available in this format. Visit the link or scan the QR code to access it online.*

   View this Interactive Content online

Query 12.3.6.2

3. *This interactive element is not available in this format. Visit the link or scan the QR code to access it online.*

   View this Interactive Content online

4. *FLASH* You saw a lightning stroke! 15 seconds go by... *BOOM* You heard thunder! How far away was the lightning stroke?
5. Where is the tornado?

Query 12.3.6.3

This interactive element is not available in this format. Visit the link or scan the QR code to access it online.

View this Interactive Content online



6. Bob just finished reading this chapter and wants to practice sheltering in the event of a tornado warning. Can you help him find the safest place to shelter in his house?

Query 12.3.6.4

This interactive element is not available in this format. Visit the link or scan the QR code to access it online.

View this Interactive Content online

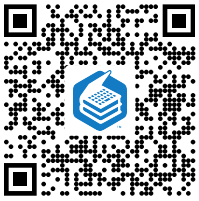


Selected Practice

Query 12.3.6.5

This interactive element is not available in this format. Visit the link or scan the QR code to access it online.

View this Interactive Content online



This page titled 12.3.6: Keeping Yourself Safe is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by LibreTexts.

13: EARTH MATERIALS

13.1: What is Geology?
13.2: Minerals and Rocks
13.3: Mineral Groups
13.4: Silicate Minerals
13.5: Mineral Properties

Chapter 13: Earth Materials — Introduction

With this chapter, we begin our study of geology and shift our attention from the atmosphere above us to the solid Earth beneath our feet. Geology examines the materials that make up Earth, the processes that shape it, and the evidence those materials preserve about our planet's history. Before we can understand processes such as volcanism, mountain building, or plate tectonics, we first need to become familiar with the basic materials from which the solid Earth is constructed.

We begin by distinguishing between **minerals and rocks**. Minerals are the fundamental building blocks of rocks, and their chemical composition and internal structure give them characteristic physical properties. Although thousands of minerals are known, they can be organized into groups based on their composition, and a relatively small number account for much of Earth's crust. In particular, **silicate minerals** are especially important because of the abundance of silicon and oxygen in the crust.

Finally, we will examine the physical properties geologists use to identify minerals. Properties such as hardness, luster, streak, cleavage, and fracture provide observable evidence that can be used to distinguish one mineral from another. These identification skills provide a foundation for the next chapter, where we will move from individual minerals to rocks and use those materials to begin reconstructing Earth's history.

This page titled 13: Earth Materials is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Robert Wagner.

13.1: What is Geology?

In its broadest sense, geology is the study of Earth—its interior and its exterior surface, the minerals, rocks and other materials that are around us, the processes that have resulted in the formation of those materials, the water that flows over the surface and through the ground, the changes that have taken place over the vastness of geological time, and the changes that we can anticipate will take place in the near future. Geology is a science, meaning that we use deductive reasoning and scientific methods to understand geological problems. It is, arguably, the most integrated of all of the sciences because it involves the understanding and application of all of the other sciences: physics, chemistry, biology, mathematics, astronomy, and others. But unlike most of the other sciences, geology has an extra dimension, that of time—deep time—billions of years of it. Geologists study the evidence that they see around them, but in most cases, they are observing the results of processes that happened thousands, millions, and even billions of years in the past. Those were processes that took place at incredibly slow rates—millimeters per year to centimeters per year—but because of the amount of time available, they produced massive results.

Geology is displayed on a grand scale in mountainous regions, perhaps nowhere better than the Rocky Mountains in Canada (Figure 13.1.1). The peak on the right is Rearguard Mountain, which is a few kilometers northeast of Mount Robson, the tallest peak in the Canadian Rockies (3,954 meters). The large glacier in the middle of the photo is the Robson Glacier. The river flowing from Robson Glacier drains into Berg Lake in the bottom right. There are many geological features portrayed here. The sedimentary rock that these mountains are made of formed in ocean water over 500 million years ago. A few hundred million years later, these beds were pushed east for tens to hundreds of kilometers by tectonic plate convergence and also pushed up to thousands of meters above sea level. Over the past two million years this area—like most of the rest of Canada—has been repeatedly glaciated, and the erosional effects of those glaciations are obvious.

The Robson Glacier is now only a small remnant of its size during the Little Ice Age of the 15th to 18th centuries, and even a lot smaller that it was just over a century ago in 1908. The distinctive line on the slope on the left side of both photos shows the elevation of the edge of the glacier a few hundred years ago. Like almost all other glaciers in the world, it receded after the 18th century because of natural climate change, is now receding even more rapidly because of human-caused climate change.



Figure 13.1.1 Rearguard Mountain and Robson Glacier in Mount Robson Provincial Park, BC. Left: Robson Glacier in 2012. Right: Robson Glacier circa 1908.

Geology is also about understanding the evolution of life on Earth; about discovering resources such as water, metals and energy; about recognizing and minimizing the environmental implications of our use of those resources; and about learning how to mitigate the hazards related to earthquakes, volcanic eruptions, and slope failures. All of these aspects of geology, and many more, are covered in this textbook.



Figure 13.1.2

There is no single method of inquiry that is specifically the “scientific method”; furthermore, scientific inquiry is not necessarily different from serious research in other disciplines. The most important thing that those involved in any type of inquiry must do is to be skeptical. As the physicist Richard Feynman once said: the first principle of science is that “you must not fool yourself—and you are the easiest person to fool.” A key feature of serious inquiry is the creation of a hypothesis (a tentative explanation) that could explain the observations that have been made, and then the formulation and testing (by experimentation) of one or more predictions that follow from that hypothesis.

For example, we might observe that most of the cobbles in a stream bed are well rounded (see photo above), and then derive the hypothesis that the rocks are rounded by transportation along the stream bed. A prediction that follows from this hypothesis is that cobbles present in a stream will become increasingly rounded as they are transported downstream. An experiment to test this prediction would be to place some angular cobbles in a stream, label them so that we can be sure to find them again later, and then return at various time intervals (over a period of years) to carefully measure their locations and roundness.

A critical feature of a good hypothesis and any resulting predictions is that they must be testable. For example, an alternative hypothesis to the one above is that an extraterrestrial organization creates rounded cobbles and places them in streams when nobody is looking. This may indeed be the case, but there is no practical way to test this hypothesis. Most importantly, there is no way to prove that it is false, because if we aren’t able to catch the aliens at work, we still won’t know if they did it!

Media Attributions

- Figure 13.1.1 (left): © Steven Earle. CC BY.
- Figure 13.1.1 (right): A.P. Coleman. Public domain. Source: Arthur P. Coleman Collection at Victoria University Library.
- Figure 13.1.2: © Steven Earle. CC BY.

This page titled 13.1: What is Geology? is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Steven Earle (BCCampus) via source content that was edited to the style and standards of the LibreTexts platform.

- 1.1: What is Geology? by Steven Earle is licensed CC BY 4.0. Original source: https://opentextbc.ca/physicalgeology2ed.

13.2: Minerals and Rocks

The rest of this chapter is devoted to a brief overview of a few of the important aspects of physical geology, starting with minerals and rocks. This is followed by a review of Earth's internal structure and the processes of plate tectonics, and an explanation of geological time.

The Earth is made up of varying proportions of the 90 naturally occurring elements—hydrogen, carbon, oxygen, magnesium, silicon, iron, and so on. In most geological materials, these combine in various ways to make minerals. Minerals will be covered in some detail in Chapter 2, but here we will briefly touch on what minerals are, and how they are related to rocks.

A mineral is a naturally occurring combination of specific elements that are arranged in a particular repeating three-dimensional structure or **lattice**.[1] The mineral **halite** is shown as an example in Figure 13.2.1.

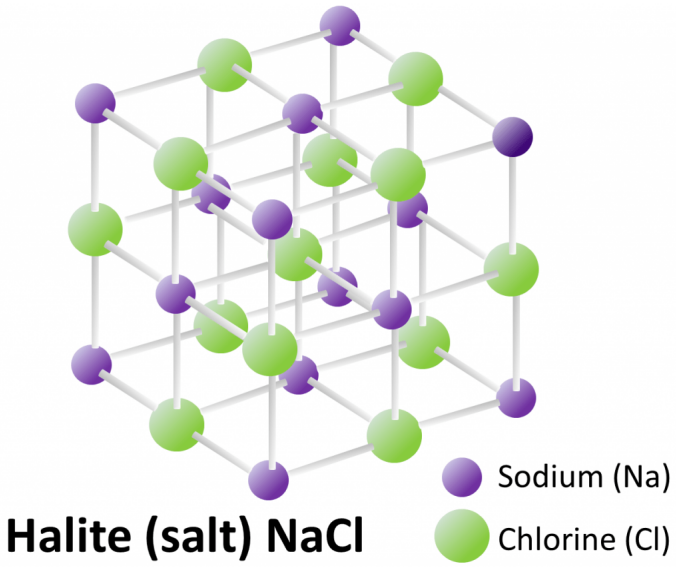


Figure 13.2.1 The lattice structure and composition of the mineral halite (common table salt).

In this case, atoms of sodium (Na: purple) alternate with atoms of chlorine (Cl: green) in all three dimensions, and the angles between the bonds are all 90°. Even in a tiny crystal, like the ones in your salt shaker, the lattices extend in all three directions for thousands of repetitions. Halite always has this composition and this structure.

Note: Element symbols (e.g., Na and Cl) are used extensively in this book. In Appendix 1, you will find a list of the symbols and names of the elements common in minerals and a copy of the periodic table. Please use those resources if you are not familiar with the element symbols.

There are thousands of minerals, although only a few dozen are mentioned in this book. In nature, minerals are found in rocks, and the vast majority of rocks are composed of at least a few different minerals. A close-up view of **granite**, a common rock, is shown in Figure 13.2.2. Although a hand-sized piece of granite may have thousands of individual mineral crystals in it, there are typically only a few different minerals, as shown here.

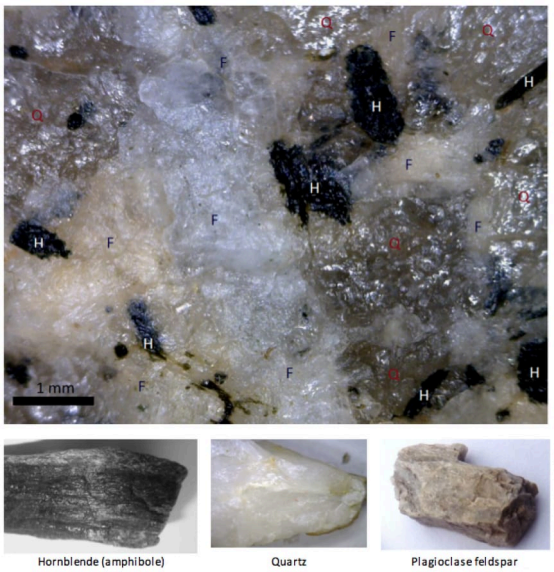


Figure 13.2.2 A close-up view of the rock granite and some of the minerals that it typically contains (H = hornblende (amphibole), Q = quartz and F = feldspar). The crystals range from about 0.1 to 3 millimeters (mm) in diameter. Most are irregular in outline, but some are rectangular.

Rocks can form in a variety of ways. Igneous rocks form from **magma** (molten rock) that has either cooled slowly underground (e.g., to produce granite) or cooled quickly at the surface after a volcanic eruption (e.g., **basalt**). Sedimentary rocks, such as **sandstone**, form when the weathered products of other rocks accumulate at the surface and are then buried by other sediments. Metamorphic rocks form when either igneous or sedimentary rocks are heated and squeezed to the point where some of their minerals are unstable and new minerals form to create a different type of rock. An example is **schist**.

A critical point to remember is the difference between a mineral and a rock. A mineral is a pure substance with a specific composition and structure, while a rock is typically a mixture of several different minerals (although a few types of rock may include only one type of mineral). Examples of minerals are feldspar, quartz, mica, halite, calcite, and amphibole. Examples of rocks are granite, basalt, sandstone, limestone, and schist.

> If you are currently taking a geology course, you'll likely be asked more than once to name a mineral or a rock that has specific characteristics or composition, or was formed in a specific environment. Please make sure that if you're asked for a **rock name** that you don't respond with a **mineral name**, and *vice versa*. Confusing minerals and rocks is one of the most common mistakes that geology students make.

Exercise 1.1 Find a piece of granite

The rock granite is very common in most parts of North America, and unless everything is currently covered with snow where you live, you should have no trouble finding a sample of it near you. The best places to look are pebbly ocean or lake beaches, a gravel bar of a creek or river, a gravel driveway, or somewhere where gravel has been used in landscaping. In Figure 13.2.3 , taken on a beach, the granitic pebbles are the ones that are predominantly light-colored with dark specks. The one in the very centre is a good example.



Figure 13.2.3

Select a sample of granite and, referring to Figure 13.2.2, see if you can identify some of the minerals in it. It may help to break it in half with a hammer to see a fresh surface, but be careful to protect your eyes if you do so. You should be able to see glassy-looking quartz, dull white plagioclase feldspar (and maybe pink potassium feldspar), and black hornblende or, in some cases, flaky black biotite mica (or both).

In addition to identifying the minerals in your granite, you might also try to describe the texture in terms of the range sizes of the mineral crystals (in millimeters) and the shapes of the crystals (some may be rectangular in outline, most will be irregular). Think about where your granite might have come from and how it got to where you found it.

See Appendix 3 for Exercise 1.1 answers.

Media Attributions

- Figure 13.2.1, 1.4.2, 1.4.3: © Steven Earle. CC BY.

---

1. Terms in bold are defined in the glossary at the end of the book. ↵

This page titled 13.2: Minerals and Rocks is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Steven Earle (BCCampus) via source content that was edited to the style and standards of the LibreTexts platform.

- 1.4: Minerals and Rocks by Steven Earle is licensed CC BY 4.0. Original source: https://opentextbc.ca/physicalgeology2ed.

13.3: Mineral Groups

Most minerals are made up of a cation (a positively charged ion) or several cations, plus an anion (a negatively charged ion (e.g., S^{2-})) or an anion complex (e.g., SO_4^{2-}). For example, in the mineral hematite (Fe_2O_3), the cation is Fe^{3+} (iron) and the anion is O^{2-} (oxygen). The two Fe^{3+} ions have an overall charge of +6 and that balances the overall charge of −6 from the three O^{2-} ions.

We group minerals into classes on the basis of their predominant anion or anion complex. These include oxides, sulfides, carbonates, silicates, and others. Silicates are by far the predominant group in terms of their abundance within the crust and mantle. (They will be discussed in Section 2.4). Some examples of minerals from the different mineral groups are given in Table 2.4.

Table 2.4 The main mineral groups and some examples of minerals in each group.

[Skip Table]

Group	Examples
Oxides	Hematite (iron oxide Fe_2O_3), corundum (aluminum oxide Al_2O_3), water ice (H_2O)
Sulphides	Galena (lead sulfide PbS), pyrite (iron sulfide FeS_2), chalcopyrite (copper-iron sulfide $CuFeS_2$)
Sulphates	Gypsum (calcium sulphate $CaSO_4{\cdot}H_2O$), barite (barium sulphate $BaSO_4$) *(Note that sulphates are different from sulfides. Sulphates have the SO_4^{-2} ion while sulfides have the S^{-2} ion)*
Halides	Fluorite (calcium flouride CaF_2), halite (sodium chloride NaCl) *(Halide minerals have halogen elements as their anion — the minerals in the second last column on the right side of the periodic table, including F, Cl, Br, etc. — see the periodic table in Appendix 1: List of Geologically Important Elements and the Periodic Table.)*
Carbonates	Calcite (calcium carbonate $CaCO_3$), dolomite (calcium-magnesium carbonate $(Ca,Mg)CO_3$)
Phosphates	Apatite ($Ca_5(PO_4)3(OH)$), Turquoise ($CuAl_6(PO_4)4(OH)_8{\cdot}5H_2O$)
Silicates	Quartz (SiO_2), feldspar (sodium-aluminum silicate $NaAlSi_3O_8$), olivine (iron or magnesium silicate $(Mg,Fe)_2SiO_4$) *(Note that in quartz the anion is oxygen, and while it could be argued, therefore, that quartz is an oxide, it is always classed with the silicates.)*
Native minerals	Gold (Au), diamond (C), graphite (C), sulfur (S), copper (Cu)

Oxide minerals have oxygen (O^{2-}) as their anion, but they exclude those with oxygen complexes such as carbonate (CO_3^{2-}), sulphate (SO_4^{2-}), and silicate (SiO_4^{4-}). The most important oxides are the iron oxides hematite and magnetite (Fe_2O_3 and Fe_3O_4, respectively). Both of these are common ores of iron. Corundum (Al_2O_3) is used as an abrasive, but can also be a gemstone in its ruby and sapphire varieties. If the oxygen is also combined with hydrogen to form the hydroxyl anion (OH^-) the mineral is known as a **hydroxide**. Some important hydroxides are limonite and bauxite, which are ores of iron and aluminium respectively. Frozen water (H_2O) is a mineral (an oxide), but liquid water is not because it doesn't have a regular lattice.

Sulphides are minerals with the S^{-2} anion, and they include galena (PbS), sphalerite (ZnS), chalcopyrite ($CuFeS_2$), and molybdenite (MoS_2), which are the most important ores of lead, zinc, copper, and molybdenum respectively. Some other sulfide minerals are pyrite (FeS_2), bornite (Cu_5FeS_4), stibnite (Sb_2S_3), and arsenopyrite (FeAsS).

Sulphates are minerals with the SO_4^{-2} anion, and these include anhydrite ($CaSO_4$) and its cousin gypsum ($CaSO_4.2H_2O$) and the sulphates of barium and strontium: barite ($BaSO_4$) and celestite ($SrSO_4$). In all of these minerals, the cation has a +2 charge, which balances the −2 charge on the sulphate ion.

The **halides** are so named because the anions include the **halogen** elements chlorine, fluorine, bromine, etc. Examples are halite (NaCl), cryolite (Na_3AlF_6), and fluorite (CaF_2).

The **carbonates** include minerals in which the anion is the CO_3^{-2} complex. The carbonate combines with +2 cations to form minerals such as calcite ($CaCO_3$), magnesite ($MgCO_3$), dolomite ($(Ca,Mg)CO_3$)[1], and siderite ($FeCO_3$). The copper minerals malachite and azurite are also carbonates.

In **phosphate** minerals, the anion is the PO_4^{-3} complex. An important phosphate mineral is apatite ($Ca_5(PO_4)_3(OH)$), which is what your teeth are made of. Note that it is called a phosphate, not a hydroxide, even though it has a hydroxyl ion.

The **silicate** minerals include the elements silicon and oxygen in varying proportions ranging from Si : O_2 to Si : O_4. These are discussed at length in Section 2.4.

Native minerals are single-element minerals, such as gold, copper, sulfur, and graphite.

Exercise 2.2 Mineral groups

We classify minerals according to the anion part of the mineral formula, and mineral formulas are always written with the anion part on the right. For example, for pyrite (FeS_2), Fe^{2+} is the cation and S^- is the anion. This helps us to know that it's a sulfide, but it is not always that obvious. Hematite (Fe_2O_3) is an oxide; that's easy, but anhydrite ($CaSO_4$) is a sulphate because SO_4^{2-} is the anion, not O. Along the same lines, calcite ($CaCO_3$) is a carbonate, and olivine (Mg_2SiO_4) is a silicate. Minerals with only one element (such as S) are native minerals, while those with an anion from the halogen column of the periodic table (Cl, F, Br, etc.) are halides. Provide group names for the following minerals:

Table 2.5 Provide group names for the following minerals

[Skip Table]

Name	Formula	Group
sphalerite	ZnS	
magnetite	Fe_3O_4	
pyroxene	$MgSiO_3$	
anglesite	$PbSO_4$	
sylvite	KCl	
silver	Ag	
fluorite	CaF_2	
ilmenite	$FeTiO_3$	
siderite	$FeCO_3$	
feldspar	$KAlSi_3O_8$	
sulfur	S	
xenotime	YPO_4	

See Appendix 3 for Exercise 2.2 answers.

1. The notations of two (or more) elements enclosed in parentheses with a comma between them: (Ca,Mg), indicates that both can be present, in varying proportions, but that there is still only one of them for each anion present. ↵

This page titled 13.3: Mineral Groups is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Steven Earle (BCCampus) via source content that was edited to the style and standards of the LibreTexts platform.

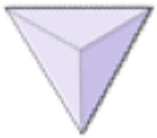
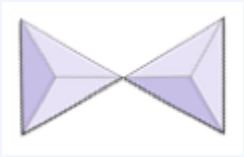
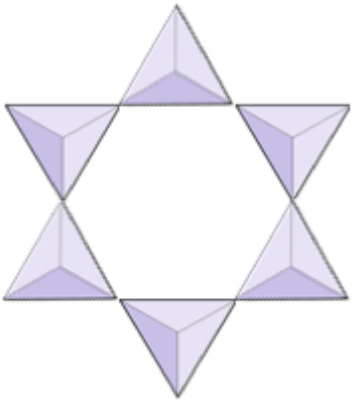
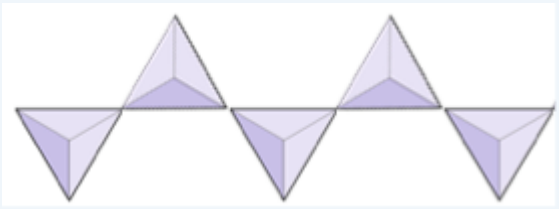
- 2.3: Mineral Groups by Steven Earle is licensed CC BY 4.0. Original source: https://opentextbc.ca/physicalgeology2ed.

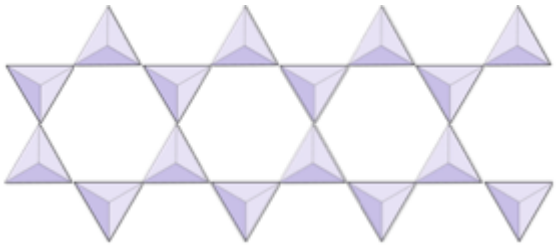
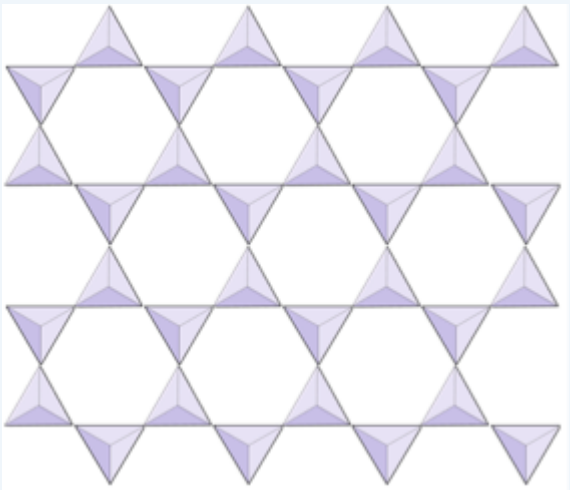
13.4: Silicate Minerals

The vast majority of the minerals that make up the rocks of Earth's crust are silicate minerals. These include minerals such as quartz, feldspar, mica, amphibole, pyroxene, olivine, and a variety of clay minerals. The building block of all of these minerals is the **silica tetrahedron**, a combination of four oxygen atoms and one silicon atom. As we've seen, it's called a tetrahedron because planes drawn through the oxygen atoms form a shape with 4 surfaces (Figure 13.4.4). Since the silicon ion has a charge of 4 and each of the four oxygen ions has a charge of –2, the silica tetrahedron has a net charge of –4.

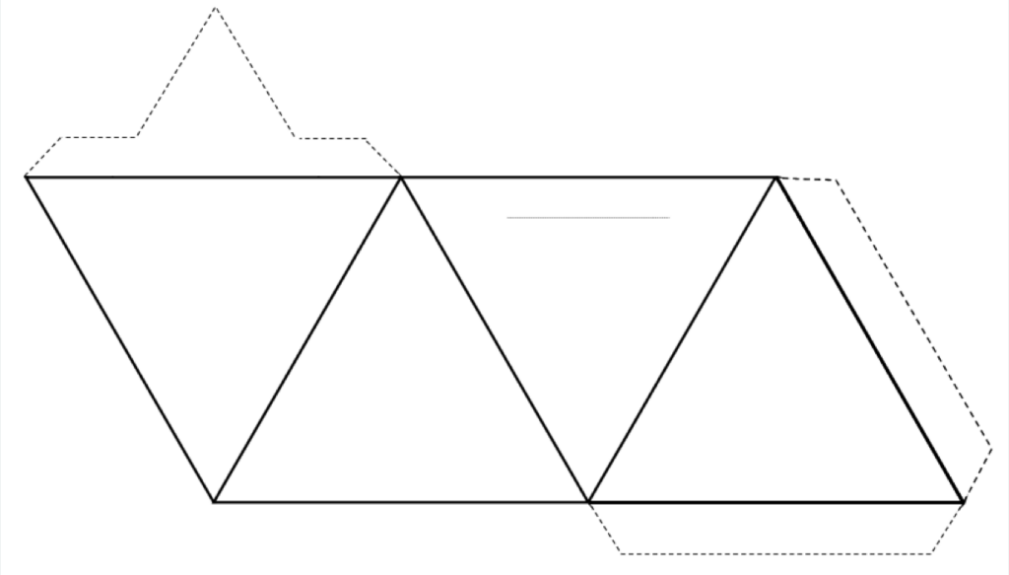
In silicate minerals, these tetrahedra are arranged and linked together in a variety of ways, from single units to complex frameworks (Table 2.6). The simplest silicate structure, that of the mineral **olivine**, is composed of isolated tetrahedra bonded to iron and/or magnesium ions. In olivine, the –4 charge of each silica tetrahedron is balanced by two **divalent** (i.e., +2) iron or magnesium cations. Olivine can be either Mg_2SiO_4 or Fe_2SiO_4, or some combination of the two $(Mg,Fe)_2SiO_4$. The divalent cations of magnesium and iron are quite close in radius (0.73 versus 0.62 angstroms[1]). Because of this size similarity, and because they are both divalent cations (both can have a charge of +2), iron and magnesium can readily substitute for each other in olivine and in many other minerals.

Table 2.6 Silicate mineral configurations. The triangles represent silica tetrahedra.

Tetrahedron Configuration Picture	Tetrahedron Configuration Name	Example Minerals
	Isolated (nesosilicates)	Olivine, garnet, zircon, kyanite
	Pairs (sorosilicates)	Epidote, zoisite
	Rings (cyclosilicates)	Tourmaline
	Single chains (inosilicates)	Pyroxenes, wollastonite

Tetrahedron Configuration Picture	Tetrahedron Configuration Name	Example Minerals
	Double chains (inosilicates)	Amphiboles
	Sheets (phyllosilicates)	Micas, clay minerals, serpentine, chlorite
3-dimensional structure	Framework (tectosilicates)	Feldspars, quartz, zeolite

Exercise 2.3 Make a Tetrahedron



Cut around the outside of the shape (solid lines and dotted lines), and then fold along the solid lines to form a tetrahedron. If you have glue or tape, secure the tabs to the tetrahedron to hold it together. If you don't have glue or tape, make a slice along the thin grey line and insert the pointed tab into the slit.

If you are doing this in a classroom, try joining your tetrahedron with others into pairs, rings, single and double chains, sheets, and even three-dimensional frameworks.

See Appendix 3 for Exercise 2.3 answers.

In olivine, unlike most other silicate minerals, the silica tetrahedra are not bonded to each other. Instead they are bonded to the iron and/or magnesium ions, in the configuration shown on Figure 13.4.1.

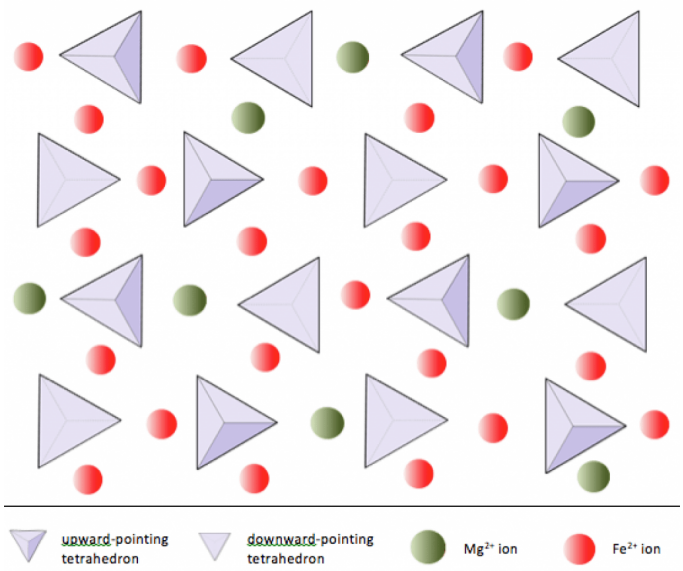


Figure 13.4.1 A depiction of the structure of olivine as seen from above. The formula for this particular olivine, which has three Fe ions for each Mg ion, could be written: $Mg_{0.5}Fe_{1.5}SiO_4$.

As already noted, the 2 ions of iron and magnesium are similar in size (although not quite the same). This allows them to substitute for each other in some silicate minerals. In fact, the ions that are common in silicate minerals have a wide range of sizes, as depicted in Figure 13.4.2. All of the ions shown are cations, except for oxygen. Note that iron can exist as both a +2 ion (if it loses two electrons during ionization) or a +3 ion (if it loses three). Fe^{2+} is known as **ferrous** iron. Fe^{3+} is known as **ferric** iron. Ionic radii are critical to the composition of silicate minerals, so we'll be referring to this diagram again.

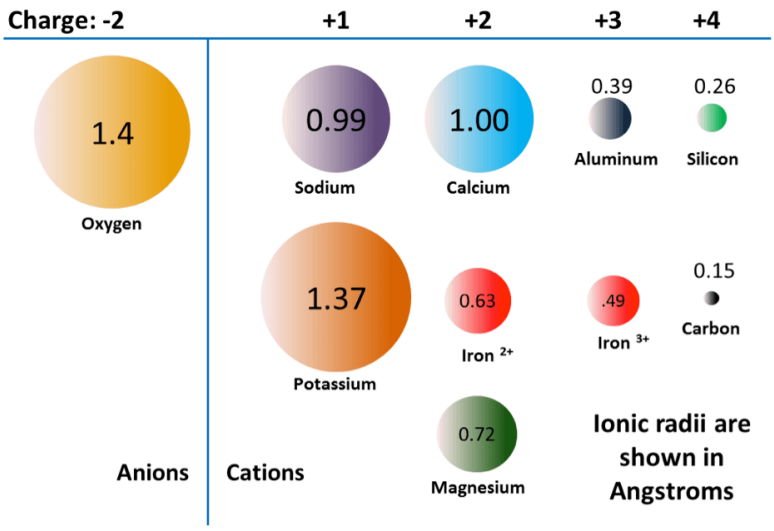


Figure 13.4.2 The ionic radii (effective sizes) in angstroms, of some of the common ions in silicate minerals.

The structure of the single-chain silicate pyroxene is shown on Figures 2.4.3 and 2.4.4. In **pyroxene**, silica tetrahedra are linked

together in a single chain, where one oxygen ion from each tetrahedron is shared with the adjacent tetrahedron, hence there are fewer oxygens in the structure. The result is that the oxygen-to-silicon ratio is lower than in olivine (3:1 instead of 4:1), and the net charge per silicon atom is less (–2 instead of –4). Therefore, fewer cations are necessary to balance that charge. Pyroxene compositions are of the type $MgSiO_3$, $FeSiO_3$, and $CaSiO_3$, or some combination of these. Pyroxene can also be written as $(Mg,Fe,Ca)SiO_3$, where the elements in the brackets can be present in any proportion. In other words, pyroxene has one cation for each silica tetrahedron (e.g., $MgSiO_3$) while olivine has two (e.g., Mg_2SiO_4). Because each silicon ion is +4 and each oxygen ion is –2, the three oxygens (–6) and the one silicon (+4) give a net charge of –2 for the single chain of silica tetrahedra. In pyroxene, the one divalent cation (2) per tetrahedron balances that –2 charge. In olivine, it takes two divalent cations to balance the –4 charge of an isolated tetrahedron.The structure of pyroxene is more “permissive” than that of olivine—meaning that cations with a wider range of ionic radii can fit into it. That’s why pyroxenes can have iron (radius 0.63 Å) or magnesium (radius 0.72 Å) or calcium (radius 1.00 Å) cations (see Figure 13.4.2 above).

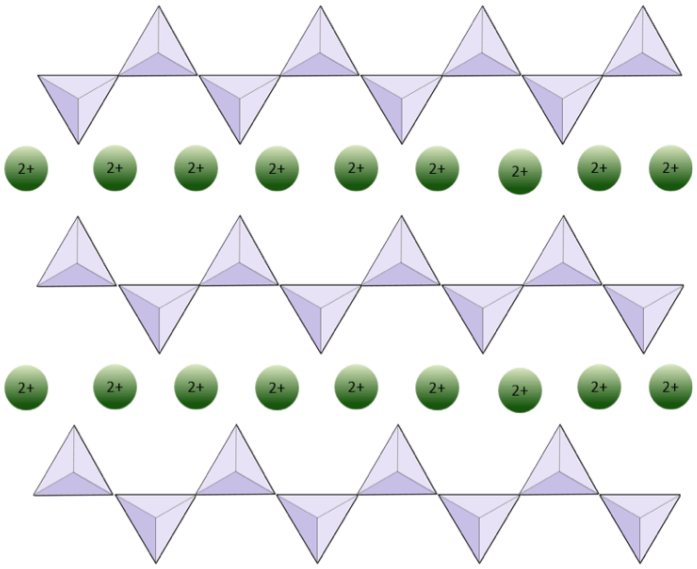


Figure 13.4.3 A depiction of the structure of pyroxene. The tetrahedral chains continue to left and right and each is interspersed with a series of divalent cations. If these are Mg ions, then the formula is $MgSiO_3$.

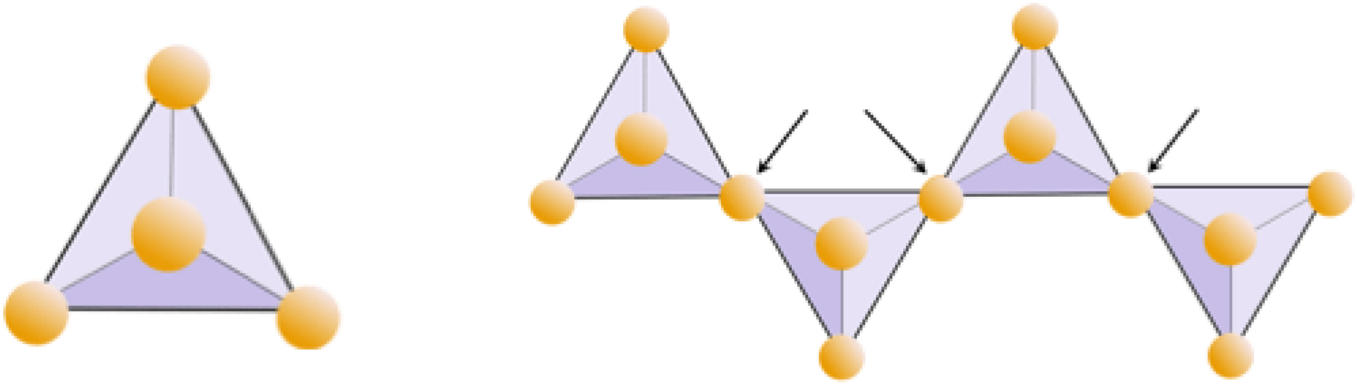
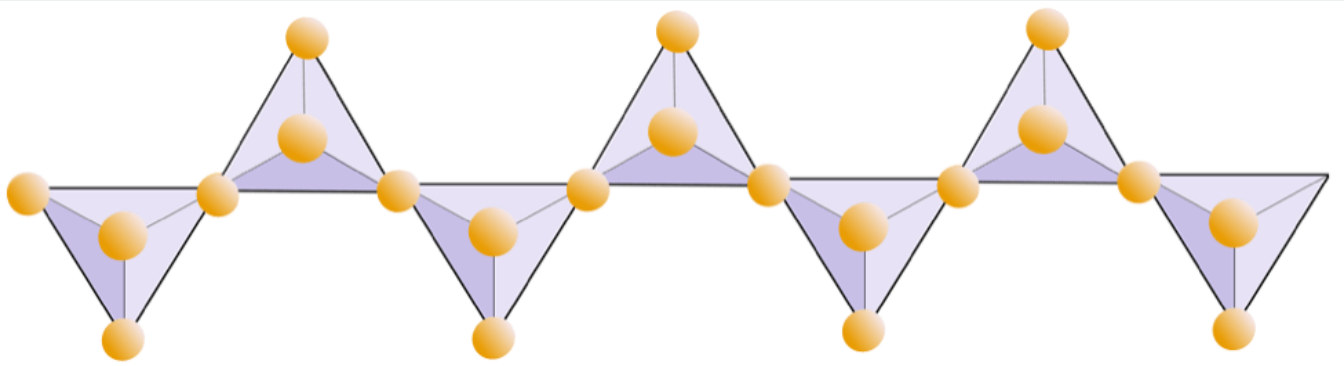


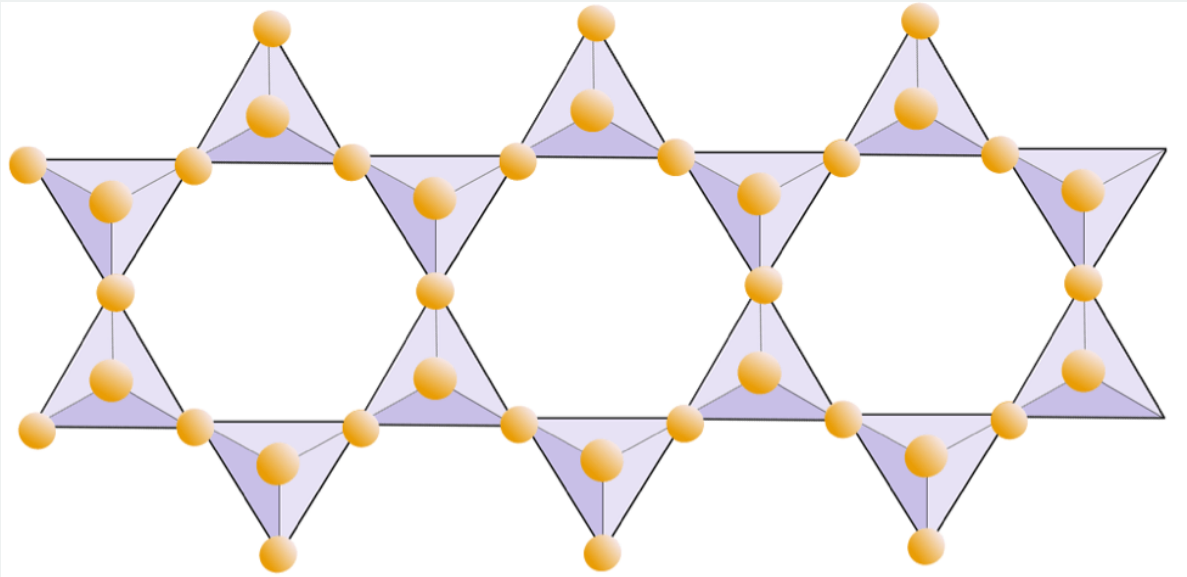
Figure 13.4.4 A single silica tetrahedron (left) with four oxygen ions per silicon ion (SiO_4). Part of a single chain of tetrahedra (right), where the oxygen atoms at the adjoining corners are shared between two tetrahedra (arrows). For a very long chain the resulting ratio of silicon to oxygen is 1 to 3 (SiO_3).

Exercise 2.4 Oxygen deprivation

The diagram below represents a single chain in a silicate mineral. Count the number of tetrahedra versus the number of oxygen ions (yellow spheres). Each tetrahedron has one silicon ion so this should give you the ratio of Si to O in single-chain silicates (e.g., pyroxene).



The diagram below represents a double chain in a silicate mineral. Again, count the number of tetrahedra versus the number of oxygen ions. This should give you the ratio of Si to O in double-chain silicates (e.g., amphibole).



See Appendix 3 for Exercise 2.4 answers.

In **amphibole** structures, the silica tetrahedra are linked in a double chain that has an oxygen-to-silicon ratio lower than that of pyroxene, and hence still fewer cations are necessary to balance the charge. Amphibole is even more permissive than pyroxene and its compositions can be very complex. Hornblende, for example, can include sodium, potassium, calcium, magnesium, iron, aluminum, silicon, oxygen, fluorine, and the hydroxyl ion (OH^-).

In **mica** structures, the silica tetrahedra are arranged in continuous sheets, where each tetrahedron shares three oxygen anions with adjacent tetrahedra. There is even more sharing of oxygens between adjacent tetrahedra and hence fewer cations are needed to balance the charge of the silica-tetrahedra structure in sheet silicate minerals. Bonding between sheets is relatively weak, and this accounts for the well-developed one-directional cleavage in micas (Figure 13.4.5). **Biotite** mica can have iron and/or magnesium in it and that makes it a **ferromagnesian** silicate mineral (like olivine, pyroxene, and amphibole). **Chlorite** is another similar mineral that commonly includes magnesium. In **muscovite** mica, the only cations present are aluminum and potassium; hence it is a non-ferromagnesian silicate mineral.

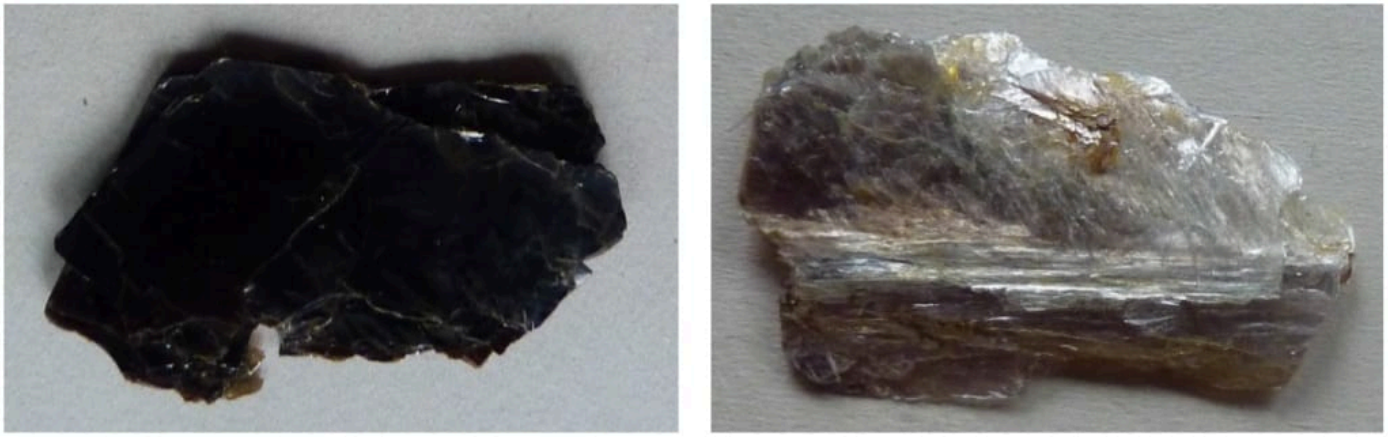


Figure 13.4.5 Biotite mica (left) and muscovite mica (right). Both are sheet silicates and split easily into thin layers along planes parallel to the sheets. Biotite is dark like the other iron- and/or magnesium-bearing silicates (e.g., olivine, pyroxene, and amphibole), while muscovite is light colored. (Each sample is about 3 cm across.)

Apart from muscovite, biotite, and chlorite, there are many other **sheet silicates** (a.k.a. **phyllosilicates**), many of which exist as clay-sized fragments (i.e., less than 0.004 millimeters). These include the clay minerals **kaolinite**, **illite**, and **smectite**, and although they are difficult to study because of their very small size, they are extremely important components of rocks and especially of soils.

All of the sheet silicate minerals also have water molecules within their structure.

Silica tetrahedra are bonded in three-dimensional frameworks in both the **feldspars** and **quartz**. These are **non-ferromagnesian minerals**—they don't contain any iron or magnesium. In addition to silica tetrahedra, feldspars include the cations aluminum, potassium, sodium, and calcium in various combinations. Quartz contains only silica tetrahedra.

The three main **feldspar** minerals are **potassium feldspar**, (a.k.a. K-feldspar or K-spar) and two types of plagioclase feldspar: **albite** (sodium only) and **anorthite** (calcium only). As is the case for iron and magnesium in olivine, there is a continuous range of compositions (solid solution series) between albite and anorthite in plagioclase. Because the calcium and sodium ions are almost identical in size (1.00 Å versus 0.99 Å) any intermediate compositions between $CaAl_2Si_3O_8$ and $NaAlSi_3O_8$ can exist (Figure 13.4.6). This is a little bit surprising because, although they are very similar in size, calcium and sodium ions don't have the same charge (Ca^{2+} versus Na^{+}). This problem is accounted for by the corresponding substitution of Al^{+3} for Si^{+4} . Therefore, albite is $NaAlSi_3O_8$ (1 Al and 3 Si) while anorthite is $CaAl_2Si_2O_8$ (2 Al and 2 Si), and plagioclase feldspars of intermediate composition have intermediate proportions of Al and Si. This is called a "coupled-substitution."

The intermediate-composition plagioclase feldspars are oligoclase (10% to 30% Ca), andesine (30% to 50% Ca), labradorite (50% to 70% Ca), and bytownite (70% to 90% Ca). **K-feldspar** ($KAlSi_3O_8$) has a slightly different structure than that of plagioclase, owing to the larger size of the potassium ion (1.37 Å) and because of this large size, potassium and sodium do not readily substitute for each other, except at high temperatures. These high-temperature feldspars are likely to be found only in volcanic rocks because intrusive igneous rocks cool slowly enough to low temperatures for the feldspars to change into one of the lower-temperature forms.

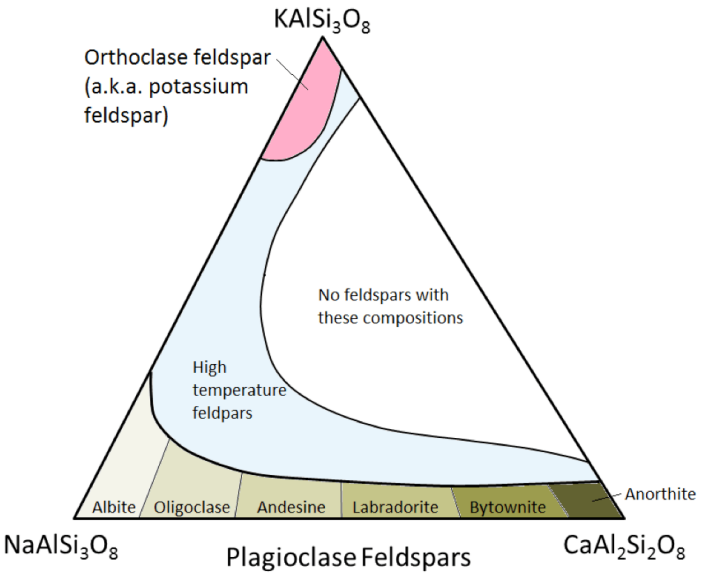


Figure 13.4.6 Compositions of the feldspar minerals.

In **quartz** (SiO_2)**,** the silica tetrahedra are bonded in a "perfect" three-dimensional framework. Each tetrahedron is bonded to four other tetrahedra (with an oxygen shared at every corner of each tetrahedron), and as a result, the ratio of silicon to oxygen is 1:2. Since the one silicon cation has a +4 charge and the two oxygen anions each have a –2 charge, the charge is balanced. There is no need for aluminum or any of the other cations such as sodium or potassium. The hardness and lack of cleavage in quartz result from the strong covalent/ionic bonds characteristic of the silica tetrahedron.

Exercise 2.5 Ferromagnesian silicates?

Silicate minerals are classified as being either ferromagnesian or non-ferromagnesian depending on whether or not they have iron (Fe) and/or magnesium (Mg) in their formula. A number of minerals and their formulas are listed below. For each one, indicate whether or not it is a *ferromagnesian silicate*.

Mineral	**Formula**	**Ferromagnesian silicate?**
olivine	$(Mg,Fe)_2SiO_4$	.
pyrite	FeS_2	.
plagioclase feldspar	$CaAl_2Si_2O_8$	.
pyroxene	$MgSiO_3$	.
hematite	Fe_2O_3	.
orthoclase feldspar	$KAlSi_3O_8$	.
quartz	SiO_2	.
amphibole	$Fe_7Si_8O_{22}(OH)_2$	.
muscovite	$K_2Al_4Si_6Al_2O_{20}(OH)_4$	.
magnetite	Fe_3O_4	.
biotite	$K_2Fe_4Al_2Si_6Al_4O_{20}(OH)_4$	.
dolomite	$(Ca,Mg)CO_3$	.
garnet	$Fe_2Al_2Si_3O_{12}$	.
serpentine	$Mg_3Si_2O_5(OH)_4$	.

See Appendix 3 for Exercise 2.5 answers.*Some of the formulas, especially the more complicated ones, have been simplified.

Image Descriptions

Figure 13.4.2 image description: The ionic radii of elements in angstroms and their charges.

Element	Ionic Radii (in angstroms)	Charge
Oxygen	1.4	–2 (Anion)
Potassium	1.37	1 (Cation)
Calcium	1.00	2 (Cation)
Sodium	0.99	1 (Cation)
Magnesium	0.72	2 (Cation)
Iron	0.63	2 (Cation)
	0.49	3 (Cation)
Aluminum	0.39	3 (Cation)
Silicon	0.26	4 (Cation)
Carbon	0.15	4 (Cation)

[Return to Figure 13.4.2]

---

1. An angstrom is the unit commonly used for the expression of atomic-scale dimensions. One angstrom is 10^{-10} meters or 0.0000000001 meters. The symbol for an angstrom is Å. ↵

This page titled 13.4: Silicate Minerals is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Steven Earle (BCCampus) via source content that was edited to the style and standards of the LibreTexts platform.

- 2.4: Silicate Minerals by Steven Earle is licensed CC BY 4.0. Original source: https://opentextbc.ca/physicalgeology2ed.

13.5: Mineral Properties

Minerals are universal. A crystal of hematite on Mars will have the same properties as one on Earth, and the same as one on a planet orbiting another star. That's good news for geology students who are planning interplanetary travel since we can use those properties to help us identify minerals anywhere. That doesn't mean that it's easy, however; identification of minerals takes a lot of practice. Some of the mineral properties that are useful for identification are as follows: color, streak, lustre, hardness, crystal habit, cleavage/fracture, density and a few others.

Colour

For most of us, color is one of our key ways of identifying objects. While some minerals have particularly distinctive colors that make good diagnostic properties, many do not, and for many, color is simply unreliable. The mineral sulfur (2.6.1 left) is always a distinctive and unique yellow. Hematite, on the other hand, is an example of a mineral for which color is not diagnostic. In some forms hematite is deep dull red, but in others it is black and shiny metallic (Figure 13.5.2). Many other minerals can have a wide range of colors (e.g., quartz, feldspar, amphibole, fluorite, and calcite). In most cases, the variations in colors are a result of varying proportions of trace elements within the mineral. In the case of quartz, for example, yellow quartz (citrine) has trace amounts of ferric iron (Fe^{3+}), rose quartz has trace amounts of manganese, purple quartz (amethyst) has trace amounts of iron, and milky quartz, which is very common, has millions of fluid inclusions (tiny cavities, each filled with water).

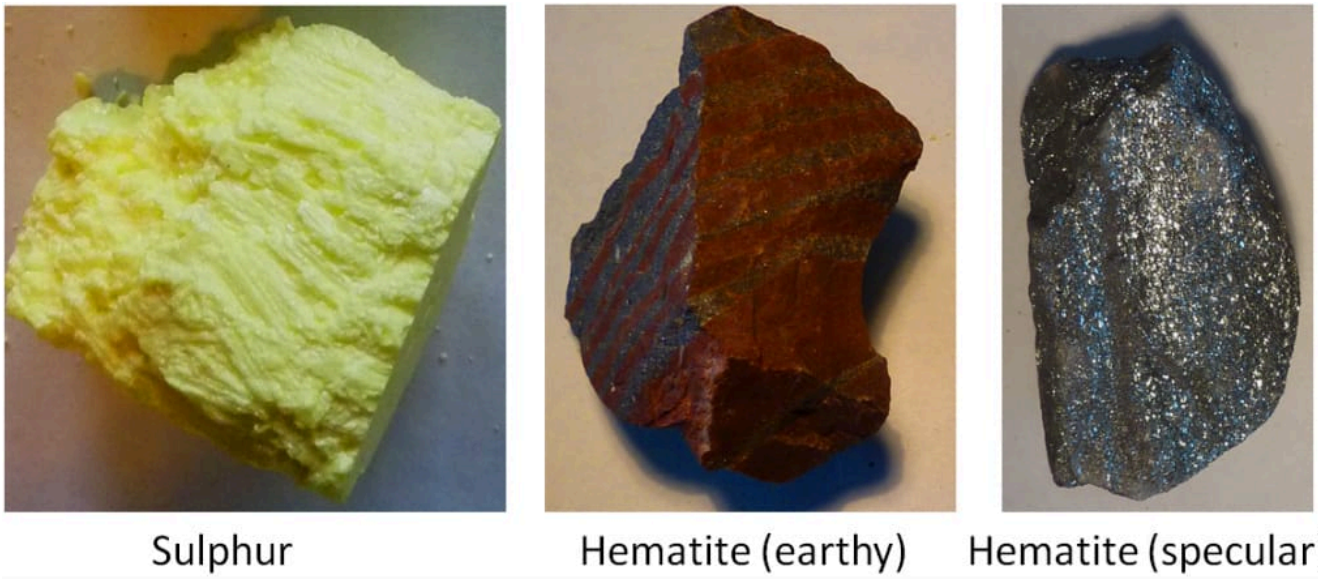


Figure 13.5.1 Examples of the colors of the minerals sulfur and hematite.

Streak

In the context of minerals, "color" is what you see when light reflects off the surface of the sample. One reason that color can be so variable is that the type of surface is variable. It may be a crystal face or a fracture surface or a cleavage plane, and the crystals may be large or small depending on the nature of the rock. If we grind a small amount of the sample to a powder we get a much better indication of its actual color. This can easily be done by scraping a corner of the sample across a streak plate (a piece of unglazed porcelain) to make a **streak**. The result is that some of the mineral gets ground to a powder and we can get a better impression of its "true" color (Figure 13.5.2).

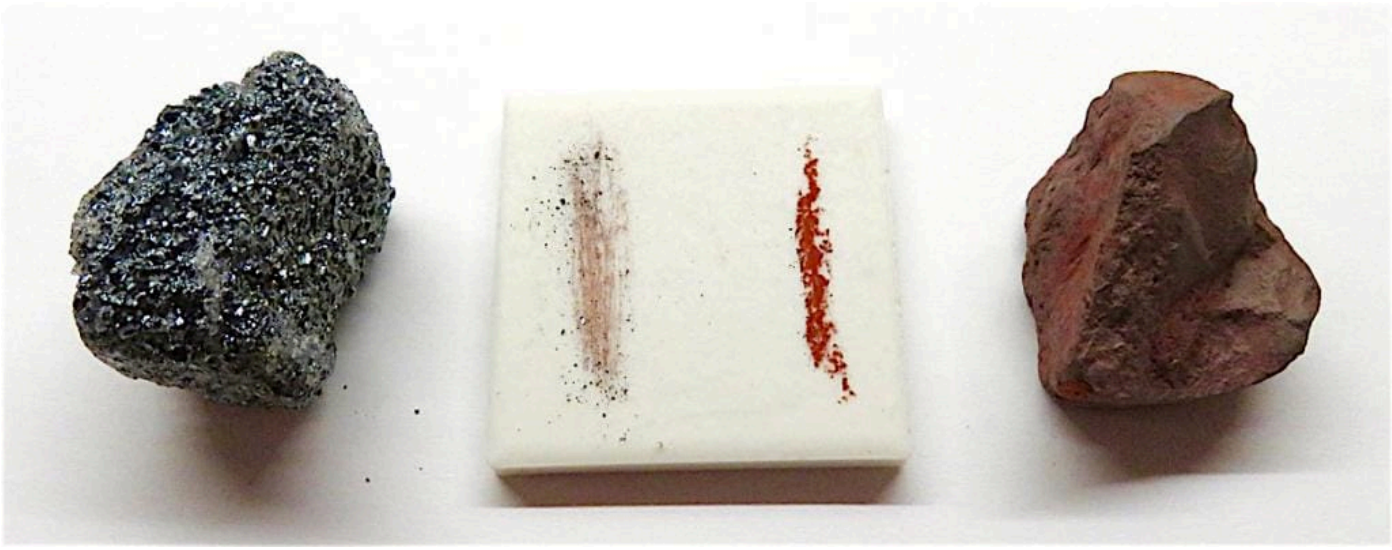


Figure 13.5.2 The streak colors of specular (metallic) hematite (left) and earthy hematite (right). Hematite leaves a distinctive reddish-brown streak whether the sample is metallic or earthy.

Lustre

Lustre is the way light reflects off the surface of a mineral, and the degree to which it penetrates into the interior. The key distinction is between **metallic** and **non-metallic lustres**. Light does not pass through metals, and that is the main reason they look "metallic." Even a thin sheet of metal—such as aluminum foil—will not allow light to pass through it. Many non-metallic minerals may look as if light will not pass through them, but if you take a closer look at a thin edge of the mineral you can see that it does. If a non-metallic mineral has a shiny, reflective surface, then it is called "glassy." If it is dull and non-reflective, it is "earthy." Other types of non-metallic lustres are "silky," "pearly," and "resinous." Lustre is a good diagnostic property since most minerals will always appear either metallic or non-metallic. There are a few exceptions to this (e.g., hematite in Figure 13.5.1).

Hardness

One of the most important diagnostic properties of a mineral is its hardness. In 1812 German mineralogist Friedrich Mohs came up with a list of 10 reasonably common minerals that had a wide range of hardnesses. These minerals are shown in Figure 13.5.3 , with the Mohs scale of hardness along the bottom axis. In fact, while each mineral on the list is harder than the one before it, the relative measured hardnesses (vertical axis) are not linear. For example apatite is about three times harder than fluorite and diamond is three times harder than corundum. Some commonly available reference materials are also shown on this diagram, including a typical fingernail (2.5), a piece of copper wire (3.5), a knife blade or a piece of window glass (5.5), a hardened steel file (6.5), and a porcelain streak plate (7). These are tools that a geologist can use to measure the hardness of unknown minerals. For example, if you have a mineral that you can't scratch with your fingernail, but you can scratch with a copper wire, then its hardness is between 2.5 and 3.5. And of course the minerals themselves can be used to test other minerals.

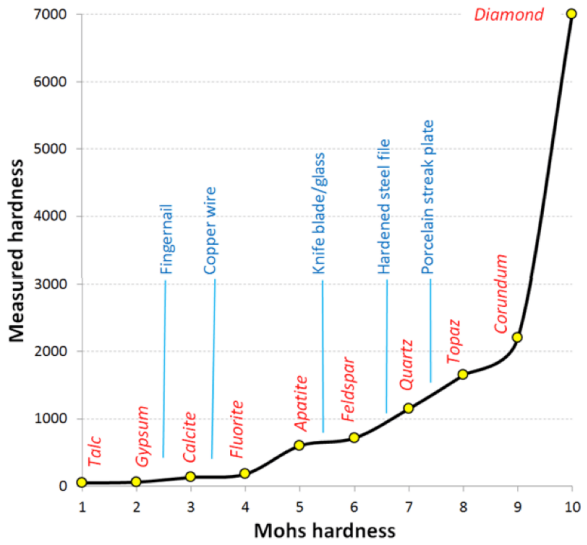


Figure 13.5.3 Minerals and reference materials in the Mohs scale of hardness. The “measured hardness” values are Vickers Hardness numbers.

Crystal Habit

When minerals form within rocks, there is a possibility that they will form in distinctive crystal shapes if they formed slowly and if they are not crowded out by other pre-existing minerals. Every mineral has one or more distinctive crystal **habits**, but it is not that common, in ordinary rocks, for the shapes to be obvious. Quartz, for example, will form six-sided prisms with pointed ends (Figure 13.5.4a), but this typically happens only when it crystallizes from a hot water solution within a cavity in an existing rock. Pyrite can form cubic crystals (Figure 13.5.4b), but can also form crystals with 12 faces, known as **dodecahedra** (“dodeca” means 12). The mineral garnet also forms dodecahedral crystals (Figure 13.5.4c).



Figure 13.5.4a Hexagonal prisms of quartz.



Figure 13.5.4b Cubic crystals of pyrite.



Figure 13.5.4c A dodecahedral crystal of garnet.

Because well-formed crystals are rare in ordinary rocks, habit isn't as useful a diagnostic feature as one might think. However, there are several minerals for which it is important. One is garnet, which is common in some metamorphic rocks and typically displays the dodecahedral shape. Another is amphibole, which forms long thin crystals, and is common in igneous rocks like granite (Figure 13.5.2).

Mineral habit is often related to the regular arrangement of the molecules that make up the mineral. Some of the terms that are used to describe habit include bladed, botryoidal (grape-like), dendritic (branched), drusy (an encrustation of minerals), equant (similar in all dimensions), fibrous, platy, prismatic (long and thin), and stubby.

Cleavage and Fracture

Crystal habit is a reflection of how a mineral grows, while cleavage and fracture describe how it breaks. Cleavage and fracture are the most important diagnostic features of many minerals, and often the most difficult to understand and identify. **Cleavage** is what we see when a mineral breaks along a specific plane or planes, while **fracture** is an irregular break. Some minerals tend to cleave along planes at various fixed orientations, some do not cleave at all (they only fracture). Minerals that have cleavage can also fracture along surfaces that are not parallel to their cleavage planes.

As we've already discussed, the way that minerals break is determined by their atomic arrangement and specifically by the orientation of weaknesses within the lattice. Graphite and the micas, for example, have cleavage planes parallel to their sheets (Figures 2.2.5 and 2.4.5), and halite has three cleavage planes parallel to the lattice directions (Figure 13.5.6).

Quartz has no cleavage because it has equally strong Si–O bonds in all directions, and feldspar has two cleavages at 90° to each other (Figure 13.5.5).

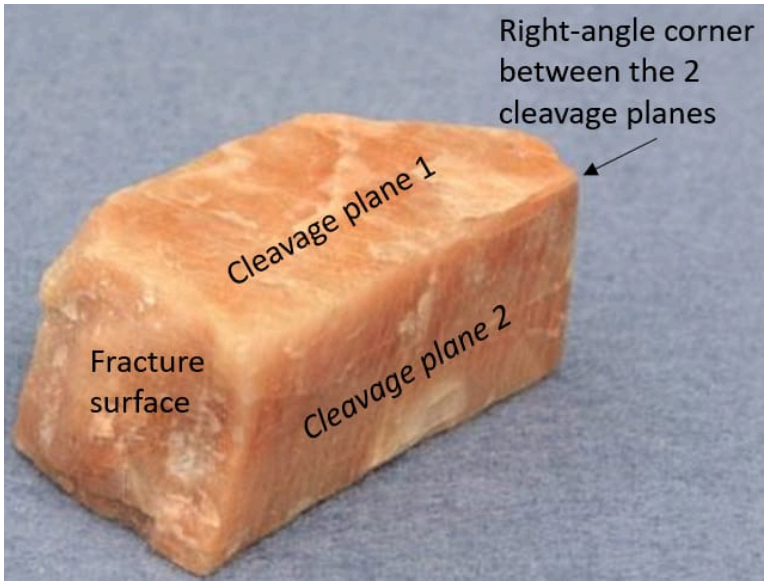


Figure 13.5.5 Cleavage and fracture in potassium feldspar

One of the main difficulties with recognizing and describing cleavage is that it is visible only in individual crystals. Most rocks have small crystals and it's very difficult to see the cleavage within those crystals. Geology students have to work hard to understand and recognize cleavage, but it's worth the effort since it is a reliable diagnostic property for most minerals.

One last thing: it is important to recognize the difference between cleavage planes and crystal surfaces. As already noted, crystal surfaces are related to how a mineral grows while cleavage planes are related to how it breaks. In most minerals cleavage planes and crystal surfaces do not align with one-another. An exception is halite, which grows in cubic crystals and has cleavage along those same planes (Figure 13.5.1 and 2.2.6). But this doesn't hold for most minerals. Quartz has crystal surfaces but no cleavage at all. Fluorite forms cubic crystals like those of halite, but it cleaves along planes that differ in orientation from the crystal surfaces. This is illustrated in Figure 13.5.6.

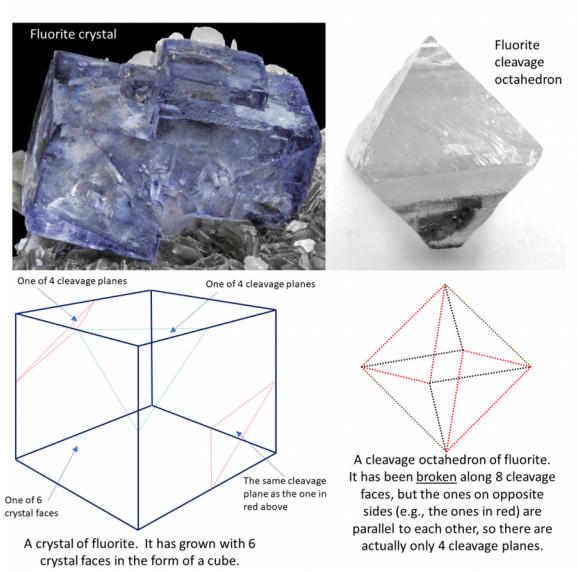


Figure 13.5.6 Crystal faces and cleavage planes in the mineral fluorite. The top-left photo shows a natural crystal of fluorite. It has crystal surfaces but you can see some future cleavage planes inside the crystal. The top-right photo shows what you can create if you take a crystal like the one on the left and carefully break it along its cleavage planes.

Density

Density is a measure of the mass of a mineral per unit volume, and it is a useful diagnostic tool in some cases. Most common minerals, such as quartz, feldspar, calcite, amphibole, and mica, have what we call “average density” (2.6 to 3.0 grams per cubic centimetre (g/cm^3)), and it would be difficult to tell them apart on the basis of their density. On the other hand, many of the metallic minerals, such as pyrite, hematite, and magnetite, have densities over 5 g/cm^3. They can easily be distinguished from the lighter minerals on the basis of density, but not necessarily from each other. A limitation of using density as a diagnostic tool is that one cannot assess it in minerals that are a small part of a rock that is mostly made up of other minerals.

Other Properties

Several other properties are also useful for identification of some minerals. For example, calcite is soluble in dilute acid and will give off bubbles of carbon dioxide. Magnetite is magnetic, so will affect a magnet. A few other minerals are weakly magnetic.

Image Descriptions

Figure 13.5.3 image description

	Talc	Gypsum	Calcite	Fluorine	Apatite	Feldspar	Quartz	Topaz	Corundum	Diamond
Measured Hardness	50	60	105	200	659	700	1100	1648	2085	7000
Mohs Hardness	1	2	3	4	5	6	7	8	9	10

[Return to Figure 13.5.3]

Media Attributions

- Figure 13.5.2: © Karla Panchuk. CC BY.
- Figure 13.5.4a: Quartz Bresil © Didier Descouens. CC BY.
- Figure 13.5.4b: Pyrite cubic crystals on marlstone © Carles Millan. CC BY-SA.
- Figure 13.5.4c: Almandine garnet © Eurico Zimbres (FGEL/UERJ) and Tom Epaminondas (mineral collector). CC BY-SA.

This page titled 13.5: Mineral Properties is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Steven Earle (BCCampus) via source content that was edited to the style and standards of the LibreTexts platform.

- 2.6: Mineral Properties by Steven Earle is licensed CC BY 4.0. Original source: https://opentextbc.ca/physicalgeology2ed.

14: ROCKS AND GEOLOGICAL TIME

14.1: The Three Rock Types

14.1.1: Classification of Igneous Rock

14.1.2: Clastic Sedimentary Rocks

14.1.3: Chemical Sedimentary Rocks

14.1.4: Sedimentary Structures and Fossils

14.1.5: Classification of Metamorphic Rocks

14.2: The Rock Cycle

14.3: Reading Earth's History

14.3.1: Geological Time

14.3.2: The Geological Time Scale

14.3.3: Relative Dating Methods

14.3.4: Isotopic Dating Methods

14.3.5: Understanding Geological Time

Chapter 14: Rocks and Geological Time — Introduction

In the previous chapter, we examined minerals—the basic building blocks of rocks. We now take the next step by looking at the rocks themselves and the processes that continually create, alter, and recycle Earth materials. Rocks are more than collections of minerals; their composition, texture, and structure preserve evidence about the conditions under which they formed.

We begin with the **three major rock types: igneous, sedimentary, and metamorphic**. Each forms through different geological processes, but these categories are not permanent. Through melting, cooling, weathering, erosion, burial, heat, and pressure, one type of rock can eventually become another. The **rock cycle** connects these processes and shows us that Earth's surface and interior are parts of a continuously changing system.

Rocks also provide one of our most important records of the past. Their relationships, structures, fossils, and radioactive isotopes allow geologists to determine the sequence of geological events and estimate when those events occurred. We will therefore finish the chapter by learning how to **read Earth's history from rocks**, moving from relative dating to numerical ages and the geological time scale. The enormous spans of time involved can be difficult to visualize, but they are essential for understanding the dynamic Earth we will explore in the final chapter.

This page titled 14: Rocks and Geological Time is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Robert Wagner.

14.1: The Three Rock Types

Rocks are classified into three major groups—igneous, sedimentary, and metamorphic—based on how they form. This section examines the processes that produce each rock type and the characteristics geologists use to identify and classify them. It also introduces sedimentary structures and fossils, which preserve important evidence about the environments and conditions that existed when rocks formed.

This page titled 14.1: The Three Rock Types is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Robert Wagner.

14.1.1: Classification of Igneous Rock

As has already been described, igneous rocks are classified into four categories: felsic, intermediate, mafic, and ultramafic, based on either their chemistry or their mineral composition. The diagram in Figure 14.1.1.1 can be used to help classify igneous rocks by their mineral composition. An important feature to note on this diagram is the red line separating the non-ferromagnesian silicates in the lower left (K-feldspar, quartz, and plagioclase feldspar) from the ferromagnesian silicates in the upper right (biotite, amphibole, pyroxene, and olivine). In classifying intrusive igneous rocks, the first thing to consider is the percentage of ferromagnesian silicates. In most igneous rocks the ferromagnesian silicate minerals are clearly darker than the others, but it is still quite difficult to estimate the proportions of minerals in a rock.

Based on the position of the red line in Figure 14.1.1.1, it is evident that felsic rocks can have between 1% and 20% ferromagnesian silicates (the red line intersects the left side of the felsic zone 1% of the distance from the top of the diagram, and it intersects the right side of the felsic zone 20% of the distance from the top). Intermediate rocks have between 20% and 50% ferromagnesian silicates, and mafic rocks have 50% to 100% ferromagnesian silicates. To be more specific, felsic rocks typically have biotite and/or amphibole; intermediate rocks have amphibole and, in some cases, pyroxene; and mafic rocks have pyroxene and, in some cases, olivine.

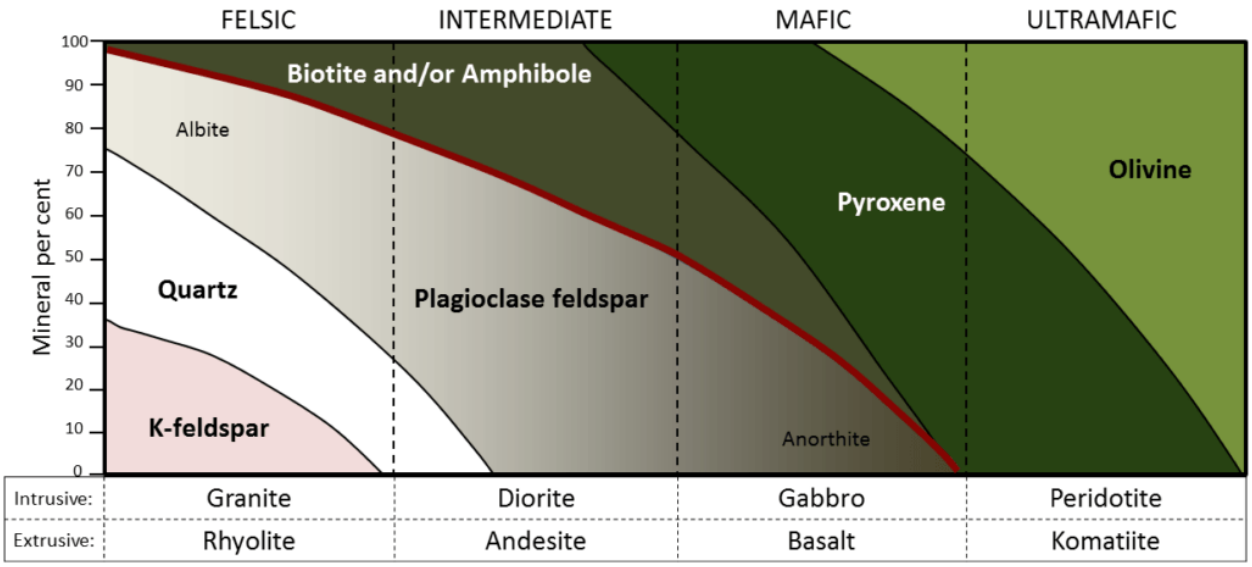


Figure 14.1.1.1 A simplified classification diagram for igneous rocks based on their mineral compositions.

If we focus on the non-ferromagnesian silicates, it is evident that felsic rocks can have from 0% to 35% K-feldspar, from 25% to 35% quartz (the vertical thickness of the quartz field varies from 25% to 35%), and from 25% to 50% plagioclase (and that plagioclase will be sodium-rich, or albitic). Intermediate rocks can have up to 25% quartz and 50% to 75% plagioclase. Mafic rocks only have plagioclase (up to 50%), and that plagioclase will be calcium-rich, or anorthitic.

Exercise 3.5 Mineral proportions in igneous rocks

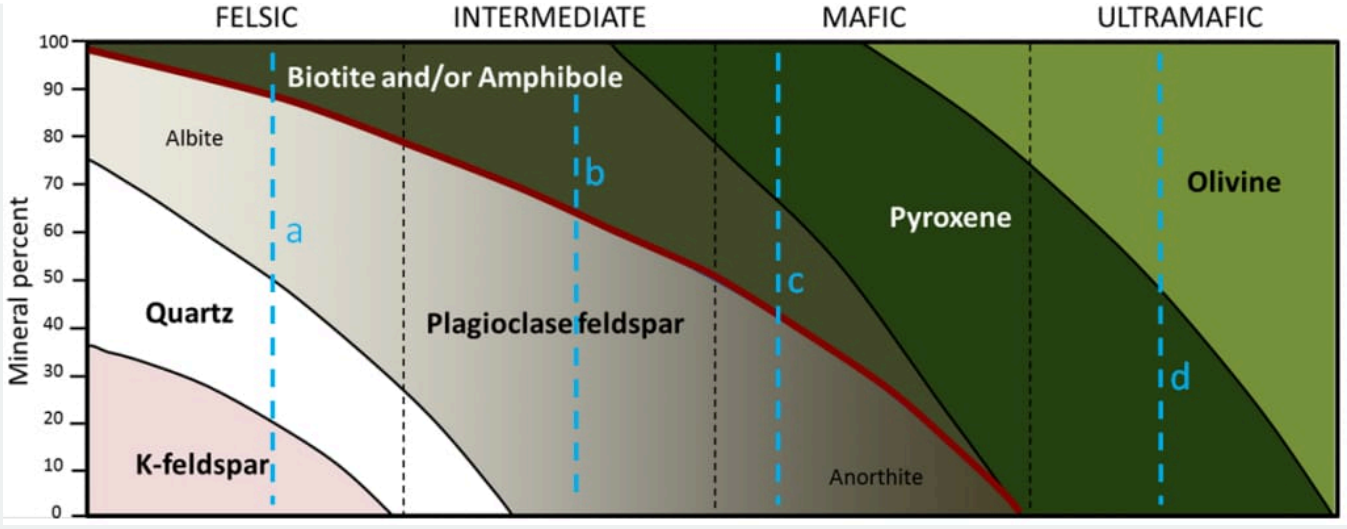


Figure 14.1.1.2

The dashed blue lines (labelled a, b, c, d) in Figure 14.1.1.2 represent four igneous rocks. Complete the table by estimating the mineral proportions (percent) of the four rocks (to the nearest 10%).

Hint: Rocks **b** and **d** are the easiest; start with those.

Rock	Biotite/amphibole	Pyroxene	Olivine	Plagioclase	Quartz	K-feldspar
a						
b						
c						
d						

See Appendix 3 for Exercise 3.5 answers.

Figure 14.1.1.3 provides a diagrammatic representation of the proportions of dark minerals in light-colored rocks. You can use that when trying to estimate the ferromagnesian mineral content of actual rocks, and you can get some practice doing that by completing Exercise 3.6. Be warned! Geology students almost universally over-estimate the proportion of dark minerals.

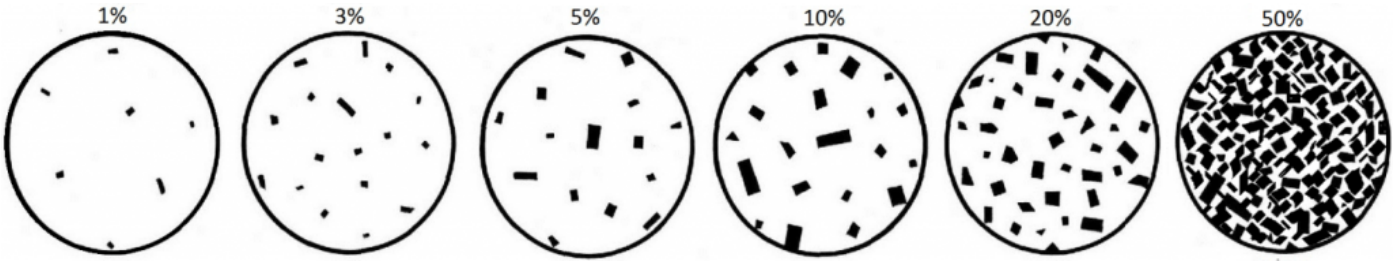
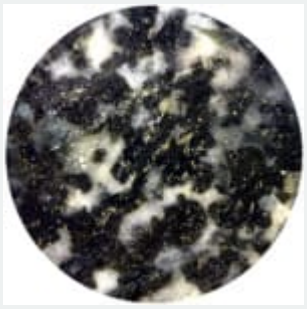
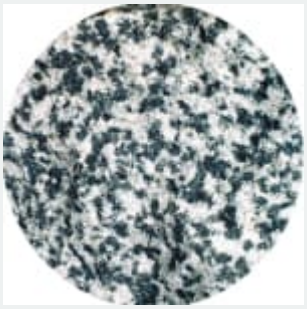


Figure 14.1.1.3 A guide to estimating the proportions of dark minerals in light-colored rocks.

Exercise 3.6 Proportions of ferromagnesian silicates

The four igneous rocks shown below have differing proportions of ferromagnesian silicates. Estimate those proportions using the diagrams in Figure 14.1.1.3, and then use Figure 14.1.1.1 to determine the likely rock name for each one.

			
___%	___%	___%	___%
_____	_____	_____	_____

See Appendix 3 for Exercise 3.6 answers.

Igneous rocks are also classified according to their textures. The textures of volcanic rocks will be discussed in Chapter 4, so here we'll only look at the different textures of intrusive igneous rocks. Almost all intrusive igneous rocks have crystals that are large enough to see with the naked eye, and we use the term **phaneritic** (from the Greek word phaneros meaning visible) to describe that. Typically that means they are larger than about 0.5 millitmeres (mm) — the thickness of a strong line made with a ballpoint pen. (If the crystals are too small to distinguish, which is typical of most volcanic rocks, we use the term **aphanitic** (from the Greek word *aphanos* – unseen) The intrusive rocks shown in Figure 14.1.1.5 are all phaneritic, as are those shown in Exercise 3.6.

In general, the size of crystals is proportional to the rate of cooling. The longer it takes for a body of magma to cool, the larger the crystals can grow. It is not uncommon to see an intrusive igneous rock with crystals up to 1 centimetre (cm) long. In some situations, especially toward the end of the cooling stage, the magma can become water rich. The presence of liquid water (still liquid at high temperatures because it is under pressure) promotes the relatively easy movement of ions, and this allows crystals to grow large, sometimes to several centimeters (Figure 14.1.1.4). Finally, as already described, if an igneous rock goes through a two-stage cooling process, its texture will be porphyritic (Figure 14.1.1.7).



Figure 14.1.1.4 A pegmatitic rock with large crystals

Image Descriptions

Figure 14.1.1.1 *image description: Mineral composition of igneous rocks*

Igneous Rocks	Felsic	Intermediate	Mafic	Ultramafic
K-feldspar	0 to 35%	0%	0%	0%
Quartz	25 to 35%	0 to 25%	0%	0%
Plagioclase feldspar	25 to 50%	50 to 70%	0 to 50%	0%
Biotite and/or Amphibole	0 to 20%	20 to 40%	0 to 30%	0%
Pyroxene	0%	0 to 20%	20 to 75%	0% to 75%
Olivine	0%	0%	0 to 25 %	25% to 100%
Intrusive	Granite	Diorite	Gabbro	Peridotite
Extrusive	Rhyolite	Andesite	Basalt	Komatiite

[Return to Figure 14.1.1.1]

Attributions

- Figure 14.1.1.1, 3.4.2, 3.4.3: © Steven Earle. CC BY.
- Figure 14.1.1.4: Pegmatite. Public domain.

This page titled 14.1.1: Classification of Igneous Rock is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Steven Earle (BCCampus) via source content that was edited to the style and standards of the LibreTexts platform.

- 3.4: Classification of Igneous Rock by Steven Earle is licensed CC BY 4.0. Original source: https://opentextbc.ca/physicalgeology2ed.

14.1.2: Clastic Sedimentary Rocks

A **clast** is a fragment of rock or mineral, ranging in size from less than a micron[1] (too small to see) to as big as an apartment block. Various types of clasts are shown in Figure 14.1.2.1 and in Exercise 5.3. The smaller ones tend to be composed of a single mineral crystal, and the larger ones are typically composed of pieces of rock. As we've seen in Chapter 5, most sand-sized clasts are made of quartz because quartz is more resistant to weathering than any other common mineral. Many of the clasts that are smaller than sand size (less than $^1/_{16}$th millimetre) are made of clay minerals. Most clasts larger than sand size (greater than 2 millimeters) are actual fragments of rock, and commonly these might be fine-grained rock like basalt or andesite, or if they are bigger, coarse-grained rock like granite or gneiss. Sedimentary rocks that are made up of "clasts" are called clastic sedimentary rocks. A comparable term is "detrital sedimentary rocks".

Grain-Size Classification

Geologists that study sediments and sedimentary rocks use the Udden-Wentworth grain-size scale for describing the sizes of the grains in these materials (Table 6.1).

Table 6.1 The Udden-Wentworth grain-size scale for classifying sediments and the grains that make up sedimentary rocks

[Skip Table]

Type	Description	Size range (millimeters)	Size range (microns)
Boulder	large	1024 and up	
medium	512 to 1024		
small	256 to 512		
Cobble	large	128 to 256	
small	64 to 128		
Pebble (Granule)	very coarse	32 to 64	
coarse	16 to 32		
medium	8 to 16		
fine	4 to 8		
very fine	2 to 4		
Sand	very coarse	1 to 2	1000 to 2000
coarse	0.5 to 1	500 to 1000	
medium	0.25 to 0.5 ($^1/_4$ to $^1/_2$ mm)	250 to 500	
fine	0.125 to 0.25 ($^1/_8$th to $^1/_4$ mm)	125 to 250	
very fine	0.063 to 0.125 (or $^1/_{16}$th to $^1/_8$th mm)	63 to 125	
Silt	very course		32 to 63
course		16 to 32	
medium		8 to 16	
fine		4 to 8	
very fine		2 to 4	
Clay	clay		0 to 2

There are six main grain-size categories; five are broken down into subcategories, with **clay** being the exception. The diameter limits for each successive subcategory are twice as large as the one beneath it. In general, a **boulder** is bigger than a toaster and difficult to lift. There is no upper limit to the size of boulder.[2] A small **cobble** will fit in one hand, a large one in two hands. A **pebble** is something that you could throw quite easily. The smaller ones—known as **granules**—are gravel size, but still you could throw one. You can't really throw a single grain of **sand**. Sand ranges from 2 millimeters down to 0.063 millimeters, and its key characteristic is that it feels "sandy" or gritty between your fingers—even the finest sand grains feel that way. **Silt** is essentially too small for individual grains to be visible, and while sand feels sandy to your fingers, silt feels smooth to your fingers but gritty in your mouth. Clay is so fine that it feels smooth even in your mouth.

Exercise 6.1 Describe the sediment on a beach

Providing that your landscape isn't covered in deep snow at present, visit a beach somewhere nearby—an ocean shore, a lake shore, or a river bank. Look carefully at the size and shape of the beach sediments. Are they sand, pebbles, or cobbles? If they are not too fine, you should be able to tell if they are well rounded or more angular.

The beach in Figure 14.1.2.1 is at Sechelt, B.C. Although there is a range of clast sizes, it's mostly made up of well-rounded cobbles interspersed with pebbles. This beach is subject to strong wave activity, especially when winds blow across the Strait of Georgia from the south. That explains why the clasts are relatively large and are well rounded.



Figure 14.1.2.1 Pebbles on an ocean beach at Sechelt, B.C.

See Appendix 3 for Exercise 6.1 answers.

If you drop a granule into a glass of water, it will sink quickly to the bottom (less than half a second). If you drop a grain of sand into the same glass, it will sink more slowly (a second or two depending on the size). A grain of silt will take several seconds to get to the bottom, and a particle of fine clay may never get there. The rate of settling is determined by the balance between gravity and friction, as shown in Figure 14.1.2.2. Large particles settle quickly because the gravitational force (which is proportional to the mass, and therefore to the volume of the particle) is much greater than the frictional resistance (which is proportional to the surface area of the particle). For smaller particles the difference between gravitational push and frictional resistance is less, so they settle slowly.

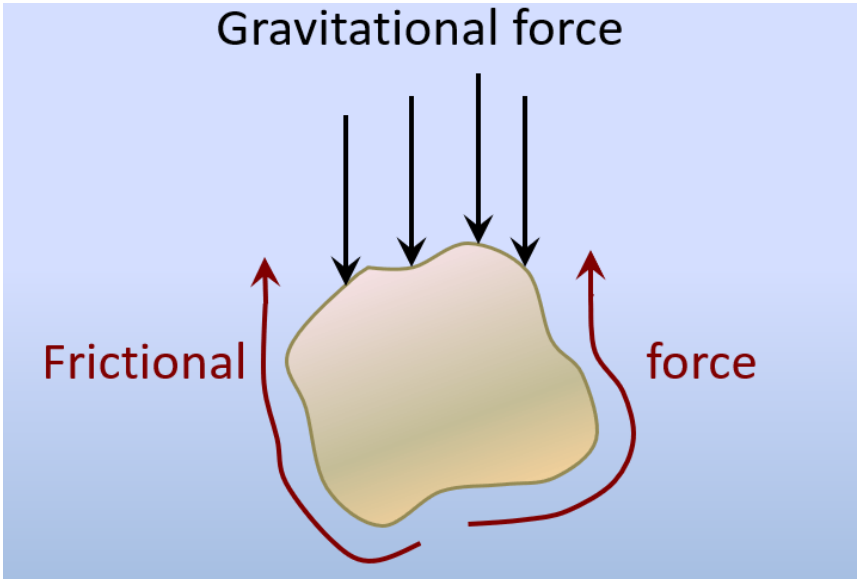


Figure 14.1.2.2 The two forces operating on a grain of sand in water. Gravity is pushing it down, and the friction between the grain and the water is resisting that downward force.

Small particles that settle slowly spend longer suspended in the water, and therefore tend to get moved farther than large particles if the water is moving.

Transportation

One of the key principles of sedimentary geology is that the ability of a moving medium (air or water) to move sedimentary particles—and keep them moving—is dependent on the velocity of flow. The faster the medium flows, the larger the particles it can move. This is illustrated in Figure 14.1.2.3. Parts of the river are moving faster than other parts, especially where the slope is greatest and the channel is narrow. Not only does the velocity of a river change from place to place, but it changes from season to season. During peak **discharge**[3] at the location of Figure 14.1.2.3, the water is high enough to flow over the embankment on the right, and it flows fast enough to move the boulders that cannot be moved during low flows.



Figure 14.1.2.3 Variations in flow velocity on the Englishman River near Parksville, B.C. When the photo was taken the river was not flowing fast enough anywhere to move the boulders and cobbles visible here. During flood events the water flows right over the snow-covered bank on the right, and is fast enough to move boulders.

Clasts within streams are moved in several different ways, as illustrated in Figure 14.1.2.4. Large **bed load** clasts are pushed

(by traction) or bounced along the bottom (by saltation), while smaller clasts are suspended in the water and kept there by the turbulence of the flow. As the flow velocity changes, different-sized clasts may be either incorporated into the flow or deposited on the bottom. At various places along a river, there are always some clasts being deposited, some staying where they are, and some being eroded and transported. This changes over time as the discharge of the river changes in response to changing weather conditions.

Other sediment transportation media, such as waves, ocean currents, and wind, operate under similar principles, with flow velocity as the key underlying factor that controls transportation and deposition.

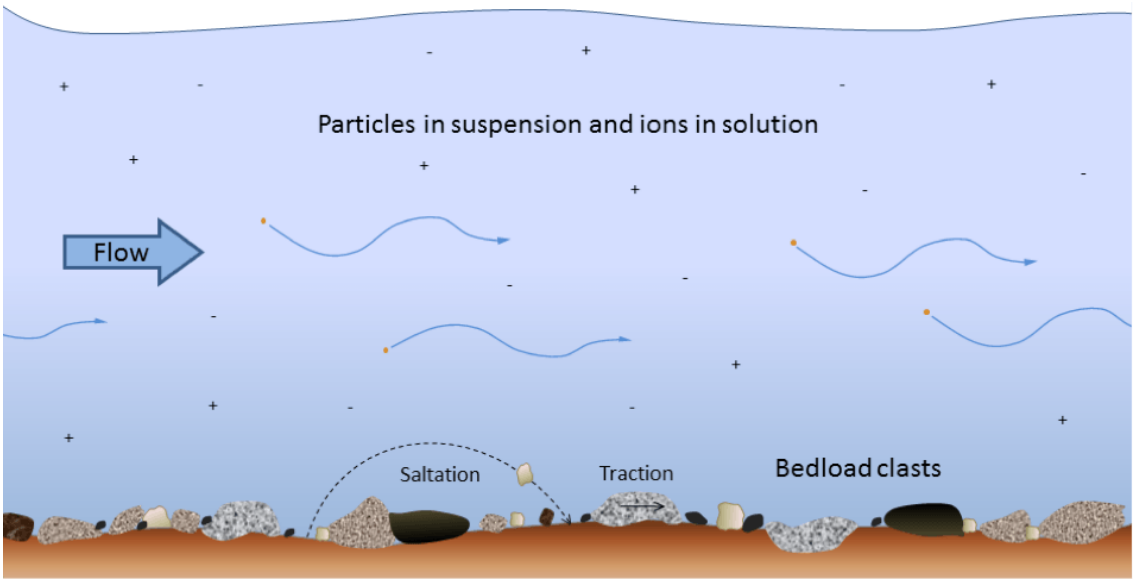


Figure 14.1.2.4 Transportation of sediment clasts by stream flow. The larger clasts, resting on the bottom (bedload), are moved by traction (sliding) or by saltation (bouncing). Smaller clasts are kept in suspension by turbulence in the flow. Ions (depicted as + and – in the image, but invisible in real life) are dissolved in the water.

Clastic sediments are deposited in a wide range of environments, including glaciers, slope failures, rivers—both fast and slow—lakes, deltas, and ocean environments—both shallow and deep. If the sedimentary deposits last long enough to get covered with other sediments they may eventually form into rocks ranging from fine mudstone to coarse breccia and conglomerate.

Lithification is the term used to describe a number of different processes that take place within a deposit of sediment to turn it into solid rock (Figure 14.1.2.5). One of these processes is burial by other sediments, which leads to compaction of the material and removal of some of the intervening water and air. After this stage, the individual clasts are touching one another. **Cementation** is the process of crystallization of minerals within the pores between the small clasts, and especially at the points of contact between clasts. Depending on the pressure, temperature, and chemical conditions, these crystals might include a range of minerals, the common ones being calcite, hematite, quartz and clay minerals.

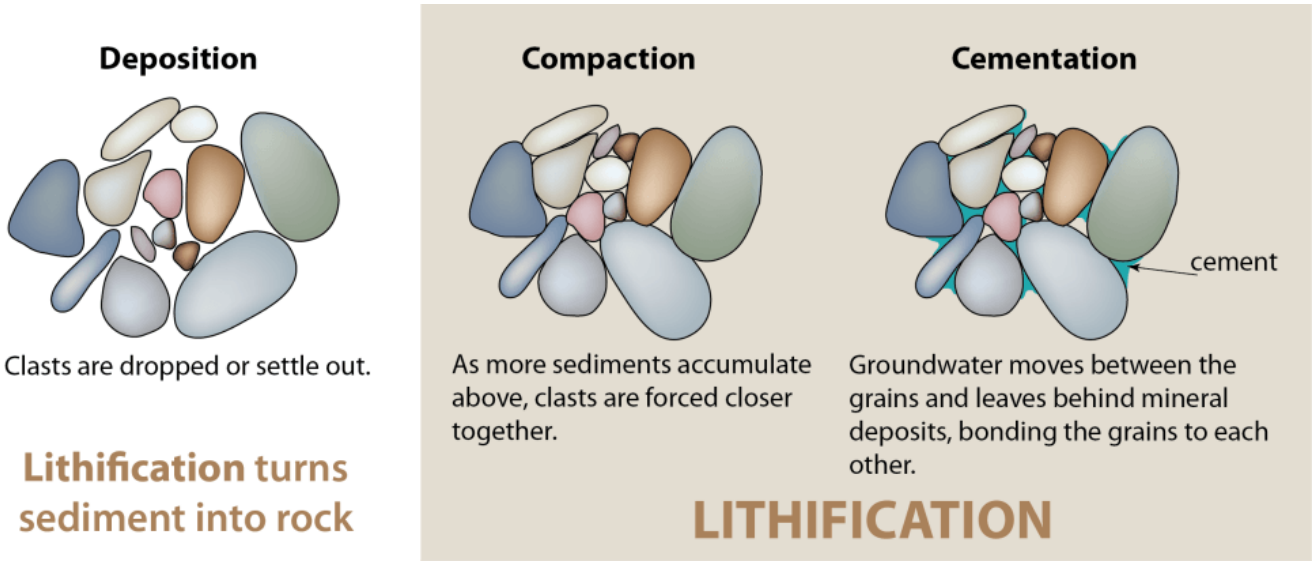


Figure 14.1.2.5 Lithification turns sediments into solid rock. Lithification involves the compaction of sediments and then the cementation of grains by minerals that precipitate from groundwater in the spaces between these grains. *Source: Karla Panchuk (2016) CC BY 4.0*

The characteristics and distinguishing features of clastic sedimentary rocks are summarized in Table 6.2. **Mudrock** is composed of at least 75% silt- and clay-sized fragments. If it is dominated by clay, it is called **claystone**. If it shows evidence of bedding or fine laminations, it is **shale**; otherwise, it is mudstone. Mudrocks form in very low energy environments, such as lakes, river backwaters, and the deep ocean.

Table 6.2 The main types of clastic sedimentary rocks and their characteristics.

[Skip Table]

Group	Examples	Characteristics
Mudrock	mudstone	Greater than 75% silt and clay, not bedded
	shale	Greater than 75% silt and clay, thinly bedded
Coal		Dominated by fragments of partially decayed plant matter often enclosed between beds of sandstone or mudrock.
Sandstone	quartz sandstone	Dominated by sand, greater than 90% quartz
	arkose	Dominated by sand, greater than 10% feldspar
	lithic wacke	dominated by sand, greater than 10% rock fragments, greater than 15% silt and clay
Conglomerate		Dominated by rounded clasts, granule size and larger
Breccia		Dominated by angular clasts, granule size and larger

Most coal forms in fluvial or delta environments where vegetation growth is vigorous and where decaying plant matter accumulates in long-lasting swamps with low oxygen levels. To avoid oxidation and breakdown, the organic matter must remain submerged for centuries or millennia, until it is covered with another layer of either muddy or sandy sediments. It is important to note that in some textbooks coal is described as an “organic sedimentary rock.” In this book, coal is included with the clastic rocks for two reasons: first, because it is made up of fragments of organic matter; and second, because coal seams (sedimentary layers) are almost always interbedded with layers of clastic rocks, such as mudrock or sandstone. In other words, coal accumulates in environments where other clastic rocks accumulate.

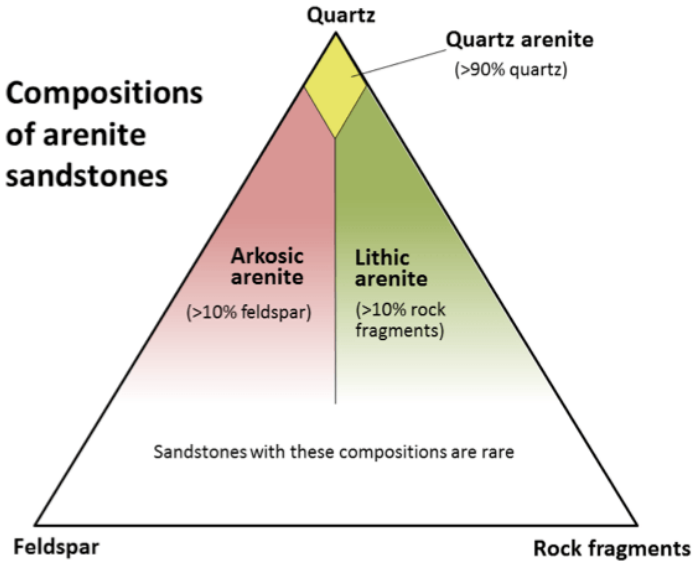


Figure 14.1.2.6 A compositional triangle for arenite sandstones, with the three most common components of sand-sized grains: quartz, feldspar, and rock fragments. Arenites have less than 15% silt or clay. Sandstones with more than 15% silt and clay are called wackes (e.g., quartz wacke, lithic wacke).

It's worth taking a closer look at the different types of sandstone because sandstone is a common and important sedimentary rock. Typical sandstone compositions are shown in Figure 14.1.2.6. Sandstones are mostly made up of sand grains of course, but they also include finer material—both silt and clay. The term **arenite** applies to a so-called clean sandstone, meaning one with less than 15% silt and clay. Considering the sand-sized grains only (the grains larger than $^{1}/_{16}$th mm), arenites with 90% or more quartz are called quartz arenites. If they have more than 10% feldspar and more feldspar than rock fragments, they are called feldspathic arenites or **arkosic arenites** (or just **arkose**). If they have more than 10% rock fragments, and more rock fragments than feldspar, they are **lithic arenites.**[4] A sandstone with more than 15% silt or clay is called a **wacke** (pronounced *wackie*). The terms *quartz wacke*, *lithic wacke*, and *feldspathic wacke* are used with limits similar to those on the arenite diagram. Another name for a lithic wacke is **greywacke**.

Some examples of sandstones, magnified in thin section are shown in Figure 14.1.2.7. (A thin section is rock sliced thin enough so that light can shine through.)

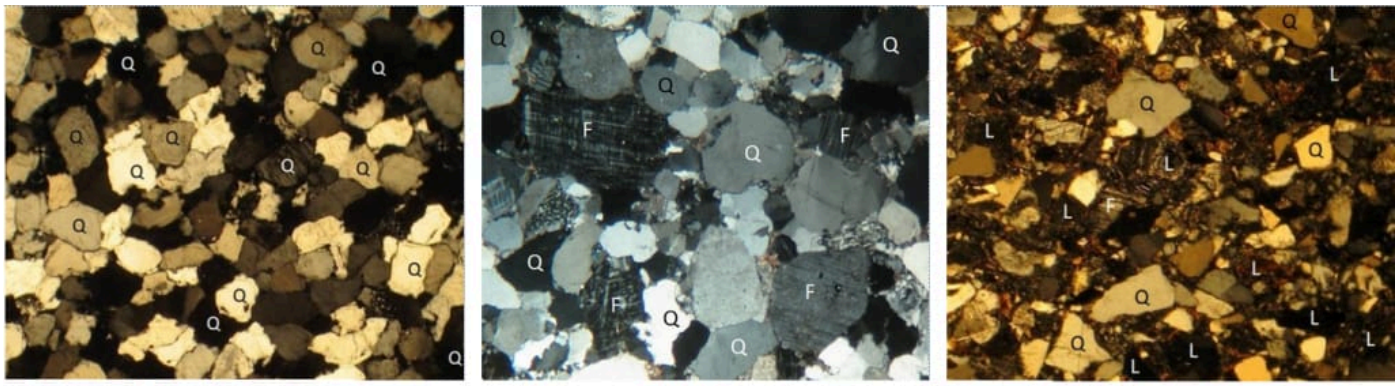


Figure 14.1.2.7 Microscope photos of three types of sandstone in thin-section. Some of the minerals are labelled: Q=quartz, F=feldspar and L= lithic (rock fragments). The quartz arenite and arkose have relatively little silt-clay matrix, while the lithic wacke has abundant matrix.

Clastic sedimentary rocks in which a significant proportion of the clasts are larger than 2 millimeters are known as **conglomerate** if the clasts are well rounded, and **breccia** if they are angular. Conglomerates form in high-energy environments such as fast-flowing rivers, where the particles can become rounded. Breccias typically form where the particles are not transported a significant distance in water, such as alluvial fans and talus slopes. Some examples of clastic sedimentary rocks are shown on Figure 14.1.2.8.



(a) Mudrock with bivalve impressions, Cretaceous Nanaimo Group, Browns River, Vancouver Island

(b) Coarse sandstone with cross-bedding, Cambrian Tapeats Formation Chino Valley, Arizona

(c) Conglomerate with imbricate (aligned, tilted down to the left) cobbles, Cretaceous Geoffrey Formation, Hornby Island, BC

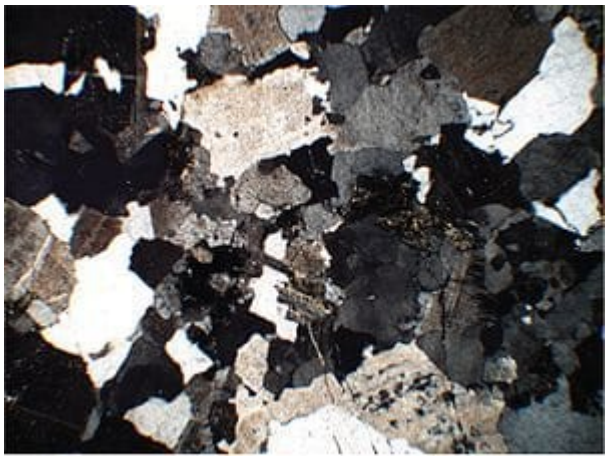
(d) Sedimentary breccia, the Pre-Cambrian Toby Formation, east of Castlegar, BC (image is approx. 1 m across)

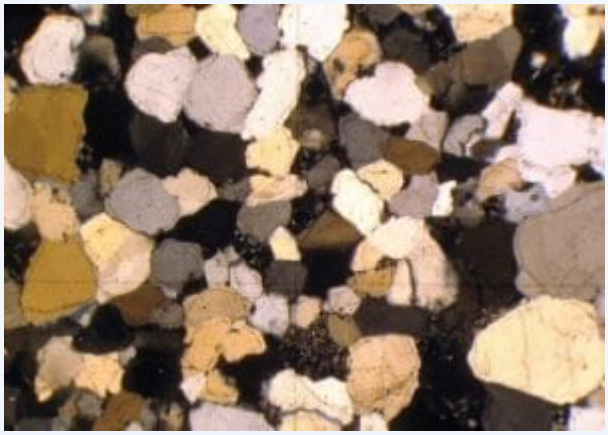
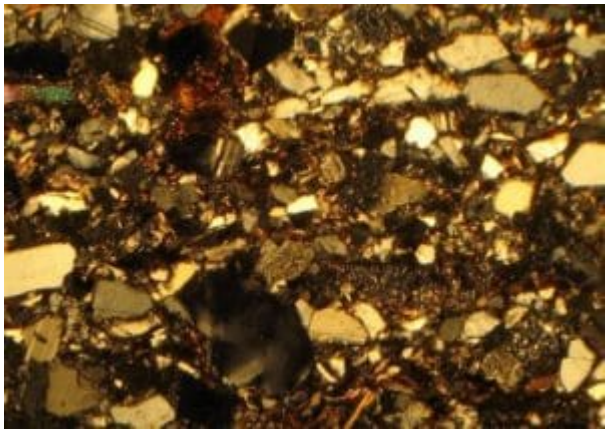
Figure 14.1.2.8 Examples of various clastic sedimentary rocks.

Exercise 6.2 Classifying sandstones

Table 6.3 below shows magnified thin sections of three sandstones, along with descriptions of their compositions. Using Table 6.1 and Figure 14.1.2.6, find an appropriate name for each of these rocks.

Table 6.3 Classifying sandstones

Magnified Thin Section	Description
	Angular sand-sized grains are approximately 85% quartz and 15% feldspar. Silt and clay make up less than 5% of the rock.

Magnified Thin Section	Description
	Rounded sand-sized grains are approximately 99% quartz and 1% feldspar. Silt and clay make up less than 2% of the rock.
	Angular sand-sized grains are approximately 70% quartz, 20% lithic, and 10% feldspar. Silt and clay make up about 20% of the rock.

See Appendix 3 for Exercise 6.2 answers.

Image Descriptions

Figure 14.1.2.8 image description: (A) Mudrock with bivalve impressions, Cretaceous Nanaimo group, Browns River, Vancouver Island. A very fine-grained rock with shell impressions. (B) Coarse sandstone with cross-bedding, Cambrian Tapeats Formation Chino Valley, Arizona. (C) Conglomerate with imbricate (aligned, tilted down to the left) cobbles, Cretaceous Geoffrey Formation, Hornby Island, BC. (D) Sedimentary breccia, the Pre-Cambrian Toby Formation, east of Castlegar, BC. [Return to Figure 14.1.2.8]

Media Attributions

- Figures 6.1.1, 6.1.2, 6.1.3, 6.1.4, 6.1.5, 6.1.6, 6.1.7, 6.1.8: © Steven Earle. CC BY.
- Exercise 6.2, first image: Aplite Red © Rudolf Pohl. CC BY-SA.

---

1. A micron is a millionth of a metre. There are 1,000 microns in a millimetre. ↩
2. The largest known free-standing rock (i.e., not part of bedrock) is Giant Rock in the Mojave Desert, California. It's about as big as an apartment building—seven stories high! ↩
3. Discharge of a stream is the volume of flow passing a point per unit time. It's normally measured in cubic meters per second (m^3/s). ↩
4. "Lithic" means "rock." Lithic clasts are rock fragments, as opposed to mineral fragments. ↩

This page titled 14.1.2: Clastic Sedimentary Rocks is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Steven Earle (BCCampus) via source content that was edited to the style and standards of the LibreTexts platform.

- 6.1: Clastic Sedimentary Rocks by Steven Earle is licensed CC BY 4.0. Original source: https://opentextbc.ca/physicalgeology2ed.

14.1.3: Chemical Sedimentary Rocks

Whereas clastic sedimentary rocks are dominated by components that have been transported as solid clasts (clay, silt, sand, etc.), chemical sedimentary rocks are dominated by components that have been transported as ions in solution (Na^+, Ca^{2+}, HCO_3^-, etc.). There is some overlap between the two because almost all clastic sedimentary rocks contain cement formed from dissolved ions, and many chemical sedimentary rocks include some clasts. Since ions can stay in solution for tens of thousands of years (some much longer), and can travel for tens of thousands of kilometers, it is virtually impossible to relate chemical sediments back to their source rocks.

Chemical weathering and chemical sedimentary rocks

> Many students confuse chemical weathering with chemical sedimentary rocks, or mistakenly assume that when and where chemical weathering is taking place, chemical sedimentary rocks will accumulate. Most ions in solution in rivers, lakes and the ocean are produced during chemical weathering, but those ions can remain in solution for millions of years, and during that time they can travel hundreds of thousands of km (yes, literally around the world, several times). They might eventually come out of solution as a result of a biological process or a change in the chemical conditions and will then become a mineral crystal that can settle to form a chemical sediment.
>
> So the calcium ions that are part of a calcite mud on the sea floor near Australia's Great Barrier Reef could literally have come from anywhere on Earth (and almost certainly came from many different places), and might have been in solution for as little as a few days or for as long as tens of millions of years.

The most common chemical sedimentary rock, by far, is **limestone**. Others include **chert**, **banded iron formation,** and **evaporites**. Biological processes are important in the formation of some chemical sedimentary rocks, especially limestone and chert. For example, limestone is made up almost entirely of fragments of marine[1] organisms that manufacture calcite for their shells and other hard parts, and most chert includes at least some of the silica tests (shells) of tiny marine organisms (such as diatoms and radiolarians).

Limestone

Almost all limestone forms in the oceans, and most of that forms on the shallow continental shelves, especially in tropical regions with coral reefs. Reefs are highly productive ecosystems populated by a wide range of organisms, many of which use calcium and bicarbonate ions in seawater to make carbonate minerals (especially calcite) for their shells and other structures. These include corals, of course, but also green and red algae, urchins, sponges, molluscs, and crustaceans. The hard parts of these organisms are eroded by waves and currents to produce carbonate fragments that accumulate in the surrounding region, as illustrated in Figure 14.1.3.1.

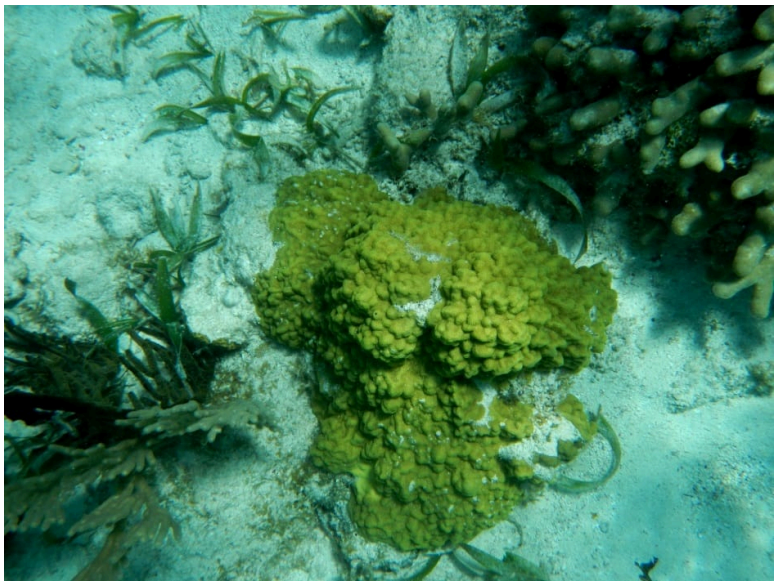


Figure 14.1.3.1 Various corals and green algae on a reef at Ambergris, Belize. The light-colored sand consists of carbonate fragments eroded from the reef organisms, and this is the type of material that will eventually become limestone.

Figure 14.1.3.2 shows a cross-section through a typical reef in a tropical environment (normally between 40° N and 40° S). Reefs tend to form in areas with clear water (e.g., not close to the mouths of large rivers), and near the edges of steep drop-offs because the reef organisms thrive on nutrient-rich upwelling currents. As the reef builds up, it is eroded by waves and currents to produce carbonate sediments that are transported into the steep offshore **fore-reef** area and the shallower inshore **back-reef** area. These sediments are dominated by reef-type carbonate fragments of all sizes, including mud. In many such areas, carbonate-rich sediments also accumulate in quiet lagoons, where mud and mollusc-shell fragments predominate (Figure 14.1.3.3 a) or in offshore areas with strong currents, where either foraminifera tests accumulate (Figure 14.1.3.3b) or calcite crystallizes inorganically to form **ooids**—spheres of calcite that form in shallow tropical ocean water with strong currents (Figure 14.1.3.3c).

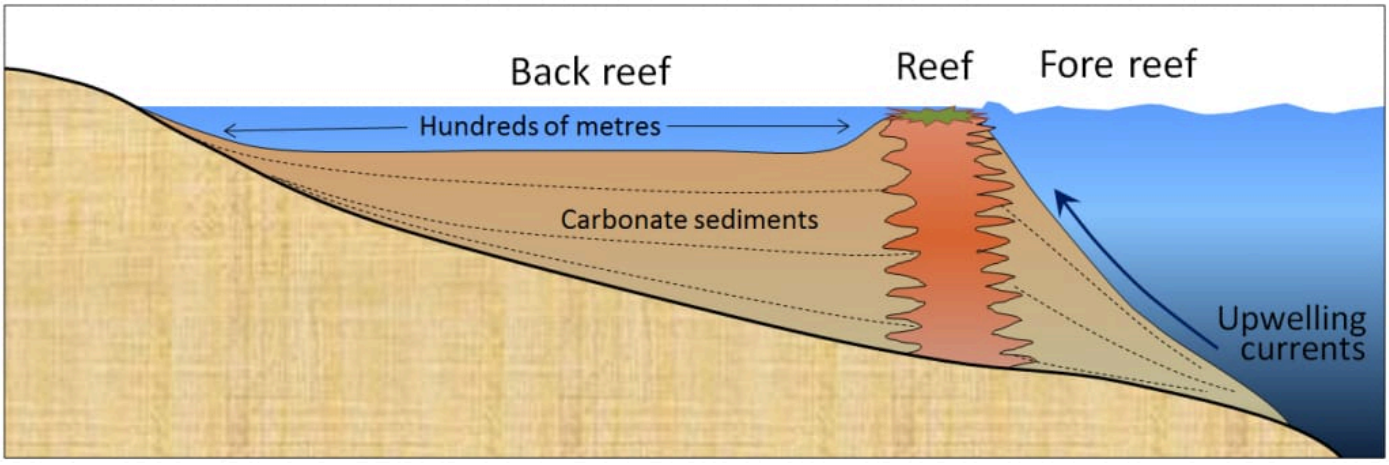


Figure 14.1.3.2 Schematic cross-section through a typical tropical reef.

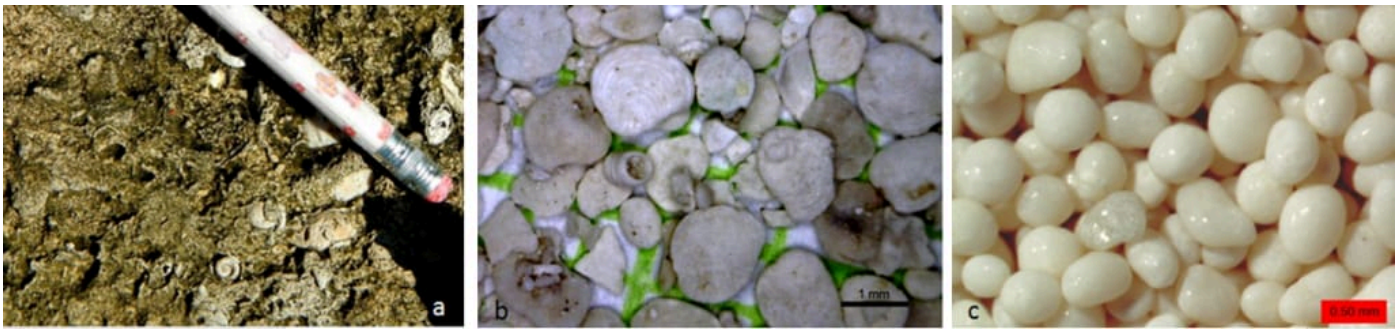


Figure 14.1.3.3 Carbonate rocks and sediments: (a) mollusc-rich limestone formed in a lagoon area at Ambergris, Belize, (b) foraminifera-rich sediment from a submerged carbonate sandbar in Belize (c) ooids from a beach at Joulters Cay, Bahamas.

Limestone also accumulates in deeper water, from the steady rain of the carbonate shells of tiny organisms that lived near the ocean surface. The lower limit for limestone accumulation is around 4,000 meters. Beneath that depth, calcite is soluble so limestone does not accumulate.

Calcite can also form on land in a number of environments. **Tufa** forms at springs (Figure 14.1.3.4) and **travertine** forms at hot springs. Similar material precipitates within limestone caves to form **stalactites**, **stalagmites,** and a wide range of other **speleothems**. Tufa, travertine and speleothems make up only a tiny proportion of all limestone.



Figure 14.1.3.4 Tufa formed at a spring at Johnston Creek, Alberta. The bedded grey rock to the left is limestone.

Dolomite ($CaMg(CO_3)_2$) is another carbonate mineral, but *dolomite* is also the name for a rock composed of the mineral dolomite (although some geologists use the term **dolostone** to avoid confusion). Dolomite rock is quite common (there's a whole Italian mountain range named after it), which is surprising since marine organisms don't make dolomite. All of the dolomite found in ancient rocks has been formed through magnesium replacing some of the calcium in the calcite in carbonate muds and sands. This process is known as **dolomitization**, and it is thought to take place where magnesium-rich water percolates through the sediments in carbonate tidal flat environments.

Chert

As we've seen, not all marine organisms make their hard parts out of calcite; some, like radiolarians and diatoms, use silica, and when they die their tiny shells (or tests) settle slowly to the bottom where they accumulate as chert. In some cases, chert is deposited along with limestone in the moderately deep ocean, but the two tend to remain separate, so chert beds within limestone are quite common (Figure 14.1.3.5), as are nodules, like the flint nodules of the Cretaceous chalk of southeastern England. In other situations, and especially in very deep water, chert accumulates on its own, commonly in thin beds.



Figure 14.1.3.5 Chert (brown layers) interbedded with Triassic Quatsino Fm. limestone on Quadra Island, B.C. All of the layers have been folded, and the chert, being insoluble and harder than limestone, stands out.

Banded iron formation

Banded iron formation (BIF) is a deep sea-floor deposit of iron oxide that is a common ore of iron (Figure 14.1.3.6). BIF forms when iron dissolved in seawater is oxidized, becomes insoluble, and sinks to the bottom in the same way that silica tests do to form chert. BIF is prevalent in rocks dating from 2400 to 1800 Ma, a result off changes in the atmosphere and oceans that took place over that time period. Photosynthetic bacteria (i.e., **cyanobacteria**, a.k.a. blue-green algae) consume carbon dioxide from the atmosphere and use solar energy to convert it to oxygen. These bacteria first evolved around 3500 Ma, and for the next billion years, almost all of that free oxygen was used up by chemical and biological processes, but by 2400 Ma free oxygen levels started to increase in the atmosphere and the oceans. Over a period of 600 million years, that oxygen gradually converted soluble ferrous iron (Fe^{2+}) to insoluble ferric iron (Fe^{3+}), which combined with oxygen to form the mineral hematite (Fe_2O_3), leading to the accumulation of BIFs on the sea floor. After 1800 Ma, little dissolved iron was left in the oceans and the formation of BIF essentially stopped.



Figure 14.1.3.6 An example of a banded iron formation with dark iron oxide layers interspersed with chert stained red by hematite. This rock is 2.1 billion years old.

Evaporites

In arid regions many lakes and inland seas have no stream outlet and the water that flows into them is removed only by evaporation. Under these conditions, the water becomes increasingly concentrated with dissolved salts, and eventually some of these salts reach saturation levels and start to crystallize (Figure 14.1.3.7). Although all evaporite deposits are unique because of differences in the chemistry of the water, in most cases minor amounts of carbonates start to precipitate when the solution is reduced to about 50% of its original volume. Gypsum ($CaSO_4 \cdot H_2O$) precipitates at about 20% of the original volume and halite (NaCl) precipitates at 10%. Other important evaporite minerals include sylvite (KCl) and borax ($Na_2B_4O_7 \cdot 10H_2O$). Sylvite is mined at numerous locations across Saskatchewan (Figure 14.1.3.8) from evaporites that were deposited during the Devonian (~385 Ma) when an inland sea occupied much of the region.



Figure 14.1.3.7 Spotted Lake, near Osoyoos, B.C. The patterns on the surface are salt. This photo was taken in May when the water was relatively fresh because of winter rains. By the end of the summer the surface of this lake is typically fully encrusted with salt deposits.



Figure 14.1.3.8 A mining machine at the face of potash ore (sylvite) in the Lanigan Mine near Saskatoon, Saskatchewan. The mineable potash layer (on the right) is about 3 meters thick.

Exercise 6.3 Making evaporite



Figure 14.1.3.9

This is an easy experiment that you can do at home. Pour about 50 mL (just less than 1/4 cup) of very hot water into a cup and add 2 teaspoons (10 mL) of salt. Stir until all or almost all of the salt has dissolved, then pour the salty water (leaving any undissolved salt behind) into a shallow wide dish or a small plate. Leave it to evaporate for a few days and observe the result. What is the size range and shape of the crystals you grew?

It may look a little like Figure 14.1.3.9. These crystals are up to about 3 millimeters across.

See Appendix 3 for Exercise 6.3 answers.

Media Attributions

- Figures 6.2.1, 6.2.2, 6.2.3ab 6.2.4, 6.2.5, 6.2.7, 6.2.9: © Steven Earle. CC BY.
- Figure 14.1.3.3c: JoultersCayOoids by Wilson44691. Public domain.
- Figure 14.1.3.6: © Andre Karwath. CC BY-SA.
- Figure 14.1.3.8: Photo courtesy of PotashCorp. All rights reserved. Used with permission.

---

1. We use the word *marine* when referring to salt water (i.e., oceanic) environments, and the word *aquatic* when referring to freshwater environments. ↵

This page titled 14.1.3: Chemical Sedimentary Rocks is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Steven Earle (BCCampus) via source content that was edited to the style and standards of the LibreTexts platform.

- 6.2: Chemical Sedimentary Rocks by Steven Earle is licensed CC BY 4.0. Original source: https://opentextbc.ca/physicalgeology2ed.

14.1.4: Sedimentary Structures and Fossils

Through careful observation over the past few centuries, geologists have discovered that the accumulation of sediments and sedimentary rocks takes place according to some important geological principles, as follows:

- The **principle of original horizontality** is that sediments accumulate in essentially horizontal layers. The implication is that tilted sedimentary layers observed to day must have been subjected to tectonic forces.
- The **principle of superposition** is that sedimentary layers are deposited in sequence, and that unless the entire sequence has been turned over by tectonic processes, the layers at the bottom are older than those at the top.
- The **principle of inclusions** is that any rock fragments in a sedimentary layer must be older than the layer. For example, the cobbles in a conglomerate must have been formed before the conglomerate was formed.
- The **principle of faunal succession** is that there is a well-defined order in which organisms have evolved through geological time, and therefore the identification of specific fossils in a rock can be used to determine its age.

In addition to these principles, that apply to all sedimentary rocks (as well as volcanic rocks), a number of other important characteristics of sedimentary processes result in the development of distinctive sedimentary features in specific sedimentary environments. By understanding the origins of these features, we can make some very useful inferences about the processes that led to deposition the rocks that we are studying.

Bedding, for example, is the separation of sediments into layers that either differ from one another in textures, composition, color, or weathering characteristics, or are separated by **partings**—narrow gaps between adjacent beds (Figure 14.1.4.1). Bedding is an indication of changes in depositional processes that may be related to seasonal differences, changes in climate, changes in locations of rivers or deltas, or tectonic changes. Partings may represent periods of non-deposition that could range from a few decades to a few millennia. Bedding can form in almost any sedimentary depositional environment.



Figure 14.1.4.1 The Triassic Sulphur Mt. Formation near Exshaw, Alberta. Bedding is defined by differences in color and texture, and also by partings (gaps) between beds that may otherwise appear to be similar.

Cross-bedding is bedding that contains angled layers within otherwise horizontal beds, and it forms when sediments are deposited by flowing water or wind. Some examples are shown in Figures 6.0.11, 6.1.7b, and 6.4.2. Cross-beds formed in streams tend to be on the scale of centimeters to tens of centimeters, while those in **aeolian** (wind deposited) sediments can be on the scale of meters to several meters.



Figure 14.1.4.2 Cross-bedded Jurassic Navajo Formation aeolian sandstone at Zion National Park, Utah. In most of the layers the cross-beds dip down toward the right, implying a consistent wind direction from right to left during deposition.

Cross-beds form as sediments are deposited on the leading edge of an advancing ripple or dune under steady state conditions (similar flow rate and same flow direction). Each layer is related to a different ripple that advances in the direction of flow, and is partially eroded by the following ripple (Figure 14.1.4.3). Cross-bedding is a very important sedimentary structure to be able to recognize because it can provide information on the process of deposition, the direction of current flows and, when analyzed in detail, on other features like the rate of flow and the amount of sediment available.

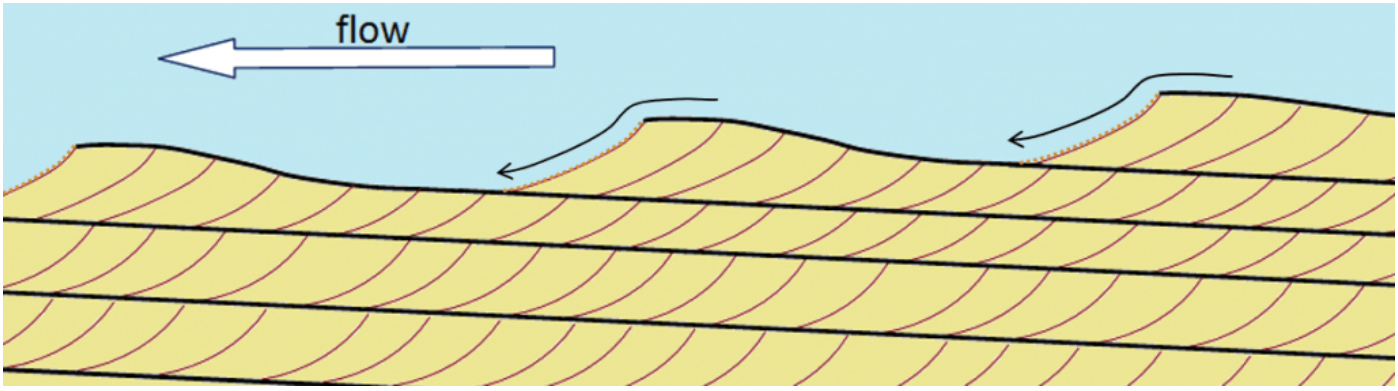


Figure 14.1.4.3 Formation of cross-beds as a series of ripples or dunes migrates with the flow. Each ripple advances forward (right to left in this view) as more sediment is deposited on its leading face (small arrows). (On each ripple the last deposited layer is represented by small dots.)

Graded bedding is characterized by a gradation in grain size from bottom to top within a single bed. “Normal” graded beds are coarse at the bottom and become finer toward the top. They are a product of deposition from a slowing current (Figure 14.1.4.4). Most graded beds form in a submarine-fan environment (see Figure 14.1.4.1), where sediment-rich flows descend periodically from a shallow marine shelf down a slope and onto the deeper sea floor. Some graded beds are reversed (coarser at the top), and this normally results from deposition by a fast-moving debris flow (see Chapter 15).

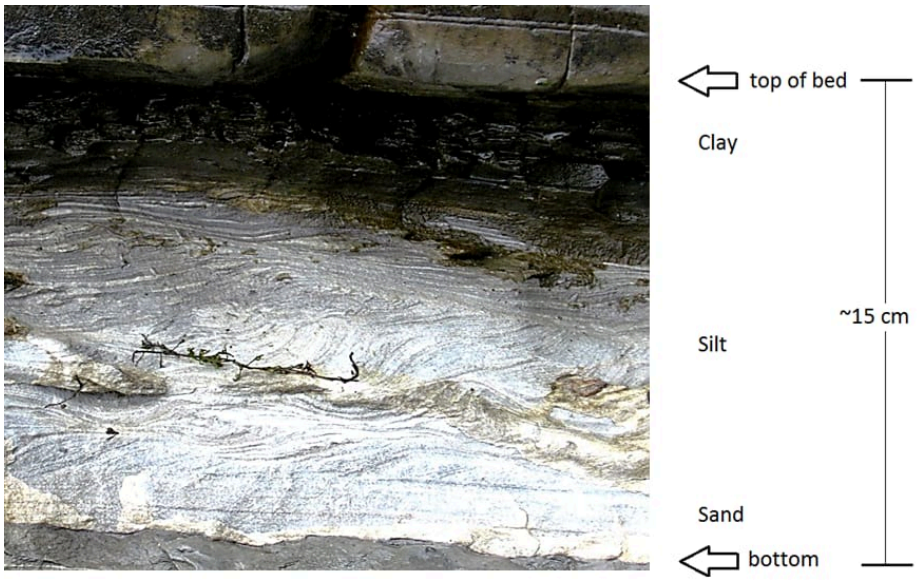


Figure 14.1.4.4 A graded turbidite bed in Cretaceous Spray Formation rocks on Gabriola Island, B.C. The lower several centimeters of sand and silt probably formed over the duration of less than an hour. The upper few centimeters of fine clay may have accumulated over several hundred years.

Ripples, which are associated with the formation of cross-bedding, may be preserved on the surfaces of sedimentary beds. Ripples can also help to determine flow direction as they tend to have their steepest surface facing in the direction of the flow (see Figure 14.1.4.3).

In a stream environment, boulders, cobbles, and pebbles can become **imbricated**, meaning that they are generally tilted in the same direction. Clasts in streams tend to tilt with their upper ends pointing downstream because this is the most stable position with respect to the stream flow (Figure 14.1.4.5 and Figure 14.1.4.7c).

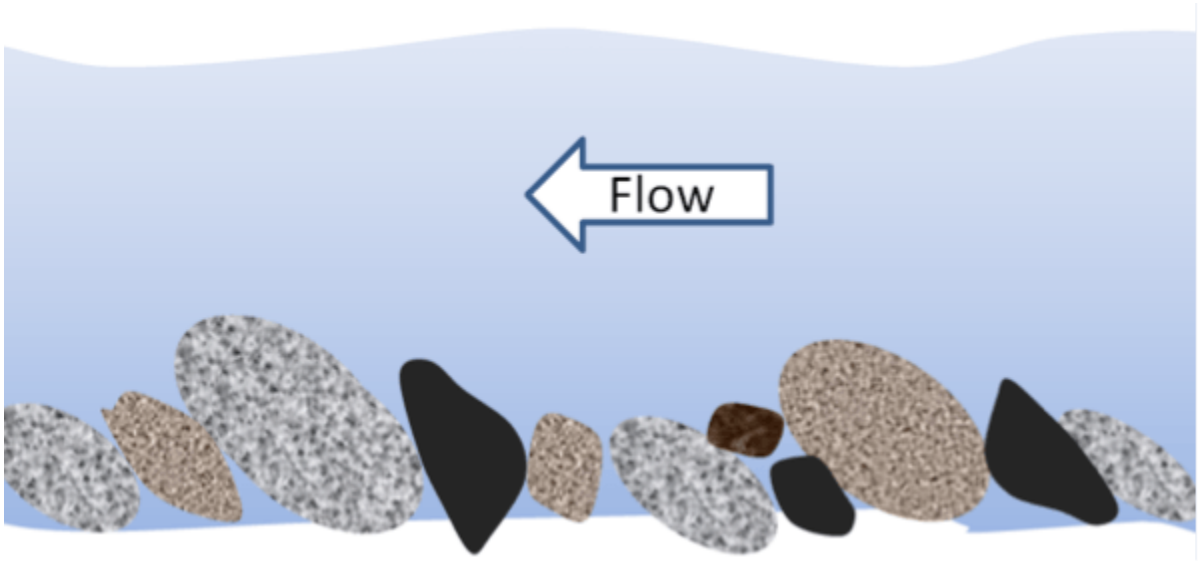


Figure 14.1.4.5 An illustration of imbrication of clasts in a fluvial environment.

Mud cracks form when a shallow body of water (e.g., a tidal flat or pond or even a puddle), into which muddy sediments have been deposited, dries up and cracks (Figure 14.1.4.6). This happens because the clay in the upper mud layer tends to shrink on drying, and so it cracks because it occupies less space when it is dry.



Figure 14.1.4.6 Mudcracks in volcanic mud at a hot-spring area near Myvatn, Iceland.

The various structures described above are critical to understanding and interpreting the conditions that existed during the formation of sedimentary rocks. In addition to these, geologists also look very closely at sedimentary grains to determine their mineralogy or lithology (in order to make inferences about the type of source rock and the weathering processes), their degree of rounding, their sizes, and the extent to which they have been sorted by transportation and depositional processes. Some of the types of differences that we might want to look for are illustrated in Figure 14.1.4.7.

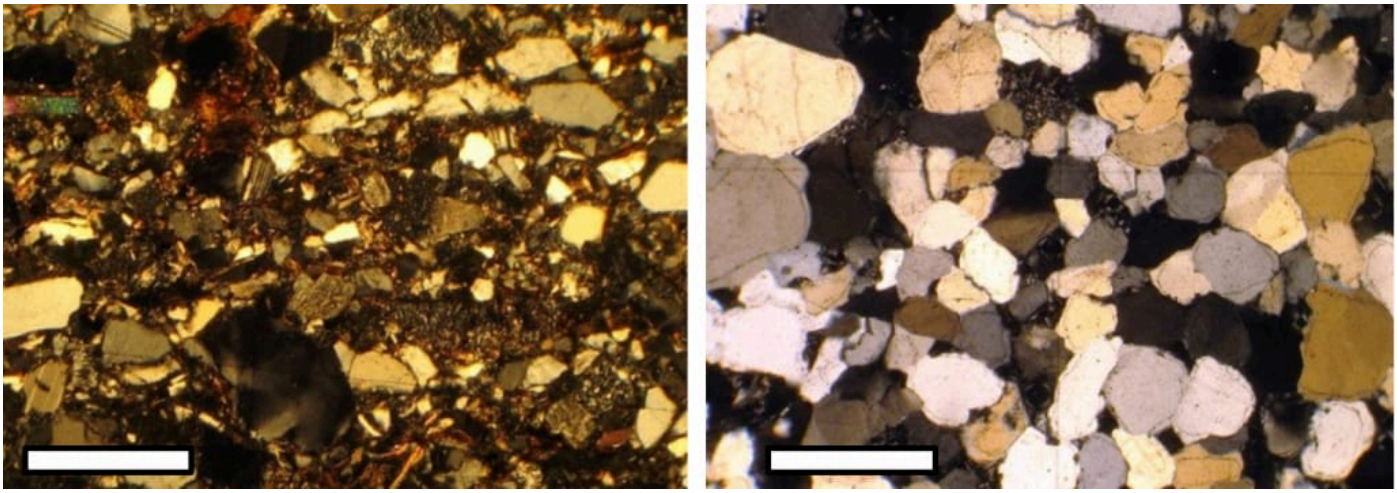


Figure 14.1.4.7 Thin section photos of two sandstones with very different grain characteristics. The one on the left has angular grains with a wide range of different types (quartz, feldspar, biotite, rock fragments), and is poorly sorted (grains range from less than 0.05 mm to ~1 mm). The one on the right has relatively well-rounded grains of quartz only, and the size range is much less (approx. 0.25 to 1 mm). (Scale bars are 1 mm.)

We won't be covering fossils in any detail in this book, but they are extremely important for understanding sedimentary rocks. Of course, fossils can be used to date sedimentary rocks, but equally importantly, they tell us a great deal about the depositional environment of the sediments and the climate at the time. For example, they can help to differentiate marine versus terrestrial environments; estimate the depth of the water; detect the existence of currents; and estimate average temperature and precipitation. For example, the tests of tiny marine organisms (mostly foraminifera) have been recovered from deep-ocean sediment cores from all over the world, and their isotopic signatures have been measured. As we'll see in Chapter 19, this has provided us with information about the changes in average global temperatures over the past 65 million years.

Exercise 6.4 Interpretation of past environments

Sedimentary rocks can tell us a great deal about the environmental conditions that existed during the time of their formation. Make some inferences about the source rock, weathering environment, type and distance of sediment transportation, and deposition conditions that existed during the formation of the following rocks:

1. **Quartz sandstone:** no feldspar, well-sorted and well-rounded quartz grains, cross-bedding
2. **Feldspathic sandstone and mudstone:** feldspar, volcanic fragments, angular grains, repetitive graded bedding from sandstone upwards to mudstone
3. **Conglomerate:** well-rounded pebbles and cobbles of granite and basalt; imbrication
4. **Breccia:** poorly sorted, angular limestone fragments; orange-red matrix

See Appendix 3 for Exercise 6.4 answers.

Media Attributions

- Figures 6.4.1, 6.4.2, 6.4.3, 6.4.4, 6.4.5, 6.4.6, 6.4.7: © Steven Earle. CC BY.

This page titled 14.1.4: Sedimentary Structures and Fossils is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Steven Earle (BCCampus) via source content that was edited to the style and standards of the LibreTexts platform.

- 6.4: Sedimentary Structures and Fossils by Steven Earle is licensed CC BY 4.0. Original source: https://opentextbc.ca/physicalgeology2ed.

14.1.5: Classification of Metamorphic Rocks

There are two main types of metamorphic rocks: those that are foliated because they have formed in an environment with either directed pressure or shear stress, and those that are not foliated because they have formed in an environment without directed pressure or relatively near the surface with very little pressure at all. Some types of metamorphic rocks, such as quartzite and marble, which can form whether there is directed-pressure or not, do not typically exhibit foliation because their minerals (quartz and calcite respectively) do not tend to show alignment (see Figure 14.1.5.8).

When a rock is squeezed under directed pressure during metamorphism it is likely to be deformed, and this can result in a textural change such that the minerals appear elongated in the direction perpendicular to the main stress (Figure 14.1.5.1). This contributes to the formation of foliation.

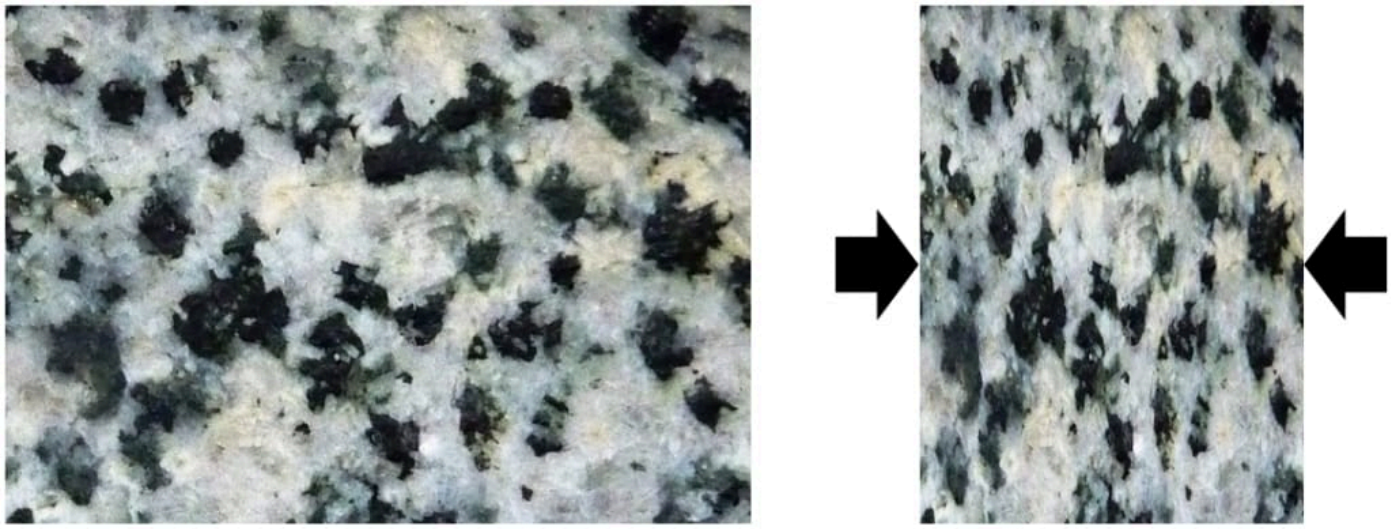


Figure 14.1.5.1 The textural effects of squeezing during metamorphism. In the original rock (left) there is no alignment of minerals. In the squeezed rock (right) the minerals have been elongated in the direction perpendicular to the squeezing.

When a rock is both heated and squeezed during metamorphism, and the temperature change is enough for new minerals to form from existing ones, there is a strong tendency for new minerals to grow with their long axes perpendicular to the direction of squeezing. This is illustrated in Figure 14.1.5.2, where the parent rock is shale, with bedding as shown. After both heating and squeezing, new minerals have formed within the rock, generally parallel to each other, and the original bedding has been largely obliterated.

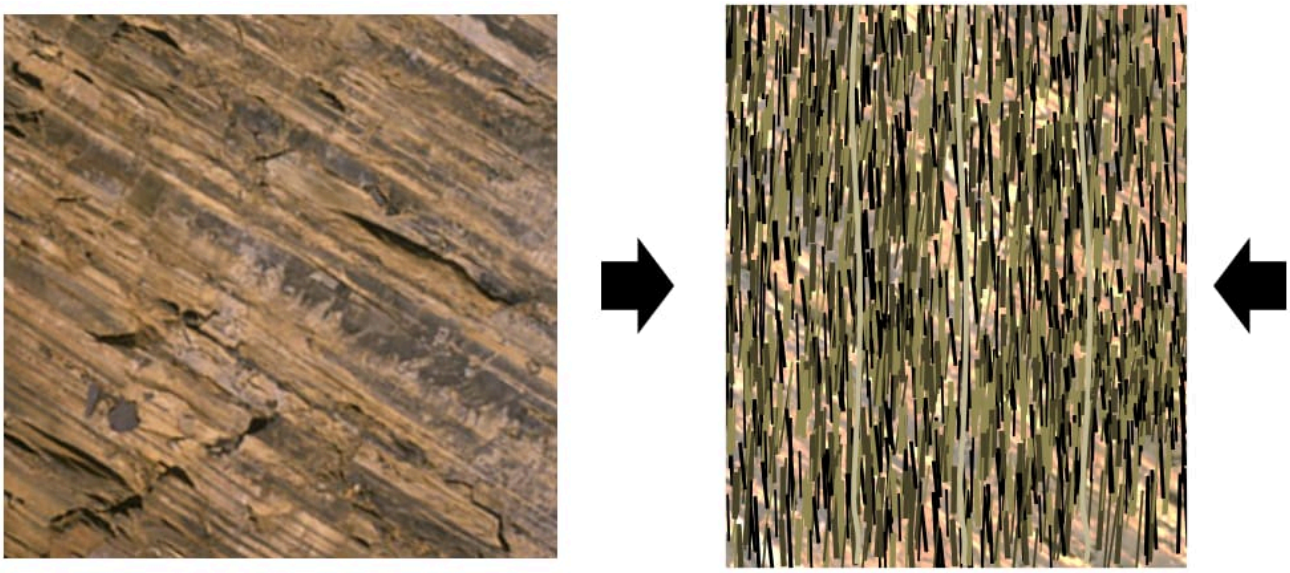


Figure 14.1.5.2 The textural effects of squeezing and mineral growth during regional metamorphism. The left diagram is shale with bedding slanting down to the right. The right diagram represents schist (derived from that shale), with mica crystals orientated perpendicular to the main stress direction and the original bedding no longer easily visible.

Figure 14.1.5.3 shows an example of this effect. This large boulder has bedding visible as dark and light bands sloping steeply down to the right. The rock also has a strong slaty foliation, which is horizontal in this view (parallel to the surface that the person is sitting on), and has developed because the rock was being squeezed during metamorphism. The rock has split from bedrock along this foliation plane, and you can see that other weaknesses are present in the same orientation.

Squeezing and heating alone (as shown in Figure 14.1.5.1) can contribute to foliation, but most foliation develops when new minerals are formed and are forced to grow perpendicular to the direction of greatest stress (Figure 14.1.5.2). This effect is especially strong if the new minerals are platy like mica or elongated like amphibole. The mineral crystals don't have to be large to produce foliation. Slate, for example, is characterized by aligned flakes of mica that are too small to see.



Figure 14.1.5.3 A slate boulder on the side of Mt. Wapta in the Rockies near Field, BC. Bedding is visible as light and dark bands sloping steeply to the right (yellow arrows). Slaty cleavage is evident from the way the rock has broken (along the flat surface that the person is sitting on) and also from lines of weakness that are parallel to that same trend (red arrows).

The various types of foliated metamorphic rocks, listed in order of the **grade** or intensity of metamorphism and the type of foliation are: **slate**, **phyllite**, **schist**, and **gneiss** (Figure 14.1.5.4). As already noted, slate is formed from the low-grade metamorphism of shale, and has microscopic clay and mica crystals that have grown perpendicular to the stress. Slate tends to

break into flat sheets. Phyllite is similar to slate, but has typically been heated to a higher temperature; the micas have grown larger and are visible as a sheen on the surface. Where slate is typically planar, phyllite can form in wavy layers. In the formation of schist, the temperature has been hot enough so that individual mica crystals are big enough to be visible, and other mineral crystals, such as quartz, feldspar, or garnet may also be visible. In gneiss, the minerals may have separated into bands of different colors. In the example shown in Figure 14.1.5.4d, the dark bands are largely amphibole while the light-colored bands are feldspar and quartz. Most gneiss has little or no mica because it forms at temperatures higher than those under which micas are stable. Unlike slate and phyllite, which typically only form from mudrock, schist, and especially gneiss, can form from a variety of parent rocks, including mudrock, sandstone, conglomerate, and a range of both volcanic and intrusive igneous rocks.

Schist and gneiss can be named on the basis of important minerals that are present. For example a schist derived from basalt is typically rich in the mineral chlorite, so we call it chlorite schist. One derived from shale may be a muscovite-biotite schist, or just a mica schist, or if there are garnets present it might be mica-garnet schist. Similarly, a gneiss that originated as basalt and is dominated by amphibole, is an amphibole gneiss or, more accurately, an **amphibolite**.



Figure 14.1.5.4 Examples of foliated metamorphic rocks: (A) Slate, (B) Phyllite, (C) Schist, (D) Gneiss.

If a rock is buried to a great depth and encounters temperatures that are close to its melting point, it may partially melt. The resulting rock, which includes both metamorphosed and igneous material, is known as **migmatite** (Figure 14.1.5.5).



Figure 14.1.5.5 Migmatite from Prague, Czech Republic

As already noted, the nature of the parent rock controls the types of metamorphic rocks that can form from it under differing metamorphic conditions. The kinds of rocks that can be expected to form at different metamorphic grades from various parent rocks are listed in Table 7.1. Some rocks, such as granite, do not change much at the lower metamorphic grades because their minerals are still stable up to several hundred degrees.

Table 7.1 A rough guide to the types of metamorphic rocks that form from different parent rocks at different grades of regional metamorphism.

Parent Rock	Very Low Grade (150-300°C)	Low Grade (300-450°C)	Medium Grade (450-550°C)	High Grade (Above 550°C)
Mudrock	slate	phyllite	schist	gneiss
Granite	no change	no change	almost no change	granite gneiss
Basalt	chlorite schist	chlorite schist	amphibolite	amphibolite
Sandstone	no change	little change	quartzite	quartzite
Limestone	little change	marble	marble	marble

Metamorphic rocks that form under either low-pressure conditions or just confining pressure do not become foliated. In most cases, this is because they are not buried deeply, and the heat for the metamorphism comes from a body of magma that has moved into the upper part of the crust. This is **contact metamorphism**. Some examples of non-foliated metamorphic rocks are **marble**, **quartzite**, and **hornfels**.

Marble is metamorphosed limestone. When it forms, the calcite crystals tend to grow larger, and any sedimentary textures and fossils that might have been present are destroyed. If the original limestone was pure calcite, then the marble will likely be white (as in Figure 14.1.5.6), but if it had various impurities, such as clay, silica, or magnesium, the marble could be “marbled” in appearance. Marble that forms during regional metamorphism—and in fact that includes most marble—may or may not develop a foliated texture, but foliation is typically not easy to see in marble.

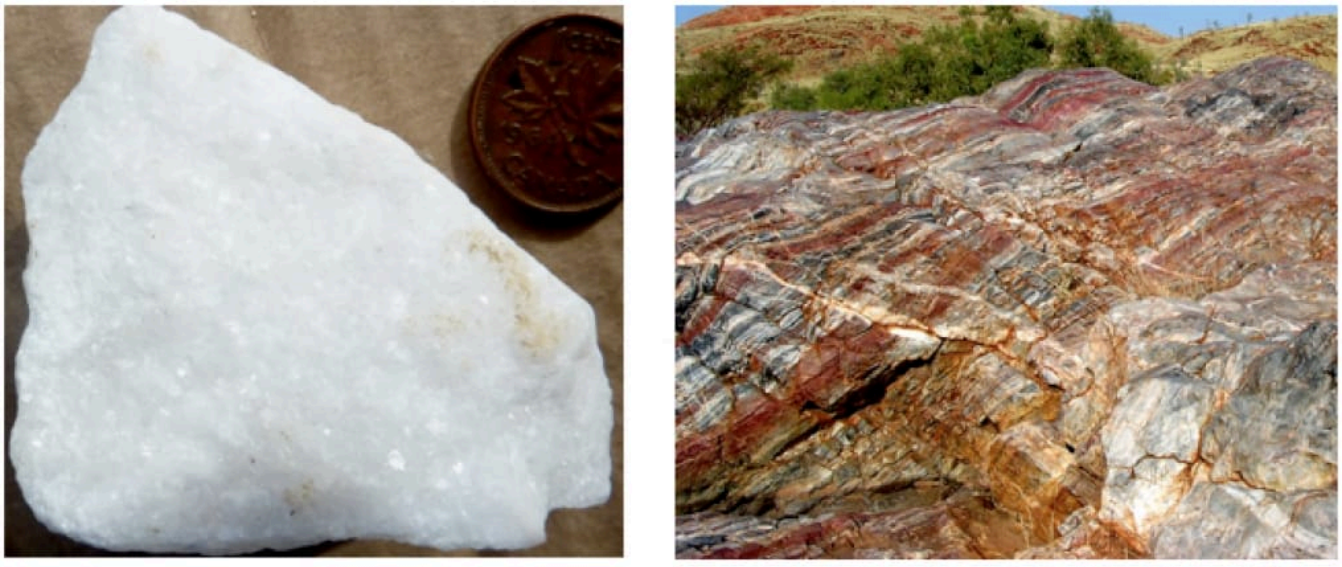


Figure 14.1.5.6 Marble with visible calcite crystals (left) and an outcrop of banded marble (right).

Quartzite is metamorphosed sandstone (Figure 14.1.5.7). It is dominated by quartz, and in many cases, the original quartz grains of the sandstone are welded together with additional silica. Most sandstone contains some clay minerals and may also include other minerals such as feldspar or fragments of rock, so most quartzite has some impurities with the quartz.



Figure 14.1.5.7 Quartzite from the Rocky Mountains, found in the Bow River at Cochrane, Alberta.

Even if formed during **regional metamorphism**, quartzite (like marble) does not tend to look foliated because quartz crystals don't align with the directional pressure. On the other hand, any clay present in the original sandstone is likely to be converted to mica during metamorphism, and any such mica is likely to align with the directional pressure. An example of this is shown in Figure 14.1.5.8. The quartz crystals show no alignment, but the micas are all aligned, indicating that there was directional pressure during regional metamorphism of this rock. Since these micas are very small, this rock would not appear to be foliated to the naked eye.

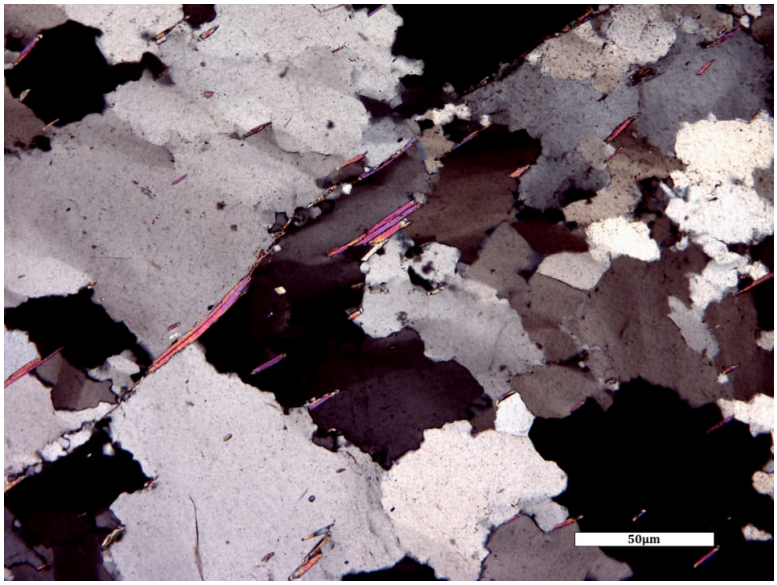


Figure 14.1.5.8 Magnified thin section of quartzite in polarized light. The irregular-shaped white, grey, and black crystals are all quartz. The small, thin, brightly colored crystals are mica. This rock is foliated, even though it might not appear to be if examined without a microscope, and so it must have formed under directed-pressure conditions.

Hornfels is another non-foliated metamorphic rock that normally forms during contact metamorphism of fine-grained rocks like mudstone or volcanic rock (Figure 14.1.5.9). In some cases, hornfels has visible crystals of minerals like biotite or andalusite. If the hornfels formed in a situation without directed pressure, then these minerals would be randomly orientated, not aligned with one-another, as they would be if formed with directed pressure.



Figure 14.1.5.9 Hornfels from the Novosibirsk region of Russia. The dark and light bands are bedding. The rock has been recrystallized during contact metamorphism and does not display foliation. (scale in centimeters).

Exercise 7.2 Naming metamorphic rocks

Provide reasonable names for the following metamorphic rocks based on the description:

1. A rock with visible crystals of mica and with small crystals of andalusite. The mica crystals are consistently parallel to one another.
2. A very hard rock with a granular appearance and a glassy lustre. There is no evidence of foliation.
3. A fine-grained rock that splits into wavy sheets. The surfaces of the sheets have a sheen to them.
4. A rock that is dominated by aligned crystals of amphibole.

See Appendix 3 for Exercise 7.2 answers.

Media Attributions

- Figures 7.2.1 7.2.2, 7.2.3, 7.2.4abd: © Steven Earle. CC BY.
- Figure 14.1.5.4c: Schist detail © Michael C. Rygel. CC BY-SA.
- Figure 14.1.5.5: Migmatite in Geopark on Albertov © Chmee2. CC BY.
- Figure 14.1.5.6 (right): An outcrop of banded marble by the USGS. Public domain.
- Figure 14.1.5.7: © Steven Earle. CC BY.
- Figure 14.1.5.8: © Sandra Johnstone. CC BY.
- Figure 14.1.5.9: Hornfels by Fed. Public domain.

This page titled 14.1.5: Classification of Metamorphic Rocks is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Steven Earle (BCCampus) via source content that was edited to the style and standards of the LibreTexts platform.

- 7.2: Classification of Metamorphic Rocks by Steven Earle is licensed CC BY 4.0. Original source: https://opentextbc.ca/physicalgeology2ed.

14.2: The Rock Cycle

The rock components of the crust are slowly but constantly being changed from one form to another and the processes involved are summarized in the **rock cycle** (Figure 14.2.1). The rock cycle is driven by two forces: (1) Earth's internal heat engine, which moves material around in the core and the mantle and leads to slow but significant changes within the crust, and (2) the hydrological cycle, which is the movement of water, ice, and air at the surface, and is powered by the sun.

The rock cycle is still active on Earth because our core is hot enough to keep the mantle moving, our atmosphere is relatively thick, and we have liquid water. On some other planets or their satellites, such as the Moon, the rock cycle is virtually dead because the core is no longer hot enough to drive mantle convection and there is no atmosphere or liquid water.

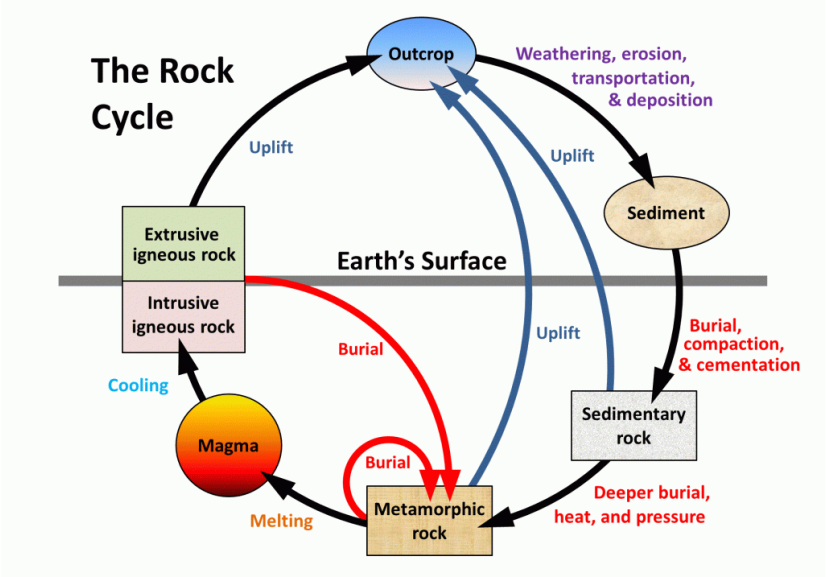


Figure 14.2.1 A schematic view of the rock cycle. The rock cycle takes place both above and below the Earth's surface. The rock deepest beneath the earth's surface, and under extreme heat and pressure, is metamorphic rock. This metamorphic rock can melt and become magma. When magma cools below the earth's surface, it becomes "intrusive igneous rock." If magma cools above the earth's surface, it is "extrusive igneous rock" and becomes part of the outcrop. The outcrop is subject to weathering and erosion, and can be moved and redeposited around the earth by forces such as water and wind. As the outcrop is eroded, it becomes sediment which can be buried, compacted, and cemented beneath the Earth's surface to become sedimentary rock. As sedimentary rock gets buried deeper and comes under increased heat and pressure, it returns to its original state as metamorphic rock. Rocks in the rock cycle do not always make a complete loop. It is possible for sedimentary rock to be uplifted back above the Earth's surface and for intrusive and extrusive igneous rock to be reburied and become metamorphic rock.

In describing the rock cycle, we can start anywhere we like, although it's convenient to start with magma. As we'll see in more detail below, magma is rock that is hot to the point of being entirely molten, with a temperature of between about 800° and 1300°C, depending on the composition and the pressure.



Figure 14.2.2 Magma forming pahoehoe basalt at Kilauea Volcano, Hawaii.

Magma can either cool slowly within the crust (over centuries to millions of years)—forming **intrusive igneous rock**, or erupt onto the surface and cool quickly (within seconds to years)—forming **extrusive igneous rock** (volcanic rock) (Figure 14.2.2). Intrusive igneous rock typically crystallizes at depths of hundreds of meters to tens of kilometers below the surface. To change its position in the rock cycle, intrusive igneous rock has to be uplifted and then exposed by the erosion of the overlying rocks.

Through the various plate-tectonics-related processes of mountain building, all types of rocks are uplifted and exposed at the surface. Once exposed, they are weathered, both physically (by mechanical breaking of the rock) and chemically (by weathering of the minerals), and the weathering products—mostly small rock and mineral fragments—are eroded, transported, and then deposited as **sediments**. Transportation and deposition occur through the action of glaciers, streams, waves, wind, and other agents, and sediments are deposited in rivers, lakes, deserts, and the ocean.

Exercise 3.1 Rock around the rock-cycle clock

Referring to the rock cycle (Figure 14.2.1), list the steps that are necessary to cycle some geological material starting with a sedimentary rock, which then gets converted into a metamorphic rock, and eventually a new sedimentary rock.

A *conservative* estimate is that each of these steps would take approximately 20 million years (some may be less, others would be more, and some could be much more). How long might it take for this entire process to be completed?

See Appendix 3 for Exercise 3.1 Answers.



Figure 14.2.3 Cretaceous-aged marine sandstone overlying marine mudstone, Gabriola Island, B.C.

Unless they are re-eroded and moved along, sediments will eventually be buried by more sediments. At depths of hundreds of meters or more, they become compressed and cemented into **sedimentary rock** (See Figure 14.2.3 for example). Again through various means, largely resulting from plate-tectonic forces, different kinds of rocks are either uplifted, to be re-eroded, or buried deeper within the crust where they are heated up, squeezed, and changed into **metamorphic rock** (Figure 14.2.4)



Figure 14.2.4 Metamorphosed and folded Triassic-aged limestone, Quadra Island, B.C.

Images Attributions

- Figure 14.2.1, 3.1.2, 3.1.3, 3.1.4: © Steven Earle. CC BY.

This page titled 14.2: The Rock Cycle is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Steven Earle (BCCampus) via source content that was edited to the style and standards of the LibreTexts platform.

- 3.1: The Rock Cycle by Steven Earle is licensed CC BY 4.0. Original source: https://opentextbc.ca/physicalgeology2ed.

14.3: Reading Earth's History

Rocks preserve a record of events that have occurred throughout Earth’s long history. This section introduces the geological time scale and examines how geologists determine the sequence and ages of past events. Relative dating methods establish whether rocks and events are older or younger than one another, while isotopic dating provides numerical ages. Together, these methods allow geologists to reconstruct Earth’s history across billions of years.

This page titled 14.3: Reading Earth's History is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by LibreTexts.

14.3.1: Geological Time

In 1788, after many years of geological study, James Hutton, one of the great pioneers of geology, wrote the following about the age of Earth: *The result, therefore, of our present enquiry is, that we find no vestige of a beginning — no prospect of an end.*[1] Of course he wasn't exactly correct, there was a beginning and there will be an end to Earth, but what he was trying to express is that geological time is so vast that we humans, who typically live for less than a century, have no means of appreciating how much geological time there is. Hutton didn't even try to assign an age to Earth, but we now know that it is approximately 4,570 million years old. Using the scientific notation for geological time, that is 4,570 **Ma** (for *mega annum* or "millions of years") or 4.57 **Ga** (for *gigaannum* or billions of years). More recent dates can be expressed in **ka** (*kiloannum*); for example, the last cycle of glaciation ended at approximately 11.7 ka or 11,700 years ago. This notation will be used for geological dates throughout this book.

Exercise 1.3 Using geological time notation

To help you understand the scientific notation for geological time—which is used extensively in this book—write the following out in numbers (for example, 3.23 Ma = 3,230,000 years).

1. 2.75 ka
2. 0.93 Ga
3. 14.2 Ma

We use this notation to describe geological events in the same way that we might say "they arrived at 2 pm." For example, we can say "this rock formed at 45 Ma." But this notation is not used to express elapsed time. We don't say: "I studied for 4 pm for that test." And we don't say: "The dinosaurs lived for 160 Ma." Instead, we could say: "The dinosaurs lived from 225 Ma to 65 Ma, which is 160 million years."

See Appendix 3 for Exercise 1.3 answers.

Unfortunately, knowing how to express geological time doesn't really help us understand or appreciate its extent. A version of the geological time scale is included as Figure 14.3.1.1. Unlike time scales you'll see in other places, or even later in this book, this time scale is linear throughout its length, meaning that 50 Ma during the **Cenozoic** is the same thickness as 50 Ma during the **Hadean**—in each case about the height of the "M" in Ma. The Pleistocene glacial epoch began at about 2.6 Ma, which is equivalent to half the thickness of the thin grey line at the top of the yellow bar marked "Cenozoic." Most other time scales have earlier parts of Earth's history compressed so that more detail can be shown for the more recent parts. That makes it difficult to appreciate the extent of geological time.

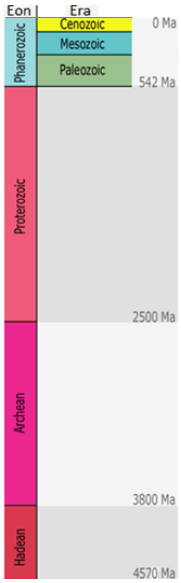


Figure 14.3.1.1 The geological time scale. 1.06: Geological Time#fig1.10

To create some context, the **Phanerozoic** Eon (the last 542 million years) is named for the time during which visible (*phaneros*) life (*zoi*) is present in the geological record. In fact, large organisms—those that leave fossils visible to the naked eye—have existed for a little longer than that, first appearing around 600 Ma, or a span of just over 13% of geological time. Animals have been on land for 360 million years, or 8% of geological time. Mammals have dominated since the demise of the dinosaurs around 65 Ma, or 1.5% of geological time, and the genus *Homo* has existed since approximately 2.8 Ma, or 0.06% (1/1,600th) of geological time.

Geologists (and geology students) need to understand geological time. That doesn't mean memorizing the geological time scale; instead, it means getting your mind around the concept that although most geological processes are extremely slow, very large and important things can happen if such processes continue for enough time.

For example, the Atlantic Ocean between Nova Scotia and northwestern Africa has been getting wider at a rate of about 2.5 centimeters (cm) per year. Imagine yourself taking a journey at that rate—it would be impossibly and ridiculously slow. And yet, since it started to form at around 200 Ma (just 4% of geological time), the Atlantic Ocean has grown to a width of over 5,000 kilometers (km)!

A useful mechanism for understanding geological time is to scale it all down into one year. The origin of the solar system and Earth at 4.57 Ga would be represented by January 1, and the present year would be represented by the last tiny fraction of a second on New Year's Eve. At this scale, each day of the year represents 12.5 million years; each hour represents about 500,000 years; each minute represents 8,694 years; and each second represents 145 years. Some significant events in Earth's history, as expressed on this time scale, are summarized on Table 1.1.

Table 1.1 A summary of some important geological dates expressed as if all of geological time was condensed into one year.

[Skip Table]

Event	Approximate Date	Calendar Equivalent
Formation of oceans and continents	4.5 to 4.4 Ga	January
Evolution of the first primitive life forms	3.8 Ga	early March
Formation of British Columbia's oldest rocks	2.0 Ga	July
Evolution of the first multi-celled animals	0.6 Ga or 600 Ma	November 15
Animals first crawled onto land	360 Ma	December 1
Vancouver Island reached North America and the Rocky Mountains were formed	90 Ma	December 25
Extinction of the non-avian dinosaurs	65 Ma	December 26
Beginning of the Pleistocene ice age	2 Ma or 2000 ka	8 p.m., December 31
Retreat of the most recent glacial ice from southern Canada	14 ka	11:58 p.m., December 31
Arrival of the first people in British Columbia	10 ka	11:59 p.m., December 31
Arrival of the first Europeans on the west coast of what is now Canada	250 years ago	2 seconds before midnight, December 31

Exercise 1.4 Take a trip through geological time

We're going on a road trip! Pack some snacks and grab some of your favourite music. We'll start in Tofino on Vancouver Island and head for the Royal Tyrrell Museum just outside of Drumheller, Alberta, 1,500 km away. Along the way, we'll talk about some important geological sites that we pass by, and we'll use the distance as a way of visualizing the extent of geological time. Of course it's just a "virtual" road trip, but it will be fun anyway. To join in, go to: Virtual Road Trip.

Once you've had a chance to do the road trip, answer these questions:

1. We need oxygen to survive, and yet the first presence of free oxygen (O_2 gas) in the atmosphere and the oceans was a "catastrophe" for some organisms. When did this happen and why was it a catastrophe?
2. Approximately how much time elapsed between the colonization of land by plants and animals?
3. Explain why the evolution of land plants was such a critical step in the evolution of life on Earth.

See Appendix 3 for Exercise 1.4 answers.

Image descriptions

Figure 14.3.1.1 image description: The Hadean eon (3800 Ma to 4570 Ma), Archean eon (2500 Ma to 3800 Ma), and Proterozoic eon (542 Ma to 2500 Ma) make up 88% of geological time. The Phanerozoic eon makes up the last 12% of geological time. The Phanerozoic eon (0 Ma to 542 Ma) contains the Paleozoic, Mesozoic, and Cenozoic eras. [Return to Figure 14.3.1.1]

Media Attributions

- Figure 14.3.1.1: © Steven Earle. CC BY.

1. Hutton, J, 1788. Theory of the Earth; or an investigation of the laws observable in the composition, dissolution, and restoration

of land upon the Globe. Transactions of the Royal Society of Edinburgh. ↵

This page titled 14.3.1: Geological Time is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Steven Earle (BCCampus) via source content that was edited to the style and standards of the LibreTexts platform.

- 1.6: Geological Time by Steven Earle is licensed CC BY 4.0. Original source: https://opentextbc.ca/physicalgeology2ed.

14.3.2: The Geological Time Scale

William Smith worked as a surveyor in the coal-mining and canal-building industries in southwestern England in the late 1700s and early 1800s. While doing his work, he had many opportunities to look at the Paleozoic and Mesozoic sedimentary rocks of the region, and he did so in a way that few had done before. Smith noticed the textural similarities and differences between rocks in different locations, and more importantly, he discovered that fossils could be used to correlate rocks of the same age. Smith is credited with formulating the **principle of faunal succession** (the concept that specific types of organisms lived during different time intervals), and he used it to great effect in his monumental project to create a geological map of England and Wales, published in 1815. For more on William Smith, including a large-scale digital copy of the famous map, see the William Smith Wikipedia page.

Inset into Smith's great geological map is a small diagram showing a schematic geological cross-section extending from the Thames estuary of eastern England all the way to the west coast of Wales. Smith shows the sequence of rocks, from the Paleozoic rocks of Wales and western England, through the Mesozoic rocks of central England, to the Cenozoic rocks of the area around London (Figure 14.3.2.1). Although Smith did not put any dates on these—because he didn't know them—he was aware of the **principle of superposition** (the idea, developed much earlier by the Danish theologian and scientist Nicholas Steno, that young sedimentary rocks form on top of older ones), and so he knew that this diagram represented a stratigraphic column. And because almost every period of the Phanerozoic is represented along that section through Wales and England, it is a primitive geological time scale.

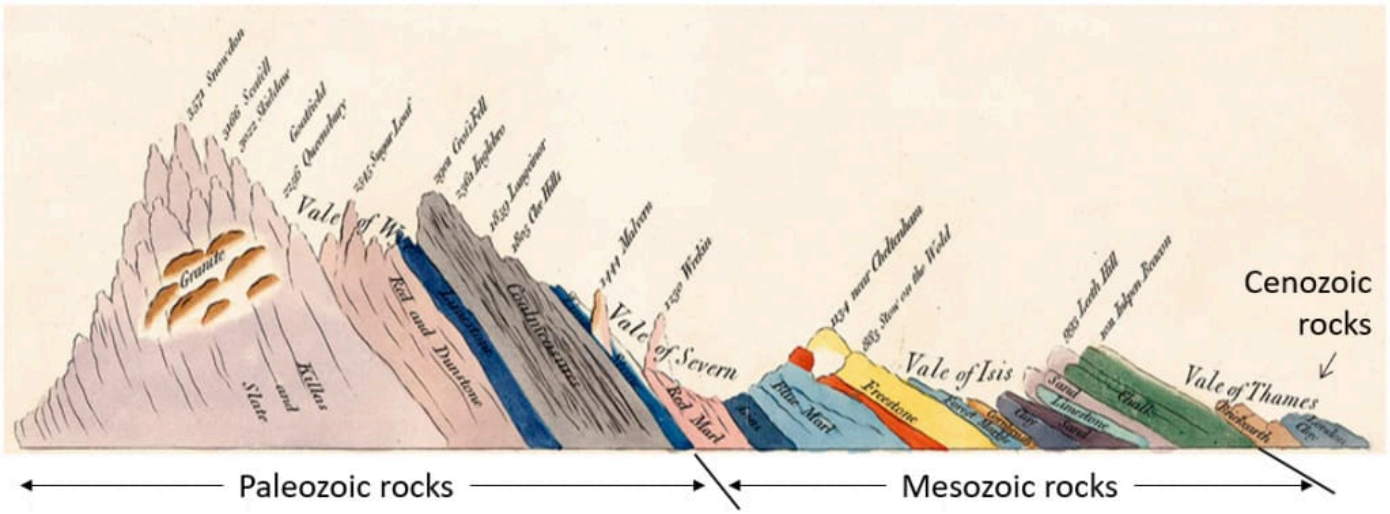


Figure 14.3.2.1 William Smith's "Sketch of the succession of strata and their relative altitudes," an inset on his geological map of England and Wales (with era names added).

Smith's work set the stage for the naming and ordering of the geological periods, which was initiated around 1820, first by British geologists, and later by other European geologists. Many of the periods are named for places where rocks of that age are found in Europe, such as Cambrian for Cambria (Wales), Devonian for Devon in England, Jurassic for the Jura Mountains in France and Switzerland, and Permian for the Perm region of Russia. Some are named for the type of rock that is common during that age, such as Carboniferous for the coal- and carbonate-bearing rocks of England, and Cretaceous for the chalks of England and France.

The early time scales were only relative because 19th century geologists did not know the ages of the rocks. That information was not available until the development of isotopic dating techniques early in the 20th century.

The geological time scale is currently maintained by the International Commission on Stratigraphy (ICS), which is part of the International Union of Geological Sciences. The time scale is continuously being updated as we learn more about the timing and nature of past geological events. You can view the ICS time scale online. It would be a good idea to print a copy (in color) to put on your wall while you are studying geology.

Geological time has been divided into four eons: Hadean (4570 to 4850 Ma), Archean (3850 to 2500 Ma), Proterozoic (2500 to 540 Ma), and Phanerozoic (540 Ma to present). As shown in Figure 14.3.2.2, the first three of these represent almost 90%

of Earth's history. The last one, the Phanerozoic (meaning "visible life"), is the time that we are most familiar with because Phanerozoic rocks are the most common on Earth, and they contain evidence of the life forms that we are familiar with to varying degrees.

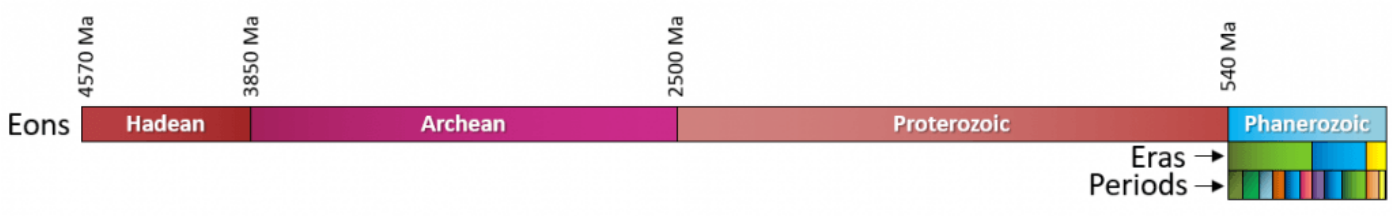


Figure 14.3.2.2 The four eons of Earth's history.

The Phanerozoic eon—the past 540 Ma of Earth's history—is divided into three eras: the Paleozoic ("early life"), the Mesozoic ("middle life"), and the Cenozoic ("new life"), and each of these is divided into a number of periods (Figure 14.3.2.3). Most of the organisms that we share Earth with evolved at various times during the Phanerozoic.

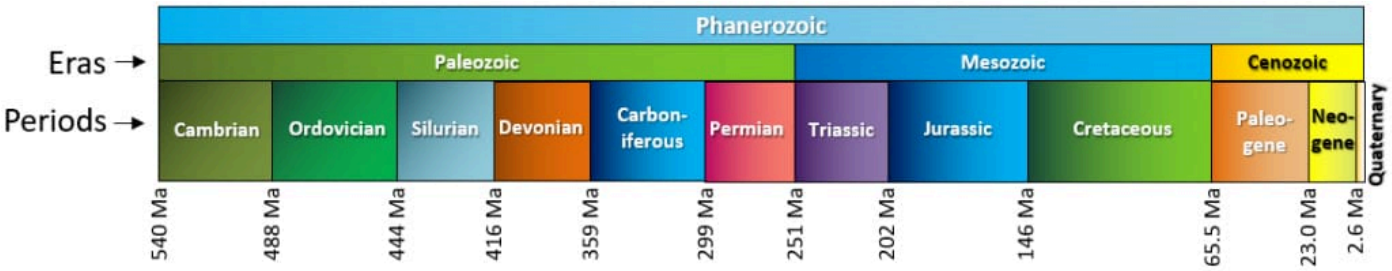


Figure 14.3.2.3 The eras (middle row) and periods (bottom row) of the Phanerozoic eon.

The Cenozoic era, which represents the past 65.5 Ma, is divided into three periods: Paleogene, Neogene, and Quaternary, and seven epochs (Figure 14.3.2.4). Dinosaurs became extinct at the start of the Cenozoic, after which birds and mammals radiated to fill the available habitats. Earth was very warm during the early Eocene and has steadily cooled ever since. Glaciers first appeared on Antarctica in the Oligocene and then on Greenland in the Miocene, and covered much of North America and Europe by the Pleistocene. The most recent of the Pleistocene glaciations ended around 11,700 years ago. The current epoch is known as the Holocene. Epochs are further divided into ages (a.k.a. stages), but we won't be going into that level of detail here.

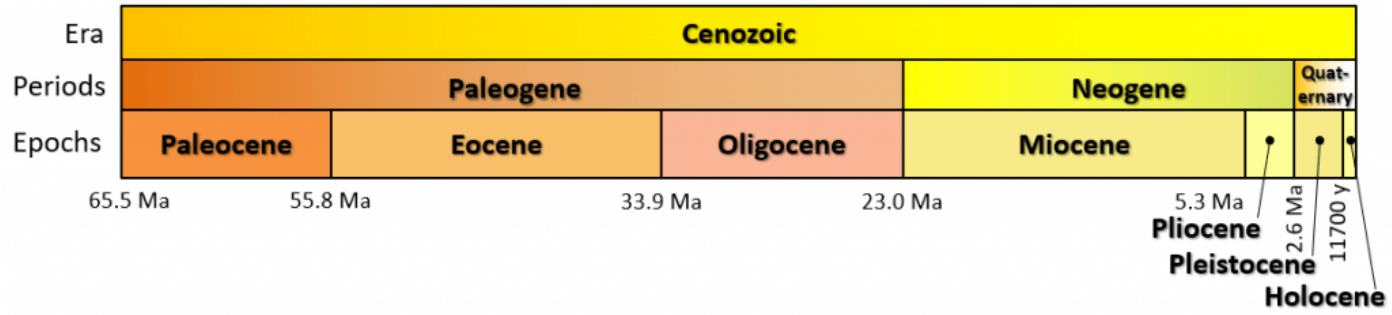


Figure 14.3.2.4 The periods (middle row) and epochs (bottom row) of the Cenozoic era.

Most of the boundaries between the periods and epochs of the geological time scale have been fixed on the basis of significant changes in the fossil record. For example, as already noted, the boundary between the Cretaceous and the Paleogene coincides exactly with a devastating mass extinction. That's not a coincidence. The dinosaurs and many other types of organisms went extinct at this time, and the boundary between the two periods marks the division between sedimentary rocks with Cretaceous organisms (including dinosaurs) below, and Paleogene organisms above.

Image Descriptions

Figure 14.3.2.3 image description: The eras and periods that make up the Phanerozoic Eon.

Era	Period	Time span
Paleozoic	Cambrian	488 to 540 Ma
Paleozoic	Ordovician	488 to 444 Ma
Paleozoic	Silurian	444 to 416 Ma
Paleozoic	Devonian	416 to 359 Ma
Paleozoic	Carboniferous	359 to 299 Ma
Paleozoic	Permian	299 to 251 Ma
Mesozoic	Triassic	251 to 202 Ma
Mesozoic	Jurassic	202 to 146 Ma
Mesozoic	Cretaceous	146 to 65.5 Ma
Cenozoic	Paleogene	65.5 to 23 Ma
Cenozoic	Neogene	23 to 2.6 Ma
Cenozoic	Quaternary	2.6 Ma to present

[Return to Figure 14.3.2.3]

Figure 14.3.2.4 image description: The periods and epochs that make up the Cenozoic era.

Period	Epoch	Time span
Paleogene	Paleocene	65.5 to 55.8 Ma
Paleogene	Eocene	55.8 to 33.9 Ma
Paleogene	Oligocene	33.9 to 23.0 Ma
Neogene	Miocene	23.0 to 5.3 Ma
Neogene	Pliocene	5.3 to 2.6 Ma
Quaternary	Pleistocene	2.6 Ma to 11,700 years ago
Quaternary	Holocene	11,700 years ago to the present

[Return to Figure 14.3.2.4]

Media Attributions

- Figure 14.3.2.1: “Sketch of the succession of strata and their relative altitudes” by William Smith. Adapted by Steven Earle. Public domain.
- Figures 8.1.2, 8.1.3, 8.1.4: © Steven Earle. CC BY.

This page titled 14.3.2: The Geological Time Scale is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Steven Earle (BCCampus) via source content that was edited to the style and standards of the LibreTexts platform.

- 8.1: The Geological Time Scale by Steven Earle is licensed CC BY 4.0. Original source: https://opentextbc.ca/physicalgeology2ed.

14.3.3: Relative Dating Methods

The simplest and most intuitive way of dating geological features is to look at the relationships between them. There are a few simple rules for doing this, some of which we've already looked at in Chapter 6. For example, the principle of superposition states that sedimentary layers are deposited in sequence, and, unless the entire sequence has been turned over by tectonic processes or disrupted by faulting, the layers at the bottom are older than those at the top. The **principle of inclusions** states that any rock fragments that are included in rock must be older than the rock in which they are included. For example, a **xenolith** in an igneous rock or a clast in sedimentary rock must be older than the rock that includes it (Figure 14.3.3.1).



Figure 14.3.3.1a A xenolith of diorite incorporated into a basalt lava flow, Mauna Kea volcano, Hawaii. The lava flow took place some time after the diorite cooled, was uplifted, and then eroded. (geological rock hammer head for scale).



Figure 14.3.3.1b Rip-up clasts of shale embedded in Gabriola Formation sandstone, Gabriola Island, B.C. The pieces of shale were eroded as the sandstone was deposited, so the shale is older than the sandstone.

The **principle of cross-cutting relationships** states that any geological feature that cuts across, or disrupts another feature must be younger than the feature that is disrupted. An example of this is given in Figure 14.3.3.2, which shows three different sedimentary layers. The lower sandstone layer is disrupted by two **faults**, so we can conclude that the faults are younger than that layer. But the faults do not appear to continue into the coal seam, and they certainly do not continue into the upper sandstone. So

we can infer that coal seam is younger than the faults (because it cuts them off), and of course the upper sandstone is youngest of all, because it lies on top of the coal seam.

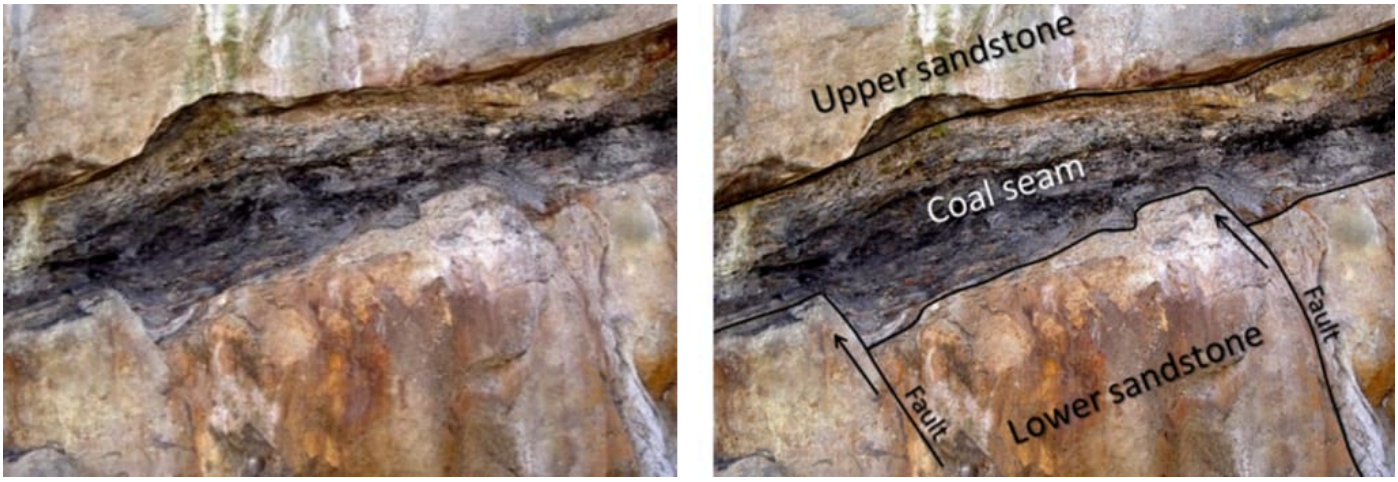


Figure 14.3.3.2 Superposition and cross-cutting relationships in Cretaceous Nanaimo Group rocks in Nanaimo, B.C. The coal seam is about 50 centimeters thick. The sequence of events is as follows: a) deposition of lower sandstone, b) faulting of lower sandstone, c) deposition of coal seam and d) deposition of upper sandstone.

Exercise 8.1 Cross-Cutting Relationships



Figure 14.3.3.3

The outcrop shown here (at Horseshoe Bay, B.C.) has three main rock types:

1. Buff/pink felsic intrusive igneous rock present as somewhat irregular masses trending from lower right to upper left
2. Dark grey metamorphosed basalt
3. A 50 centimeters wide light-grey felsic intrusive igneous dyke extending from the lower left to the middle right – offset in several places

Using the principle of cross-cutting relationships outlined above, determine the relative ages of these three rock types.

(The near-vertical stripes are blasting drill holes. The image is about 7 meters across.)

See Appendix 3 for Exercise 8.1 answers.

An **unconformity** represents an interruption in the process of deposition of sedimentary rocks. Recognizing unconformities is important for understanding time relationships in sedimentary sequences. An example of an unconformity is shown in Figure 14.3.3.4. The Proterozoic rocks of the Grand Canyon Group have been tilted and then eroded to a flat surface prior to deposition of the younger Paleozoic rocks. The difference in time between the youngest of the Proterozoic rocks and the oldest of the Paleozoic rocks is close to 300 million years. Tilting and erosion of the older rocks took place during this time, and if there was any deposition going on in this area, the evidence of it is now gone.



Figure 14.3.3.4 The great angular unconformity in the Grand Canyon, Arizona. The tilted rocks at the bottom are part of the Proterozoic Grand Canyon Group (aged 825 to 1,250 Ma). The flat-lying rocks at the top are Paleozoic (540 to 250 Ma). The boundary between the two represents a time gap of nearly 300 million years.

There are four types of unconformities, as summarized in Table 8.1, and illustrated in Figure 14.3.3.5.

Table 8.1 The characteristics of the four types of unconformities

Unconformity Type	Description
Nonconformity	A boundary between non-sedimentary rocks (below) and sedimentary rocks (above)
Angular unconformity	A boundary between two sequences of sedimentary rocks where the underlying ones have been tilted (or folded) and eroded prior to the deposition of the younger ones (as in Figure 14.3.3.4)
Disconformity	A boundary between two sequences of sedimentary rocks where the underlying ones have been eroded (but not tilted) prior to the deposition of the younger ones (as in Figure 14.3.3.2)
Paraconformity	A time gap in a sequence of sedimentary rocks that does not show up as an angular unconformity or a disconformity

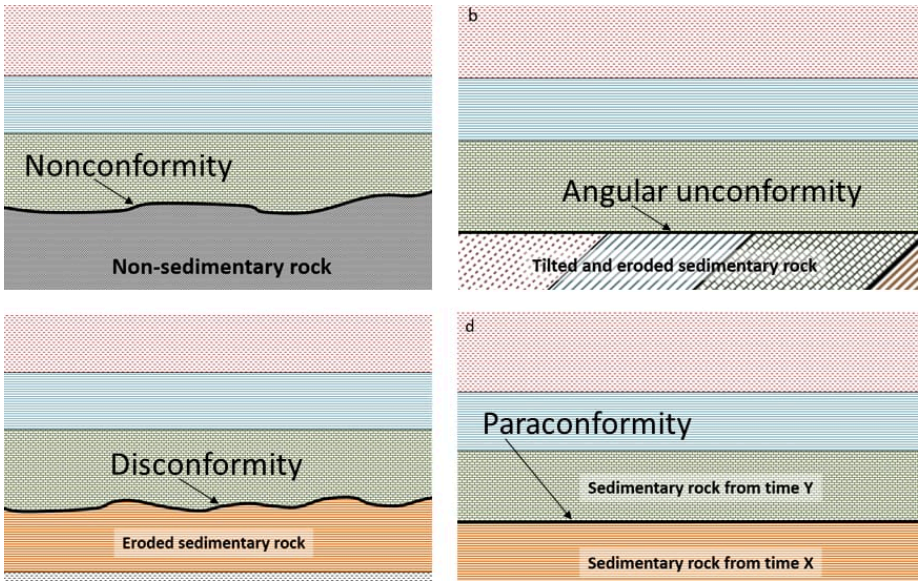


Figure 14.3.3.5 The four types of unconformities: (a) a nonconformity between older non-sedimentary rock and sedimentary rock, (b) an angular unconformity, (c) a disconformity between layers of sedimentary rock, where the older rock has been eroded but not tilted, and (d) a paraconformity where there is a long period (typically millions of years) of non-deposition between two parallel layers.

Media Attributions

- Figures 8.2.1ab, 8.2.2, 8.2.3, 8.2.4, 8.2.5: © Steven Earle. CC BY.

This page titled 14.3.3: Relative Dating Methods is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Steven Earle (BCCampus) via source content that was edited to the style and standards of the LibreTexts platform.

- 8.2: Relative Dating Methods by Steven Earle is licensed CC BY 4.0. Original source: https://opentextbc.ca/physicalgeology2ed.

14.3.4: Isotopic Dating Methods

Originally fossils only provided us with relative ages because, although early paleontologists understood biological succession, they did not know the absolute ages of the different organisms. It was only in the early part of the 20th century, when isotopic dating methods were first applied, that it became possible to discover the absolute ages of the rocks containing fossils. In most cases, we cannot use isotopic techniques to directly date fossils or the sedimentary rocks they are found in, but we can constrain their ages by dating igneous rocks that cut across sedimentary rocks, or volcanic layers that lie within sedimentary layers.

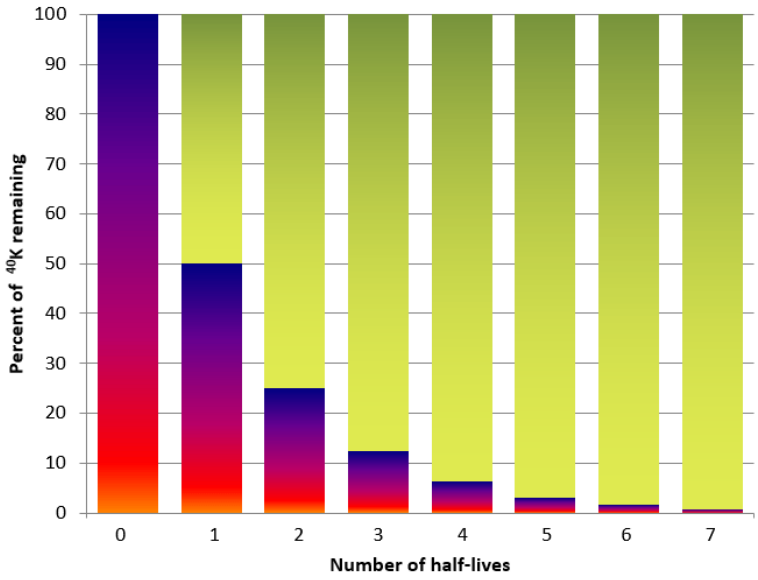


Figure 14.3.4.1 The decay of ^{40}K over time. Each half-life is 1.3 billion years, so after 3.9 billion years (three half-lives) 12.5% of the original ^{40}K will remain. The red-blue bars represent ^{40}K and the green-yellow bars represent ^{40}Ar.

Isotopic dating of rocks, or the minerals in them, is based on the fact that we know the decay rates of certain unstable **isotopes** of elements and that these rates have been constant over geological time. It is also based on the premise that when the atoms of an element decay within a mineral or a rock, they stay there and don't escape to the surrounding rock, water, or air. One of the isotope pairs widely used in geology is the decay of ^{40}K to ^{40}Ar (potassium-40 to argon-40). ^{40}K is a radioactive isotope of potassium that is present in very small amounts in all minerals that have potassium in them. It has a half-life of 1.3 billion years, meaning that over a period of 1.3 Ga one-half of the ^{40}K atoms in a mineral or rock will decay to ^{40}Ar, and over the next 1.3 Ga one-half of the remaining atoms will decay, and so on (Figure 14.3.4.1).

In order to use the K-Ar dating technique, we need to have an igneous or metamorphic rock that includes a potassium-bearing mineral. One good example is granite, which normally has some potassium feldspar (Figure 14.3.4.2). Feldspar does not have any argon in it when it forms. Over time, the ^{40}K in the feldspar decays to ^{40}Ar. Argon is a gas and the atoms of ^{40}Ar remain embedded within the crystal, unless the rock is subjected to high temperatures after it forms. The sample must be analyzed using a very sensitive mass-spectrometer, which can detect the differences between the masses of atoms, and can therefore distinguish between ^{40}K and the much more abundant ^{39}K. Biotite and hornblende are also commonly used for K-Ar dating.

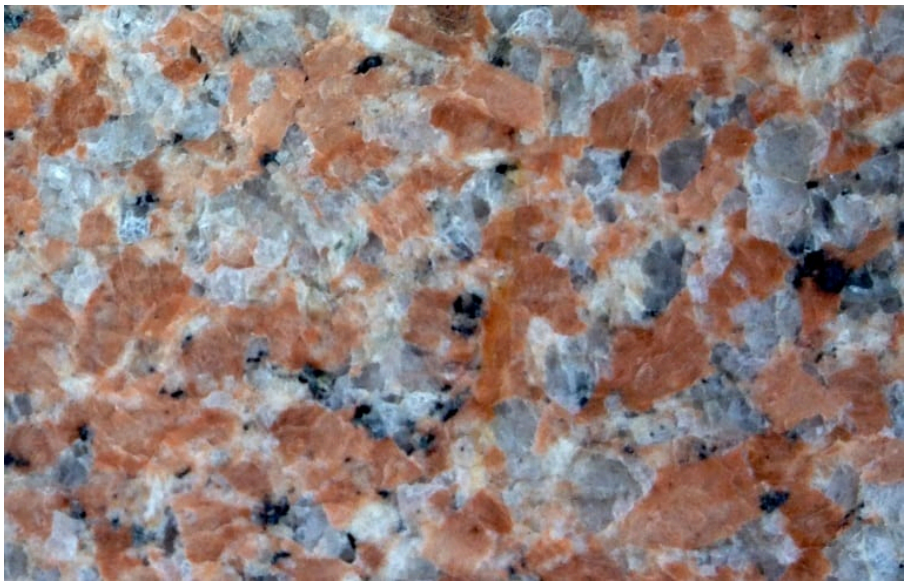


Figure 14.3.4.2 Crystals of potassium feldspar (pink) in a granitic rock are candidates for isotopic dating using the K-Ar method because they contained potassium and no argon when they formed.

Why can't we use isotopic dating techniques to accurately date sedimentary rocks?



Figure 14.3.4.3

An important assumption that we have to be able to make when using isotopic dating is that when the rock formed none of the daughter isotope was present (e.g., ^{40}Ar in the case of the K-Ar method). A clastic sedimentary rock is made up of older rock and mineral fragments, and when the rock forms it is almost certain that all of the fragments already have daughter isotopes in them. Furthermore, in almost all cases, the fragments have come from a range of source rocks that all formed at different times. If we dated a number of individual grains in the sedimentary rock, we would likely get a range of different dates, all older than the age of the rock. That could be useful information, but it would not provide an accurate date for the rock in question.

It might be possible to directly date some chemical sedimentary rocks isotopically, but there are no useful isotopes that can be used on old chemical sedimentary rocks. Radiocarbon dating can be used on sediments or sedimentary rocks that contain carbon, but it cannot be used on materials older than about 60 ka.

K-Ar is just one of many isotope-pairs that are useful for dating geological materials. Some of the other important pairs are listed in Table 8.2, along with the age ranges that they apply to and some comments on their applications. When radiometric techniques are applied to metamorphic rocks, the results normally tell us the date of metamorphism, not the date when the parent rock

formed.

Table 8.2 A few of the isotope systems that are widely used for dating geological materials

[Skip Table]

Isotope System	Half-Life	Useful Range	Comments
Potassium-argon	1.3 Ga	10 Ka to 4.57 Ga	Widely applicable because most rocks have some potassium
Uranium-lead	4.5 Ga	1 Ma to 4.57 Ga	The rock must have uranium-bearing minerals, but most have enough.
Rubidium-strontium	47 Ga	10 Ma to 4.57 Ga	Less precision than other methods at old dates
Carbon-nitrogen (a.k.a. radiocarbon dating)	5,730 years	100 to 60,000 years	Sample must contain wood, bone, or carbonate minerals; can be applied to young sediments

Exercise 8.3 Isotopic dating

Assume that a feldspar crystal from the granite shown in Figure 14.3.4.2 was analyzed for ^{40}K and ^{40}Ar. The proportion of ^{40}K remaining is 0.91. Using the decay curve shown on the graph below, estimate the age of the rock.

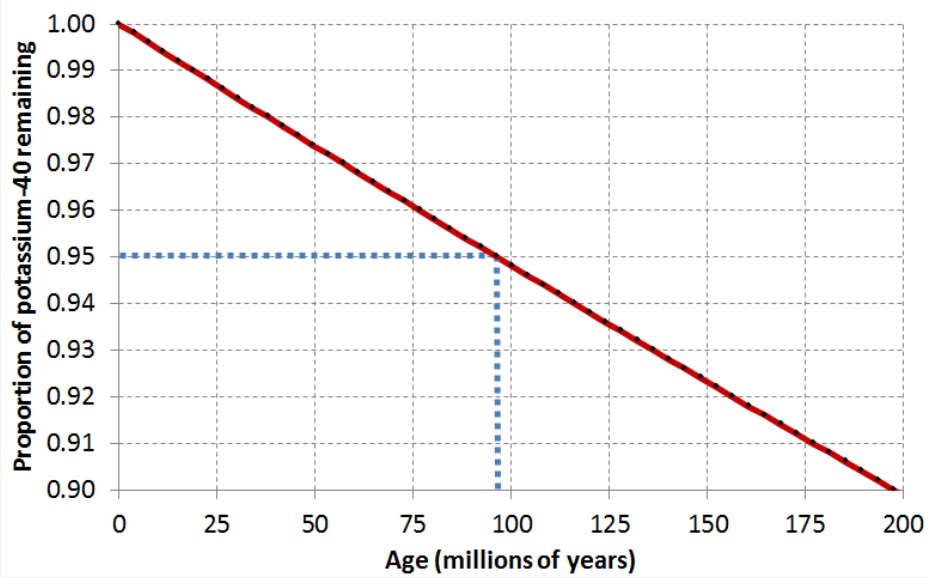


Figure 14.3.4.4

An example is provided (in blue) for a ^{40}K proportion of 0.95, which is equivalent to an age of approximately 96 Ma. This is determined by drawing a horizontal line from 0.95 to the decay curve line, and then a vertical line from there to the time axis.

See Appendix 3 for Exercise 8.3 answers.

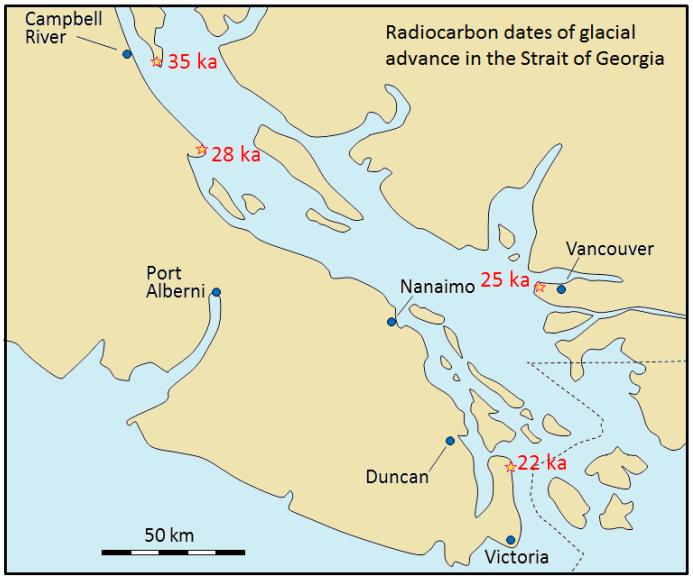


Figure 14.3.4.5 Radiocarbon dates on wood fragments in glacial sediments in the Strait of Georgia.

Radiocarbon dating (using ^{14}C) can be applied to many geological materials, including sediments and sedimentary rocks, but the materials in question must be younger than 60 ka. Fragments of wood incorporated into young sediments are good candidates for carbon dating, and this technique has been used widely in studies involving late Pleistocene glaciers and glacial sediments. An example is shown in Figure 14.3.4.5; radiocarbon dates from wood fragments in glacial sediments have been used to estimate the timing of the last glacial advance along the Strait of Georgia. It is evident that the ice-front of the major glacier that occupied the Strait of Georgia was near to Campbell River at around 35 ka, near to Nanaimo and Vancouver at about 25 ka, and had reached the Victoria area by around 22 ka.

Over the past decade there has been increasing use of U-Pb dating to study sedimentary rocks, not necessarily to find out the age of the rock, but to discover something about its history and origins. All clastic sedimentary rocks contain some tiny clasts of the silicate mineral zircon ($ZrSiO_4$), derived from the weathering of the sediment parent rocks. Zircon always has some uranium in it (but no lead) so it is a good candidate for U-Pb dating, and it isn't too difficult to separate the grains of zircon from the other grains in a sandstone. The procedure is to isolate a few hundred tiny zircons from a rock sample, and then carry out U-Pb dating on each one of them. An example of the types of results obtained are shown on Figure 14.3.4.6. All of the samples are from Nanaimo Gp. rocks on Vancouver Island and nearby Salt Spring Island.

The three samples from Vancouver Island have zircons aged around 90 Ma, 118 Ma and 150 Ma. The Salt Spring Island sample has some zircons aged around 150 Ma, but most are much older, at 200 Ma and 340 to 360 Ma. It is interpreted that the younger zircons (90 to 150 Ma) are mostly derived from granitic rocks in the Coast Range, while the older ones (>200 Ma) are from older rocks on Vancouver Island (Huang, 2018).

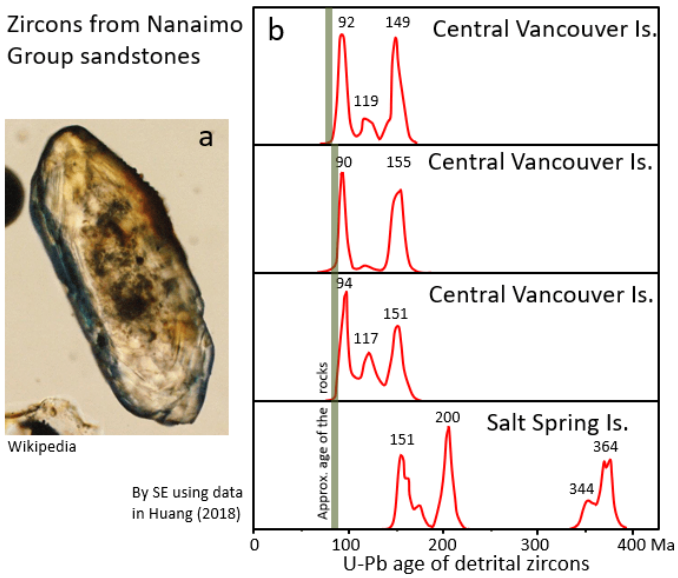


Figure 14.3.4.6 U-Pb dates for zircon samples from the Nanaimo Gp. (after Krause, 2018) a: a typical zircon clast (this one is about 1/4 mm long). b: plots of zircon ages for 4 sandstone samples.

Image Descriptions

Figure 14.3.4.1 image description: Decay of 40K over time.

Number of half-lives	Percent of 40K remaining	Percent of 40Ar
0	100	0
1	50	50
2	25	75
3	12.5	87.5
4	6.25	93.75
5	3.125	96.875
6	1.5625	98.4375
7	0.78125	99.21875

[Return to Figure 14.3.4.1]

Figure 14.3.4.4 image description: isotopic dating graph

Proportion of Potassium-40 remaining	Age (in millions of Years)
1.00	0
0.99	19
0.98	37
0.97	55
0.96	75
0.95	96
0.94	114
0.93	134

Proportion of Potassium-40 remaining	Age (in millions of Years)
0.92	156
0.91	175
0.90	194

[Return to Figure 14.3.4.4]

Media Attributions

- Figures 8.4.1, 8.4.2, 8.4.3, 8.4.4: © Steven Earle. CC BY.
- Figure 14.3.4.5: © Steven Earle. CC BY. From J. Clague, 1976, Quadra Sand and its relation to late Wisconsin glaciation of southeast British Columbia, Can. J. Earth Sciences, V. 13, p. 803-815.
- Figure 14.3.4.6 (left): "Zircon microscope" © Chd. CC BY-SA.
- Figure 14.3.4.6b: © Steven Earle. CC BY. From data in Huang, C, 2018, Refining the chronostratigraphy of the lower Nanaimo Group, Vancouver Island, Canada, using Detrital Zircon Geochronology, MSc thesis, Department of Earth Science, Simon Fraser University, 74 p.

This page titled 14.3.4: Isotopic Dating Methods is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Steven Earle (BCCampus) via source content that was edited to the style and standards of the LibreTexts platform.

- 8.4: Isotopic Dating Methods by Steven Earle is licensed CC BY 4.0. Original source: https://opentextbc.ca/physicalgeology2ed.

14.3.5: Understanding Geological Time

It's one thing to know the facts about geological time—how long it is, how we measure it, how we divide it up, and what we call the various periods and epochs—but it is quite another to really understand geological time. The problem is that our lives are short and our memories are even shorter. Our experiences span only a few decades, so we really don't have a way of knowing what 11,700 years means. What's more, it's hard for us to understand how 11,700 years differs from 65.5 million years, or even from 1.8 billion years. It's not that we can't comprehend what the numbers mean—we can all get that figured out with a bit of practice—but even if we do know the numerical meaning of 65.5 Ma, we can't really appreciate how long ago it was.

You may be wondering why it's so important to really "understand" geological time. There are some very good reasons. One is so that we can fully understand how geological processes that seem impossibly slow can produce anything of consequence. For example, we are familiar with the concept of driving from one major city to another: a journey of several hours at around 100 kilometers per hour. Continents move toward each other at rates of a fraction of a millimetre per day, or something in the order of 0.000000001 kilometers per hour, and yet, at this impossibly slow rate (try walking at that speed!), they can move thousands of kilometers. Sediments typically accumulate at even slower rates—less than a millimetre per year—but still they are thick enough to be thrust up into monumental mountains and carved into breathtaking canyons.

Another reason is that for our survival on this planet, we need to understand issues like extinction of endangered species and **anthropogenic** (human-caused) climate change. Some people, who don't understand geological time, are quick to say that the climate has changed in the past, and that what is happening now is no different. And it certainly has changed in the past—many times. For example, from the Eocene (50 Ma) to the present day, Earth's climate cooled by about 12°C. That's a huge change that ranks up there with many of the important climate changes of the distant past, and yet the rate of change over that time was only 0.000024°C/century. Anthropogenic climate change has been 1.1°C over the past century,[1]; that is 45,800 times faster than the rate of natural climate change since the Eocene!

One way to wrap your mind around geological time is to put it into the perspective of single year, because we all know how long it is from one birthday to the next. At that rate, each hour of the year is equivalent to approximately 500,000 years, and each day is equivalent to 12.5 million years.

If all of geological time is compressed down to a single year, Earth formed on January 1, and the first life forms evolved in late March (roughly 3,500 Ma). The first large life forms appeared on November 13 (roughly 600 Ma), plants appeared on land around November 24, and amphibians on December 3. Reptiles evolved from amphibians during the first week of December and dinosaurs and early mammals evolved from reptiles by December 13, but the dinosaurs, which survived for 160 million years, were gone by Boxing Day (December 26). The Pleistocene Glaciation got started at around 6:30 p.m. on New Year's Eve, and the last glacial ice left southern Canada by 11:59 p.m.

It's worth repeating: on this time scale, the earliest ancestors of the animals and plants with which we are familiar did not appear on Earth until mid-November, the dinosaurs disappeared after Christmas, and most of Canada was periodically locked in ice from 6:30 to 11:59 p.m. on New Year's Eve. As for people, the first to inhabit B.C. got here about one minute before midnight, and the first Europeans arrived about two seconds before midnight.

It is common for the popular press to refer to distant past events as being "prehistoric." For example, dinosaurs are reported as being "prehistoric creatures," even by the esteemed National Geographic Society.[2] The written records of our history date back to about 6,000 years ago, so anything prior to that can be considered "prehistoric." But to call the dinosaurs prehistoric is equivalent to—and about as useful as—saying that Singapore is beyond the city limits of Kamloops! If we are going to become literate about geological time, we have to do better than calling dinosaurs, or early horses (54 Ma), or even early humans (2.8 Ma), "prehistoric."

Exercise 8.5 What happened on your birthday?

Using the "all of geological time compressed to one year" concept, determine the geological date that is equivalent to your birthday. First go to Day Number of the Year Calculator to find out which day of the year your birth date is. Then divide that number by 365, and multiply that number by 4,570 to determine the time (in millions since the *beginning* of geological time). Finally subtract that number from 4,570 to determine the date back from the present.

For example, April Fool's Day (April 1) is day 91 of the year: 91/365 = 0.2493. 0.2493 x 4,570 = 1,139 million years from the start of time, and 4,570 – 1,193 = 3,377 Ma is the geological date.

Finally, go to the Foundation for Global Community's "Walk through Time" website to find out what was happening on your day. The nearest date to 3,377 Ma is 3,400 Ma. Bacteria ruled the world at 3,400 Ma, and there's a discussion about their lifestyles.

See Appendix 3 for Exercise 8.5 answers.

---

1. Climate change data from NASA Goddard Institute for Space Studies: http://data.giss.nasa.gov/gistemp/tabledata_v3/GLB.Ts.txt↵
2. http://science.nationalgeographic.com/science/prehistoric-world/ ↵

This page titled 14.3.5: Understanding Geological Time is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Steven Earle (BCCampus) via source content that was edited to the style and standards of the LibreTexts platform.

- 8.6: Understanding Geological Time by Steven Earle is licensed CC BY 4.0. Original source: https://opentextbc.ca/physicalgeology2ed.

15: THE DYNAMIC EARTH

15.1: Understanding Earth Through Seismology

15.2: Plate Tectonics

15.2.1: Alfred Wegener- The Father of Plate Tectonics

15.2.2: Plate, Plate Motions, and Plate Boundary Processes

15.2.3: Mechanisms for Plate Motion

15.3: Earthquakes

15.3.1: What is an Earthquake?

15.3.2: Earthquakes and Plate Tectonics

15.3.3: Measuring Earthquakes

15.3.4: The Impacts of Earthquakes

15.4: Volcanoes

15.4.1: Plate Tectonics and Volcanism

15.4.2: Magma Composition and Eruption Style

15.4.3: Types of Volcanoes

15.4.4: Volcanic Hazards

Chapter 15: The Dynamic Earth — Introduction

Earth may appear solid and unchanging on the timescale of our everyday lives, but it is an active and dynamic planet. Processes occurring deep within Earth continually reshape its surface, moving continents, building mountains, producing earthquakes, and creating volcanoes. Many of these changes occur too slowly for us to observe directly, while others can transform landscapes in a matter of minutes.

Our understanding of this dynamic Earth begins beneath the surface. Because we cannot directly explore most of Earth's interior, scientists use seismic waves produced by earthquakes to investigate its internal structure. The behavior of these waves provides evidence about the different layers within Earth and helps us understand the conditions and processes occurring far below the crust.

That internal activity is closely connected to **plate tectonics**, the unifying theory that explains much of Earth's large-scale geological activity. Earth's outermost layers are divided into plates that move slowly relative to one another. Their interactions help explain the locations of mountain ranges, ocean basins, earthquakes, and volcanoes. The theory developed from many different lines of evidence and provides a framework for understanding Earth as an interconnected system rather than a collection of separate geological features.

In this chapter, we will first examine how seismology allows us to investigate Earth's interior and then explore the evidence and mechanisms behind plate tectonics. We will apply those ideas to two of the most visible consequences of a dynamic planet: **earthquakes and volcanoes**. Together, these topics show how processes operating within Earth continue to shape the planet beneath our feet.

This page titled 15: The Dynamic Earth is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Robert Wagner.

15.1: Understanding Earth Through Seismology

Seismology is the study of vibrations within Earth. These vibrations are caused by various events: earthquakes, extraterrestrial impacts, explosions, storm waves hitting the shore, and tidal effects. Of course, seismic techniques have been most widely applied to the detection and study of earthquakes, but there are many other applications, and arguably seismic waves provide the most important information that we have concerning Earth's interior. Before going any deeper into Earth, however, we need to take a look at the properties of seismic waves. The types of waves that are useful for understanding Earth's interior are called **body waves**, meaning that, unlike the surface waves on the ocean, they are transmitted through Earth materials.

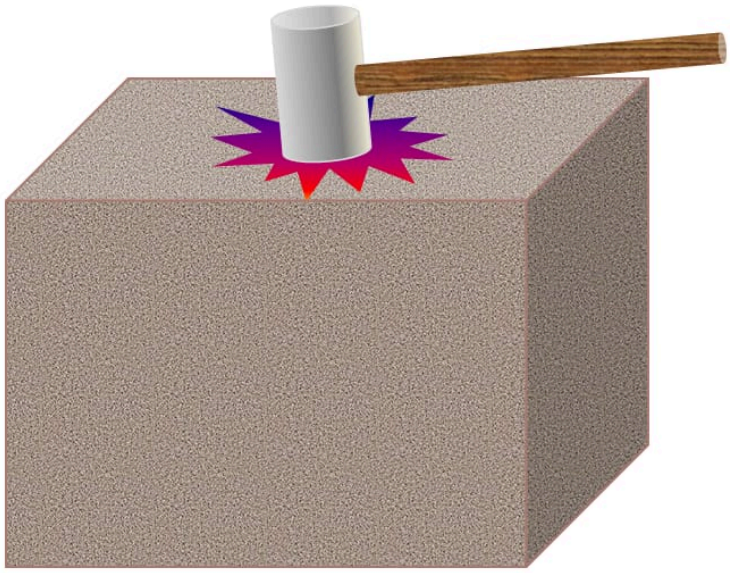


Figure 15.1.1 Hitting a large block of rock with a heavy hammer will create seismic waves within the rock. Please don't try this at home!

Imagine hitting a large block of strong rock (e.g., granite) with a heavy sledgehammer (Figure 15.1.1). At the point where the hammer strikes it, a small part of the rock will be compressed by a fraction of a millimetre. That compression will transfer to the neighbouring part of the rock, and so on through to the far side of the rock—all in a fraction of a second. This is known as a compression wave, and it can be illustrated by holding a loose spring (like a Slinky) that is attached to something (or someone) at the other end. If you give it a sharp push so the coils are compressed, the compression propagates (travels) along the length of the spring and back (Figure 15.1.2). You can think of a compression wave as a "push" wave—it's called a **P wave** (although the "P" stands for "primary" because P waves arrive first at seismic stations).

When we hit a rock with a hammer, we also create a different type of body wave, one that is characterized by back-and-forth vibrations (as opposed to compressions). This is known as a shear wave (**S wave**, where the "S" stands for "secondary"), and an analogy would be what happens when you flick a length of rope with an up-and-down motion. As shown in Figure 15.1.2, a wave will form in the rope, which will travel to the end of the rope and back.

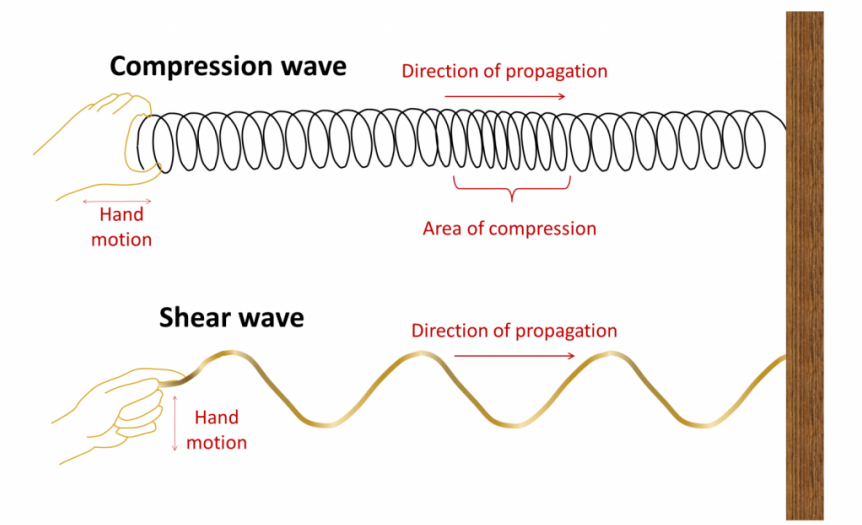


Figure 15.1.2 A compression wave can be illustrated by a spring (like a Slinky) that is given a sharp push at one end. A shear wave can be illustrated by a rope that is given a quick flick.

Compression waves and shear waves travel very quickly through geological materials. As shown in Figure 15.1.3, typical P wave velocities are between 0.5 kilometers per second (km/s) and 2.5 km/s in unconsolidated sediments, and between 3.0 km/s and 6.5 km/s in solid crustal rocks. Of the common rocks of the crust, velocities are greatest in basalt and granite. S waves are slower than P waves, with velocities between 0.1 km/s and 0.8 km/s in soft sediments, and between 1.5 km/s and 3.8 km/s in solid rocks.

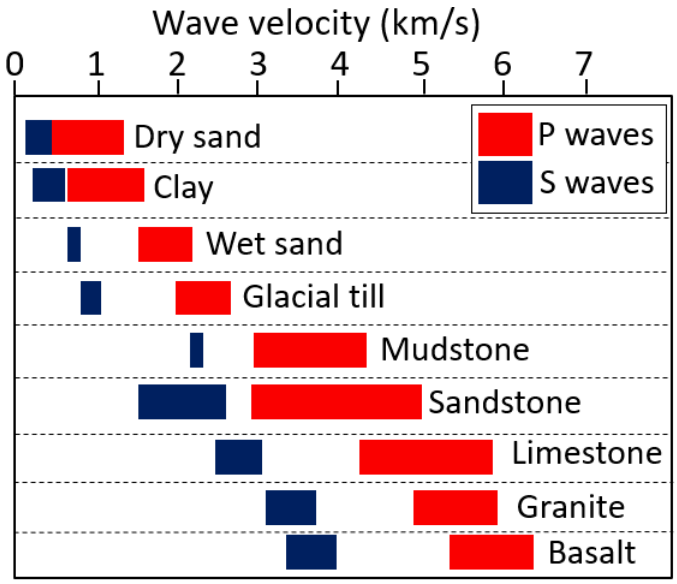


Figure 15.1.3 Typical velocities of P-waves (red) and S-waves (blue) in sediments and in solid crustal rocks.

Exercise 9.1 How soon will seismic waves get here?

Imagine that a strong earthquake takes place on Vancouver Island within Strathcona Park (west of Courtenay). Assuming that the crustal average P wave velocity is 5 km per second, how long will it take (in seconds) for the first seismic waves (P waves) to reach you in the following places (distances from the epicentre are shown)?

1. Nanaimo (120 km away)
2. Surrey (200 km away)
3. Kamloops (390 km away)

See Appendix 3 for Exercise 9.1 answers.

Mantle rock is generally denser and stronger than crustal rock and both P- and S-waves travel faster through the mantle than they do through the crust. Moreover, seismic-wave velocities are related to how tightly compressed a rock is, and the level of compression increases dramatically with depth. Finally, seismic waves are affected by the phase state of rock. They are slowed if there is any degree of melting in the rock. If the material is completely liquid, P waves are slowed dramatically and S waves are stopped altogether.

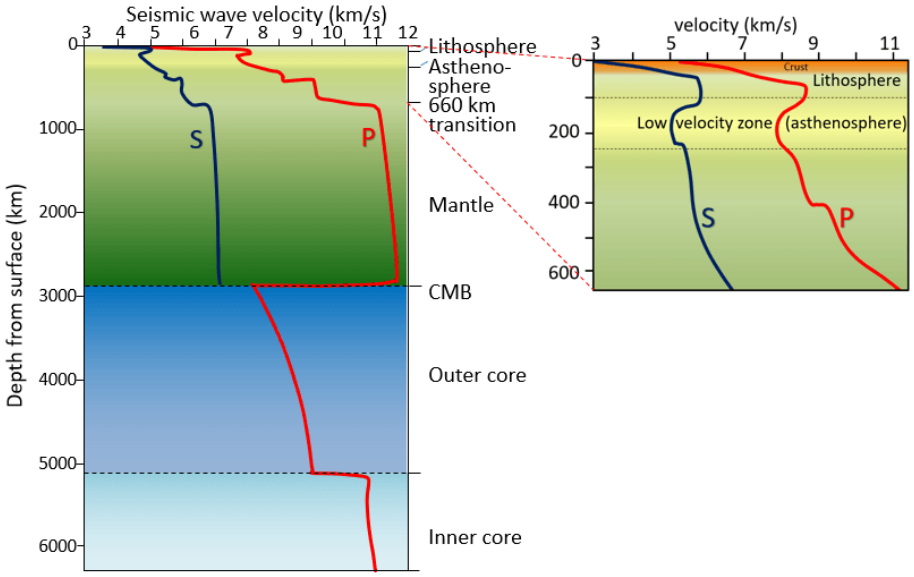


Figure 15.1.4 P wave (red) and S wave (blue) velocity variations with depth in Earth. The diagram on the right shows an expanded view of the upper 660 kilometers of the curves in the diagram on the left.

As shown on the right-hand part of Figure 15.1.4, the upper approximately 100 km of the Earth is known as the lithosphere. This includes the rigid upper part of the mantle (or lithospheric mantle) and the crust. The next 150 km is the asthenosphere or low velocity zone (because seismic waves are slowed as they pass through that material). As we'll see below, that part of the mantle is close to it's melting point and in some regions may be partially molten.

Accurate seismometers have been used for earthquake studies since the late 1800s, and systematic use of seismic data to understand Earth's interior started in the early 1900s. The rate of change of seismic waves with depth in Earth (as shown in Figure 15.1.4) has been determined over the past several decades by analyzing seismic signals from large earthquakes at seismic stations around the world. Small differences in arrival time of signals at different locations have been interpreted to show that:

- Velocities are greater in mantle rock than in the crust.
- Velocities generally increase with pressure, and therefore with depth.
- Velocities slow in the area between a 100 and 250 kilometre depth (called the "low-velocity zone"; equivalent to the asthenosphere).
- Velocities increase dramatically at 660 kilometre depth (because of a mineralogical transition).
- Velocities slow in the region just above the core-mantle boundary (the D" (d-double-prime) layer or "ultra-low-velocity zone").
- S waves do not pass through the outer part of the core.
- P wave velocities increase dramatically at the boundary between the liquid outer core and the solid inner core.

One of the first discoveries about Earth's interior made through seismology was in 1909 when Croatian seismologist Andrija Mohorovičić (pronounced *Moho-ro-vi-chich*) realized that at certain distances from an earthquake, two separate sets of seismic waves arrived at a seismic station within a few seconds of each other. He reasoned that the waves that went down into the mantle, traveled through the mantle, and then were bent upward back into the crust, reached the seismic station first because although they had farther to go, they traveled faster through mantle rock (as shown in Figure 15.1.5). The boundary between the crust and the mantle is known as the **Mohorovičić discontinuity** (or Moho). Its depth is between 30 and 40 kilometers beneath most of the continental crust, and between 5 and 10 kilometers beneath the oceanic crust.

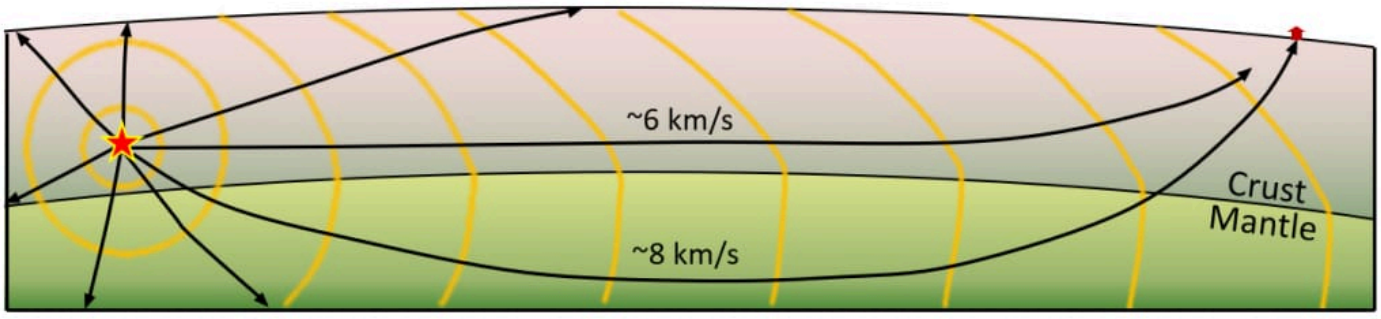


Figure 15.1.5 Depiction of seismic waves emanating from an earthquake (red star). Some waves travel through the crust to the seismic station (at about 6 km/s), while others go down into the mantle (where they travel at around 8 km/s) and are bent upward toward the surface, reaching the station before the ones that traveled only through the crust.

Our current understanding of the patterns of seismic wave transmission through Earth is summarized in Figure 15.1.6. Because of the gradual increase in density (and therefore rock strength) with depth, all waves are refracted (toward the lower density material) as they travel through homogenous parts of Earth and thus tend to curve outward toward the surface. Waves are also refracted at boundaries within Earth, such as at the Moho, at the core-mantle boundary (CMB), and at the outer-core/inner-core boundary.

S waves do not travel through liquids—they are stopped at the CMB—and there is an S wave shadow on the side of Earth opposite a seismic source. The angular distance from the seismic source to the shadow zone is 103° on either side, so the total angular distance of the shadow zone is 154°. We can use this information to infer the depth to the CMB.

P waves do travel through liquids, so they can make it through the liquid part of the core. Because of the refraction that takes place at the CMB, waves that travel through the core are bent away from the surface, and this creates a P wave shadow zone on either side, from 103° to 150°. This information can be used to discover the differences between the inner and outer parts of the core.

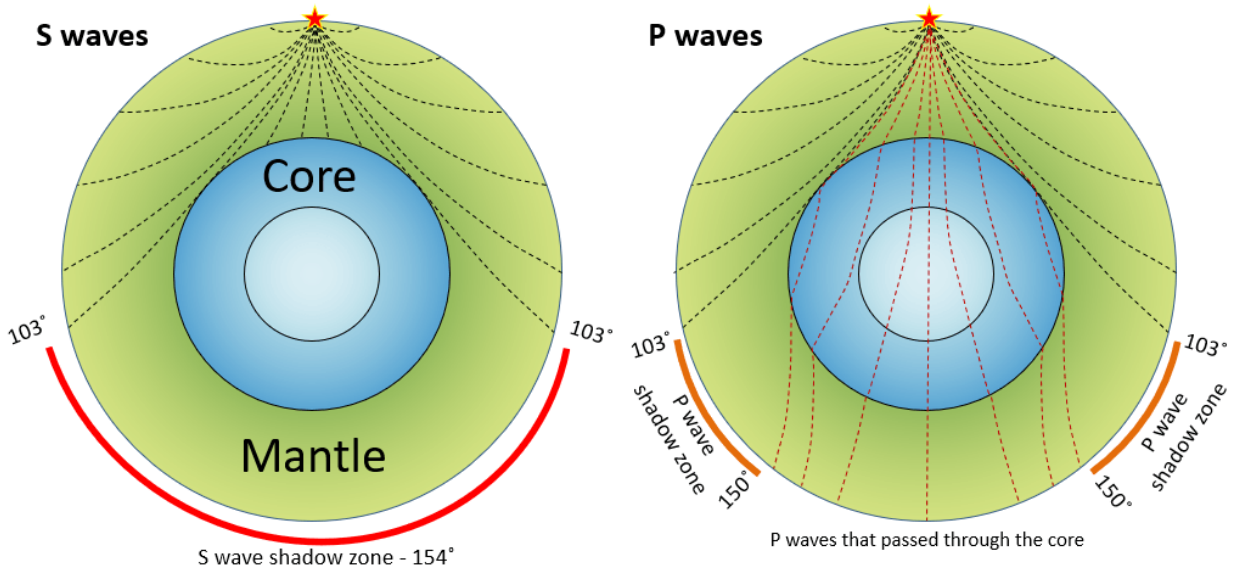


Figure 15.1.6 Patterns of seismic wave propagation through Earth's mantle and core. S waves do not travel through the liquid outer core, so they leave a shadow on Earth's far side where they cannot get to. P waves do travel through the core, but because the waves that enter the core are refracted, there are also P wave shadow zones.

Exercise 9.2 Liquid Cores in Other Planets

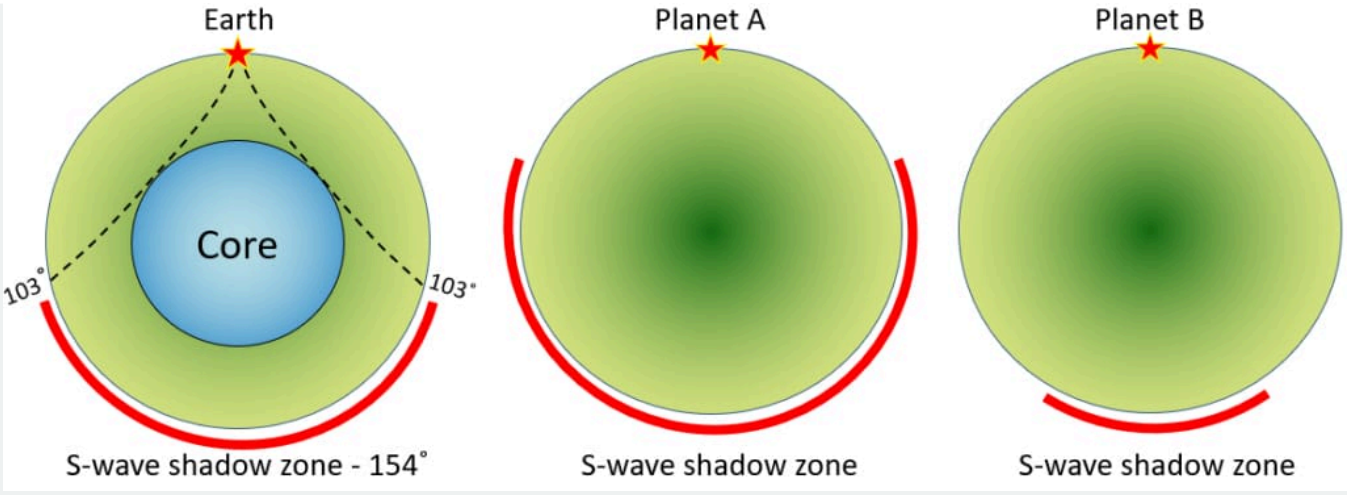


Figure 15.1.7

We know that other planets must have (or at least did have) liquid cores like ours, and we could use seismic data to find out how big they are. The S wave shadow zones on planets A and B are shown. Using the same method used for Earth (on the left), sketch in the outlines of the cores for these two other planets.

See Appendix 3 for Exercise 9.2 answers.

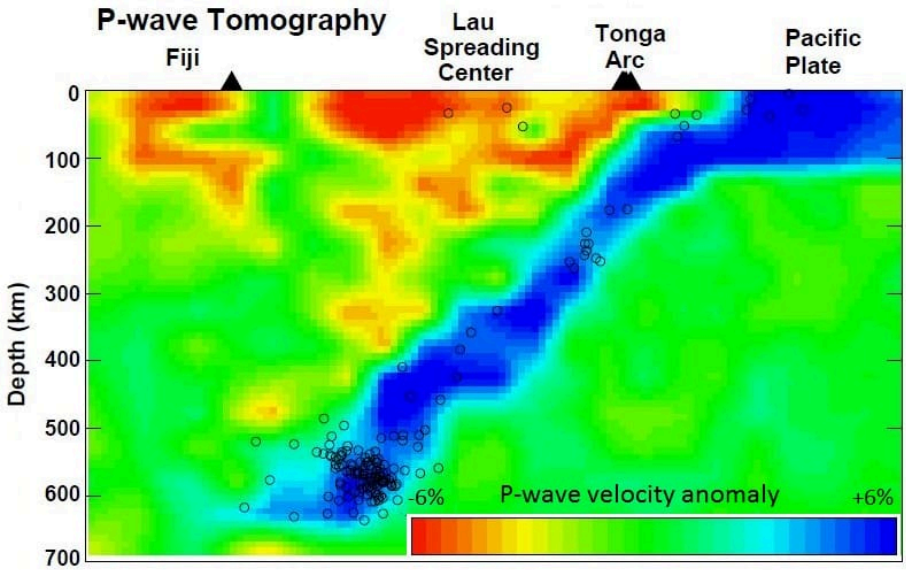


Figure 15.1.8 P-wave tomographic profile of area in the southern Pacific Ocean from southeast of Tonga to Fiji. Blue represents rock that has relatively high seismic velocities, while yellow and red represent rock with low velocities. Open circles are earthquakes used in the study.

Using data from many seismometers and hundreds of earthquakes, it is possible to create a two- or three-dimensional image of the seismic properties of part of the mantle. This technique is known as seismic tomography, and an example of the result is shown in Figure 15.1.8.

The Pacific Plate subducts beneath Tonga and appears in Figure 15.1.8 as a 100 kilometre thick slab of cold (blue-colored) oceanic crust that has pushed down into the surrounding hot mantle. The cold rock is more rigid than the surrounding hot mantle rock, so it is characterized by slightly faster seismic velocities. There is volcanism in the Lau spreading centre and also in the Fiji area, and the warm rock in these areas has slower seismic velocities (yellow and red colors).

Image descriptions

Figure 15.1.3 image description: Wave velocity in different materials in kilometers per second.

Material	S Wave (kilometers per second)	P Wave (kilometers per second)
Dry sand	0.1 to 0.4	0.4 to 1.3
Clay	0.2 to 0.6	0.6 to 1.6
Wet sand	0.7 to 0.8	1.5 to 2.2
Till	0.8 to 1.0	1.9 to 2.6
Mudstone	2.1 to 2.3	3.0 to 4.3
Sandstone	1.4 to 2.5	3.0 to 5.0
Limestone	2.4 to 3.1	4.2 to 5.8
Granite	3.0 to 3.7	4.9 to 5.9
Basalt	3.3 to 4.0	5.2 to 6.2

[Return to Figure 15.1.3]

Figure 15.1.4 image description: P-wave and S-wave velocity variations with depth in Earth.

Layer	Depth from surface (km)	S-Wave velocity (kilometers per second)	P-Wave velocity (kilometers per second)
Crust	0 to 30	3.0 to 4.6	5.3 to 7.0
Lithosphere	30 to 100	4.6 to 5.8	7.0 to 8.7
Asthenosphere	100 to 250	5.0 to 5.9	7.8 to 8.5
Mantle	250 to 2890	5. to 7.0	8.2 to 12.6
Outer core	2890 to 5100	0	8.0 to 10.1
Inner core	5100 to 6370	0	11.8 to 12.0

[Return to Figure 15.1.4]

Media Attributions

- Figures 9.1.1, 9.1.2, 9.1.4, 9.1.5, 9.1.6, 9.1.7: © Steven Earle. CC BY.
- Figure 15.1.3: “P Wave Velocity, m/s” and “Shear Wave Velocity, m/s” by the US Environment Protection Agency. Edited by Steven Earle. Public domain.
- Figure 15.1.8: “P-wave Tomography” by D. Zhao, Y. Xu, D.A. Wiens, L. Dorman, J. Hildebrand, and S. Webb. (Science, p. 278, 254-257, 1997). Used with permission.

This page titled 15.1: Understanding Earth Through Seismology is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Steven Earle (BCCampus) via source content that was edited to the style and standards of the LibreTexts platform.

- 9.1: Understanding Earth Through Seismology by Steven Earle is licensed CC BY 4.0. Original source: https://opentextbc.ca/physicalgeology2ed.

15.2: Plate Tectonics

This section introduces plate tectonics, the theory that Earth's lithosphere is divided into moving plates. It traces the development of the idea from Alfred Wegener's proposal of continental drift to the modern theory of plate tectonics, and examines the processes occurring at divergent, convergent, and transform plate boundaries. It also considers the forces and mechanisms responsible for plate motion.

This page titled 15.2: Plate Tectonics is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Robert Wagner.

15.2.1: Alfred Wegener- The Father of Plate Tectonics



Figure 15.2.1.1 Alfred Wegener a few years before his death in 1930.

Alfred Wegener (1880-1930) (Figure 15.2.1.1) earned a PhD in astronomy at the University of Berlin in 1904, but he had always been interested in geophysics and meteorology and spent most of his academic career working in meteorology. In 1911 he happened on a scientific publication that included a description of the existence of matching Permian-aged terrestrial fossils in various parts of South America, Africa, India, Antarctica, and Australia (Figure 15.2.1.2).

Wegener concluded that this distribution of terrestrial organisms could only exist if these continents were joined together during the Permian, and he coined the term **Pangea** ("all land") for the supercontinent that he thought included all of the present-day continents.

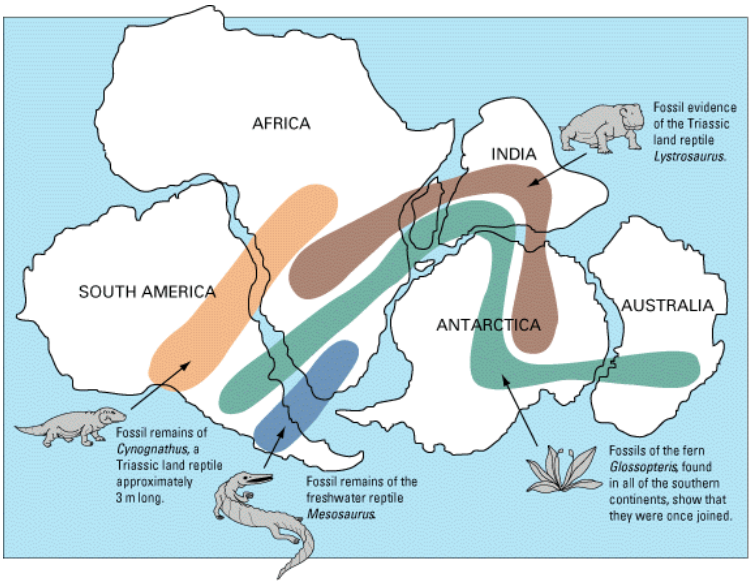


Figure 15.2.1.2 The distribution of several Permian terrestrial fossils that are present in various parts of continents that are now separated by oceans.

Wegener pursued his theory with determination—combing the libraries, consulting with colleagues, and making observations—looking for evidence to support it. He relied heavily on matching geological patterns across oceans, such as sedimentary strata in South America matching those in Africa (Figure 15.2.1.3), North American coalfields matching those in

Europe, and the mountains of Atlantic Canada matching those of northern Britain—both in morphology and rock type.

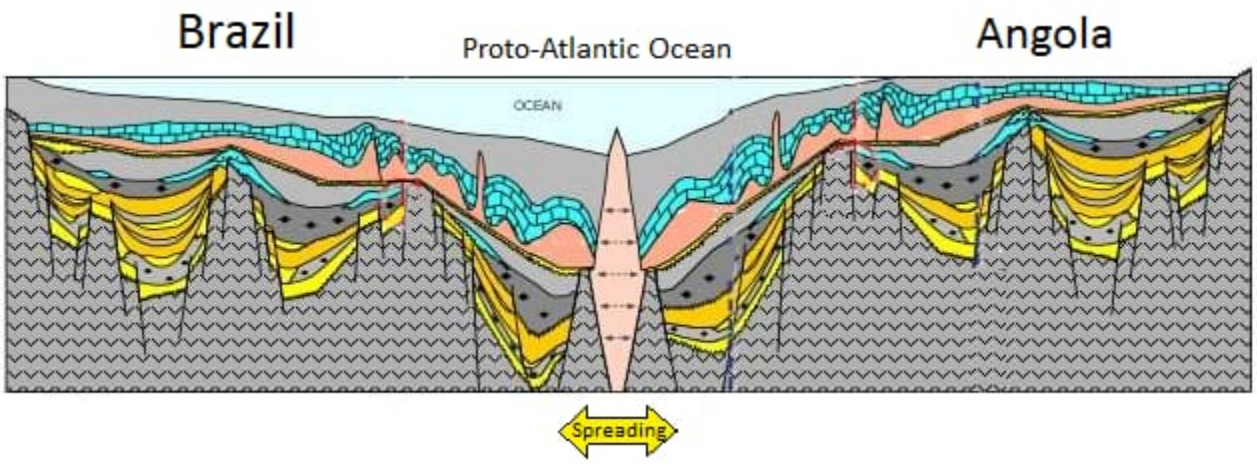


Figure 15.2.1.3 A cross-section showing the geological similarities between parts of Brazil (South America) on the left and Angola (Africa) on the right. The pink layer is a salt deposit, which is now known to be common in areas of continental rifting.

Wegener referred to the evidence for the Carboniferous and Permian (~300 Ma) Karoo Glaciation in South America, Africa, India, Antarctica, and Australia (Figure 15.2.1.4). He argued that this could only have happened if these continents were once all connected as a single supercontinent. He also cited evidence (based on his own astronomical observations) that showed that the continents were moving with respect to each other, and determined a separation rate between Greenland and Scandinavia of 11 meters per year, although he admitted that the measurements were not accurate. In fact they weren't even close—the separation rate is actually about 2.5 centimeters per year!

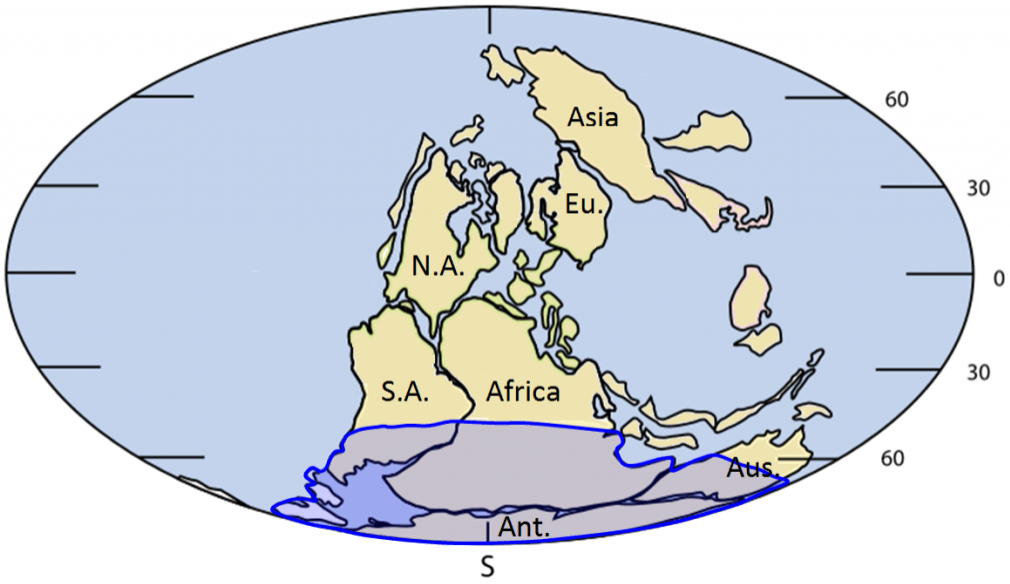


Figure 15.2.1.4 The distribution of the Carboniferous and Permian Karoo Glaciation (outlined in blue).

Wegener first published his ideas in 1912 in a short book called *Die Entstehung der Kontinente* (*The Origin of Continents*), and then in 1915 in *Die Entstehung der Kontinente und Ozeane* (*The Origin of Continentsand Oceans*). He revised this book several times up to 1929. It was translated into French, English, Spanish, and Russian in 1924.

In fact the continental fits were not perfect and the geological matchups were not always consistent, but the most serious problem of all was that Wegener could not conceive of a credible mechanism for moving the continents around. It was understood by this time that the continents were primarily composed of **sialic** material (**SIAL**: silicon and aluminum dominated, similar to "felsic"), and that the ocean floors were primarily **simatic** (**SIMA**: silicon and magnesium dominated, similar to "mafic"). Wegener proposed that the continents were like icebergs floating on the heavier SIMA crust, but the only forces that he could invoke to propel continents around were *poleflucht,* the effect of Earth's rotation pushing objects toward the equator, and the lunar and

solar tidal forces, which tend to push objects toward the west. It was quickly shown that these forces were far too weak to move continents, and without any reasonable mechanism to make it work, Wegener's theory was quickly dismissed by most geologists of the day.

Alfred Wegener died in Greenland in 1930 while carrying out studies related to glaciation and climate. At the time of his death, his ideas were tentatively accepted by only a small minority of geologists, and soundly rejected by most. However, within a few decades that was all to change. For more about his extremely important contributions to Earth science, visit the NASA website to see a collection of articles on Alfred Wegener.

Image Descriptions

Figure 15.2.1.2 **image description:** Fossils found across different continents suggest that these continents were once joined as a super-continent. Fossil remains of *Cynognathus* (a terrestrial reptile) and *Mesosaurus* (a freshwater reptile) have been found in South America and Africa. Fossil evidence of the *Lystrosaurus*, a land reptile from the Triassic period, has been found in India, Africa, and Antarctica. Fossils of the fern *Glossopteris* have been found in Australia, Antarctica, India, Africa, and South America. When you position these continents so they fit together, the areas where these fossils were found line up. [Return to Figure 15.2.1.2]

Media Attributions

- Figure 15.2.1.1: "Alfred Wegener ca.1924-30." Public domain.
- Figure 15.2.1.2: "Snider-Pellegrini Wegener fossil map" by Osvaldocangaspadilla. Public domain.
- Figure 15.2.1.3: © Steven Earle. CC BY. Based on "Angola -Brazil sub-sea geology" by Cobalt International Energy can be found at U.S. Energy Information Administration: Country Analysis Brief: Angola (May 2016) [PDF].
- Figure 15.2.1.4: "Karoo Glaciation" © GeoPotinga. Adapted by Steven Earle. CC BY-SA.

This page titled 15.2.1: Alfred Wegener- The Father of Plate Tectonics is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Steven Earle (BCCampus) via source content that was edited to the style and standards of the LibreTexts platform.

- 10.1: Alfred Wegener- The Father of Plate Tectonics by Steven Earle is licensed CC BY 4.0. Original source: https://opentextbc.ca/physicalgeology2ed.

15.2.2: Plate, Plate Motions, and Plate Boundary Processes

Continental drift and sea-floor spreading became widely accepted around 1965 as more and more geologists started thinking in these terms. By the end of 1967 the Earth's surface had been mapped into a series of plates (Figure 15.2.2.1). The major plates are Eurasia, Pacific, India, Australia, North America, South America, Africa, and Antarctic. There are also numerous small plates (e.g., Juan de Fuca, Cocos, Nazca, Scotia, Philippine, Caribbean), and many very small plates or sub-plates. For example the Juan de Fuca Plate is actually three separate plates (Gorda, Juan de Fuca, and Explorer) that all move in the same general direction but at slightly different rates.

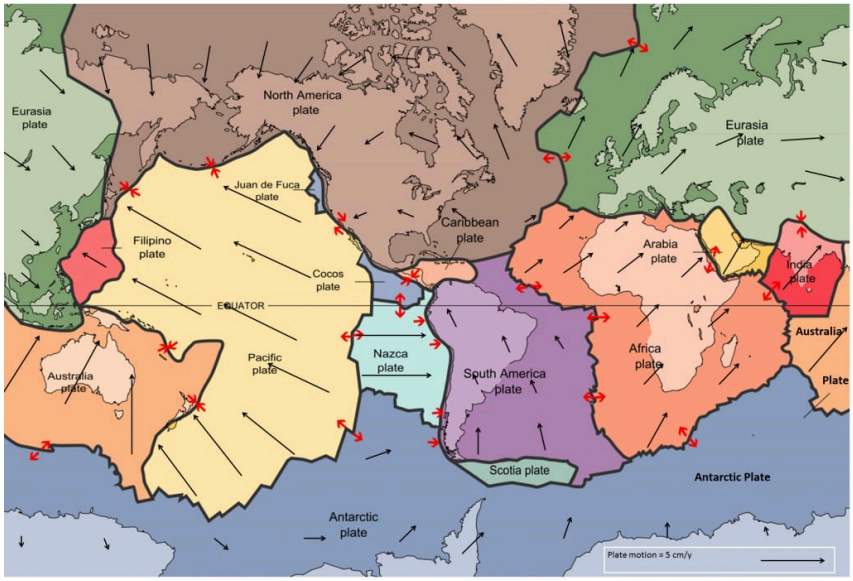


Figure 15.2.2.1 A map showing 15 of the Earth's tectonic plates and the approximate rates and directions of plate motions.

Rates of motions of the major plates range from less than 1 cm/y to over 10 cm/y. The Pacific Plate is the fastest, followed by the Australian and Nazca Plates. The North American Plate is one of the slowest, averaging around 1 cm/y in the south up to almost 4 cm/y in the north.

Plates move as rigid bodies, so it may seem surprising that the North American Plate can be moving at different rates in different places. The explanation is that plates move in a rotational manner. The North American Plate, for example, rotates counter-clockwise; the Eurasian Plate rotates clockwise.

Boundaries between the plates are of three types: **divergent** (i.e., moving apart), **convergent** (i.e., moving together), and **transform** (moving side by side). Before we talk about processes at plate boundaries, it's important to point out that there are never gaps between plates. The plates are made up of crust and the lithospheric part of the mantle (Figure 15.2.2.2), and even though they are moving all the time, and in different directions, there is never a significant amount of space between them. Plates are thought to move along the lithosphere-asthenosphere boundary, as the asthenosphere is the zone of partial melting. It is assumed that the relative lack of strength of the partial melting zone facilitates the sliding of the lithospheric plates.

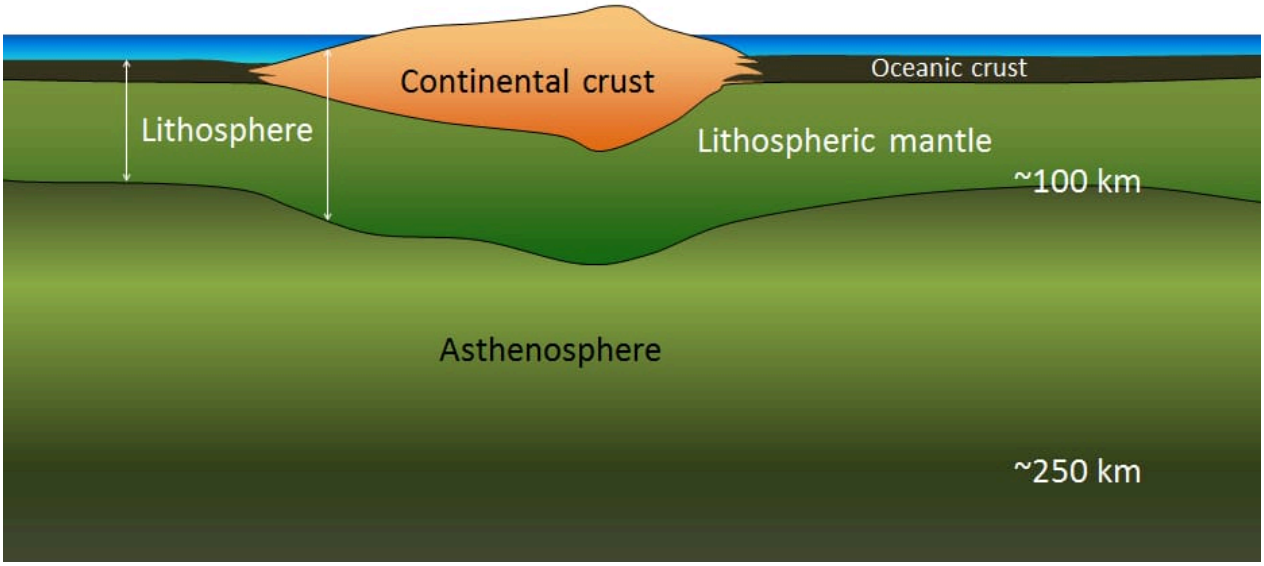


Figure 15.2.2.2 The crust and upper mantle. Tectonic plates consist of lithosphere, which includes the crust and the lithospheric (rigid) part of the mantle.

At spreading centres, the lithospheric mantle may be very thin because the upward convective motion of hot mantle material generates temperatures that are too high for the existence of a significant thickness of rigid lithosphere (Figure 15.2.2.8). The fact that the plates include both crustal material and lithospheric mantle material makes it possible for a single plate to be made up of both oceanic and continental crust. For example, the North American Plate includes most of North America, plus half of the northern Atlantic Ocean. Similarly the South American Plate extends across the western part of the southern Atlantic Ocean, while the European and African plates each include part of the eastern Atlantic Ocean. The Pacific Plate is almost entirely oceanic, but it does include the part of California west of the San Andreas Fault.

Divergent Boundaries

Divergent boundaries are spreading boundaries, where new oceanic crust is created from magma derived from partial melting of the mantle caused by decompression as hot mantle rock from depth is moved toward the surface (Figure 15.2.2.3). The triangular zone of partial melting near the ridge crest is approximately 60 km thick and the proportion of magma is about 10% of the rock volume, thus producing crust that is about 6 km thick. Most divergent boundaries are located at the oceanic ridges (although some are on land), and the crustal material created at a spreading boundary is always oceanic in character; in other words, it is mafic igneous rock (e.g., basalt or gabbro, rich in ferromagnesian minerals). Spreading rates vary considerably, from 2 cm/y to 6 cm/y in the Atlantic, to between 12 cm/y and 20 cm/y in the Pacific. (Note that spreading rates are typically double the velocities of the two plates moving away from a ridge.)

Some of the processes taking place in this setting include:

- Magma from the mantle pushing up to fill the voids left by divergence of the two plates
- **Pillow lavas** forming where magma is pushed out into seawater (Figure 15.2.2.4)
- Vertical **sheeted dykes** intruding into cracks resulting from the spreading
- Magma cooling more slowly in the lower part of the new crust and forming gabbro bodies

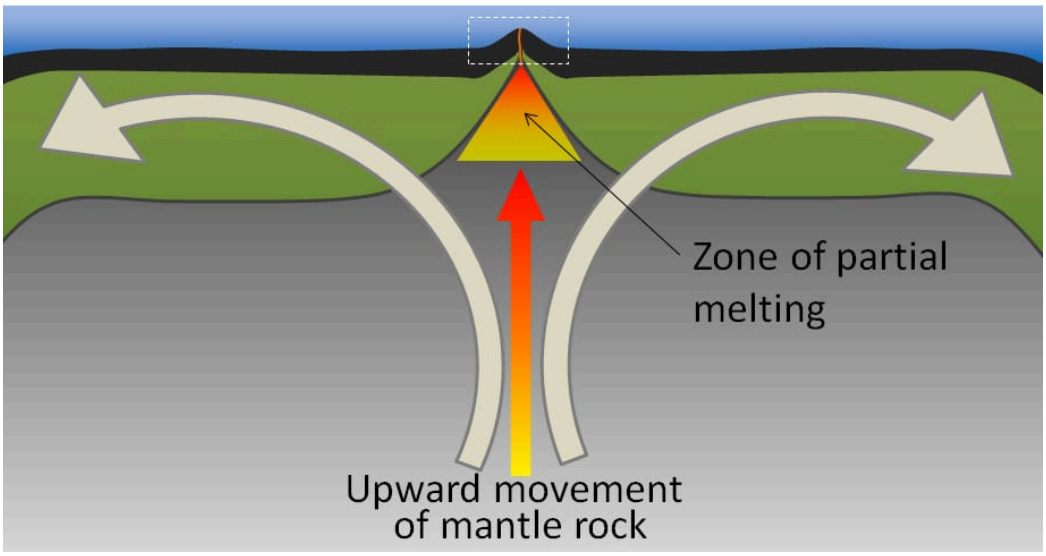


Figure 15.2.2.3 The general processes that take place at a divergent boundary. The area within the dashed white rectangle is shown in Figure 15.2.2.4.

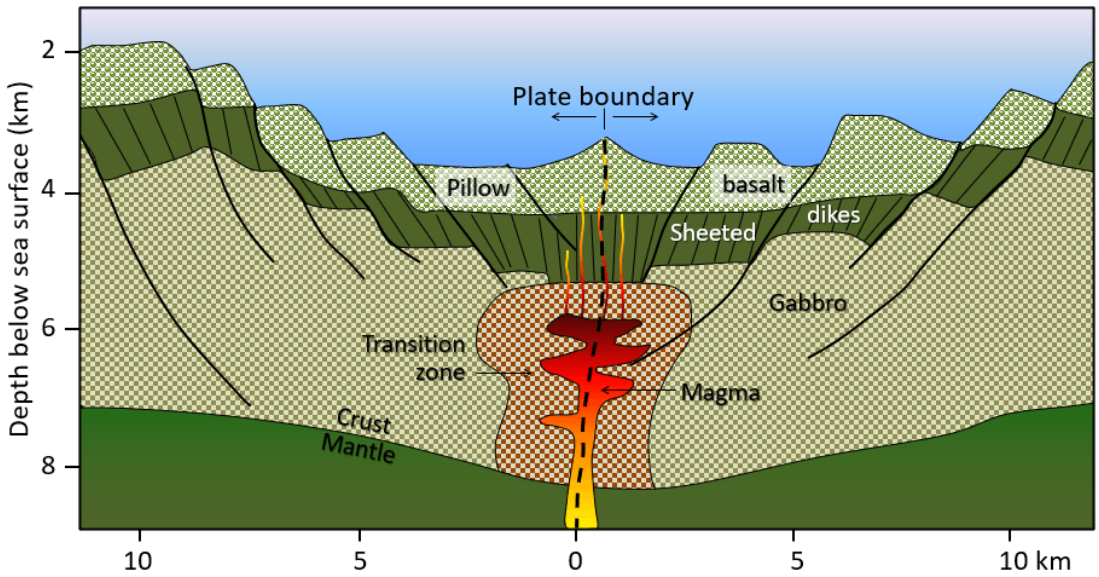


Figure 15.2.2.4 Depiction of the processes and materials formed at a divergent boundary.

Spreading is hypothesized to start within a continental area with up-warping or doming related to an underlying mantle plume or series of mantle plumes. The buoyancy of the mantle plume material creates a dome within the crust, causing it to fracture in a radial pattern, with three arms spaced at approximately 120° (Figure 15.2.2.5). When a series of mantle plumes exists beneath a large continent, the resulting rifts may align and lead to the formation of a rift valley (such as the present-day Great Rift Valley in eastern Africa). It is suggested that this type of valley eventually develops into a linear sea (such as the present-day Red Sea), and finally into an ocean (such as the Atlantic). It is likely that as many as 20 mantle plumes, many of which still exist, were responsible for the initiation of the rifting of Pangea along what is now the mid-Atlantic ridge (see Figure 15.2.2.10).

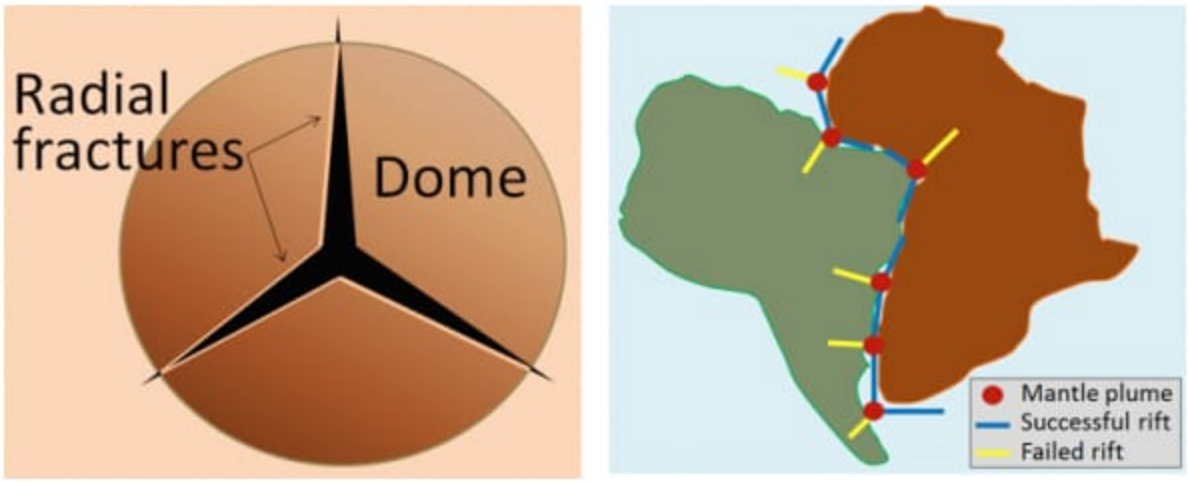


Figure 15.2.2.5 Depiction of the process of dome and three-part rift formation (left) and of continental rifting between the African and South American parts of Pangea at around 200 Ma (right).

Convergent Boundaries

Convergent boundaries, where two plates are moving toward each other, are of three types, depending on whether oceanic or continental crust is present on either side of the boundary. The types are ocean-ocean, ocean-continent, and continent-continent.

At an ocean-ocean convergent boundary, one of the plates (oceanic crust and lithospheric mantle) is pushed, or subducted**,** under the other. Often it is the older and colder plate that is denser and subducts beneath the younger and hotter plate. There is commonly an ocean trench along the boundary. The subducted lithosphere descends into the hot mantle at a relatively shallow angle close to the subduction zone, but at steeper angles farther down (up to about 45°). As discussed in the context of subduction-related volcanism in Chapter 4, the significant volume of water within the subducting material is released as the subducting crust is heated. Most of this water is present within the sheet silicate mineral serpentine which is derived from alteration of pyroxene and olivine near the spreading ridge shortly after the rock's formation. It is released when the oceanic crust is heats and then rises and mixes with the overlying mantle. The addition of water to the hot mantle lowers the rocks's melting point and leads to the formation of magma (flux melting) (Figure 15.2.2.6). The magma, which is lighter than the surrounding mantle material, rises through the mantle and the overlying oceanic crust to the ocean floor where it creates a chain of volcanic islands known as an island arc. A mature island arc develops into a chain of relatively large islands (such as Japan or Indonesia) as more and more volcanic material is extruded and sedimentary rocks accumulate around the islands.

As described above in the context of Benioff zones (Figure 15.2.2.6), earthquakes take place close to the boundary between the subducting crust and the overriding crust. The largest earthquakes occur near the surface where the subducting plate is still cold and strong.

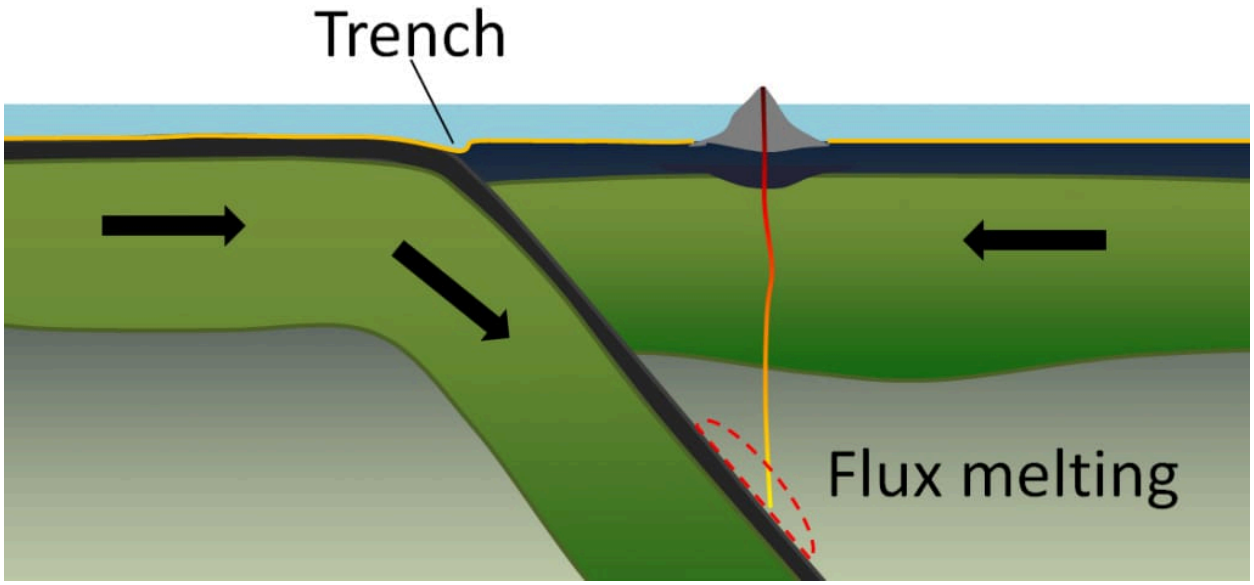


Figure 15.2.2.6 Configuration and processes of an ocean-ocean convergent boundary.

Examples of ocean-ocean convergent zones are subduction of the Pacific Plate beneath the North America Plate south of Alaska (Aleutian Islands) and beneath the Philippine Plate west of the Philippines, subduction of the India Plate beneath the Eurasian Plate south of Indonesia, and subduction of the Atlantic Plate beneath the Caribbean Plate (see Figure 15.2.2.1).

At an ocean-continent convergent boundary, the oceanic plate is pushed under the continental plate in the same manner as at an ocean-ocean boundary. Sediment that has accumulated on the **continental slope** is thrust up into an accretionary wedge, and compression leads to thrusting within the continental plate (Figure 15.2.2.7). The mafic magma produced adjacent to the subduction zone rises to the base of the continental crust and leads to partial melting of the crustal rock. The resulting magma ascends through the crust, producing a mountain chain with many volcanoes.

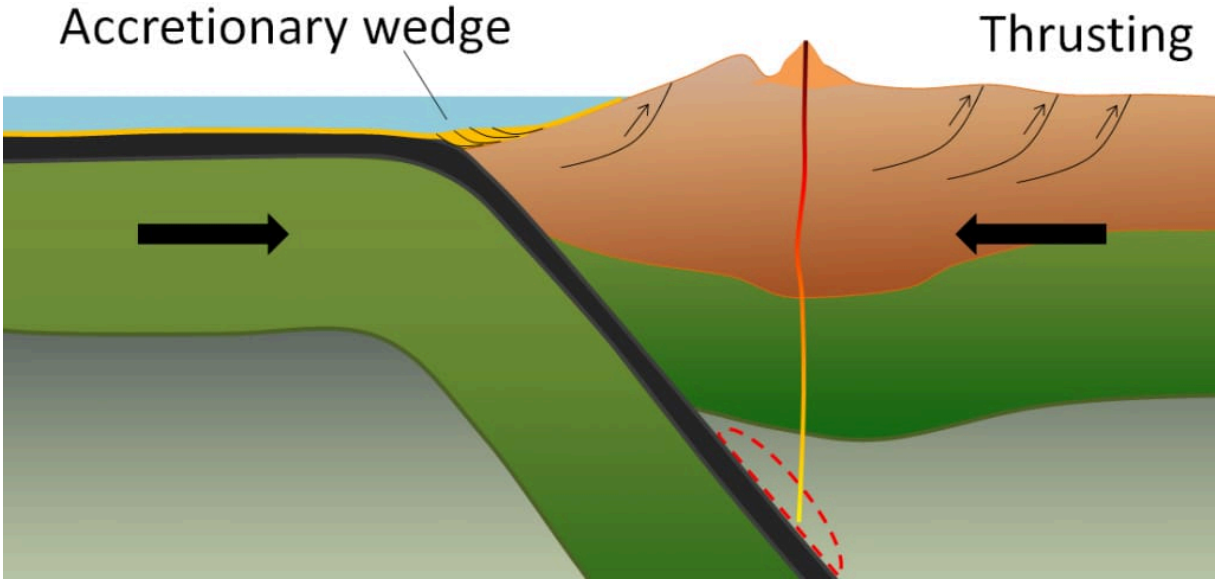


Figure 15.2.2.7 Configuration and processes of an ocean-continent convergent boundary.

Examples of ocean-continent convergent boundaries are subduction of the Nazca Plate under South America (which has created the Andes Range) and subduction of the Juan de Fuca Plate under North America (creating the mountains Garibaldi, Baker, St. Helens, Rainier, Hood, and Shasta, collectively known as the Cascade Range).

A continent-continent collision occurs when a continent or large island that has been moved along with subducting oceanic crust collides with another continent (Figure 15.2.2.8). The colliding continental material will not be subducted because it is too light (i.e., because it is composed largely of light continental rocks [SIAL]), but the root of the oceanic plate will eventually break off

and sink into the mantle. There is tremendous deformation of the pre-existing continental rocks, and creation of mountains from that rock, from any sediments that had accumulated along the shores (i.e., within geosynclines) of both continental masses, and commonly also from some ocean crust and upper mantle material.

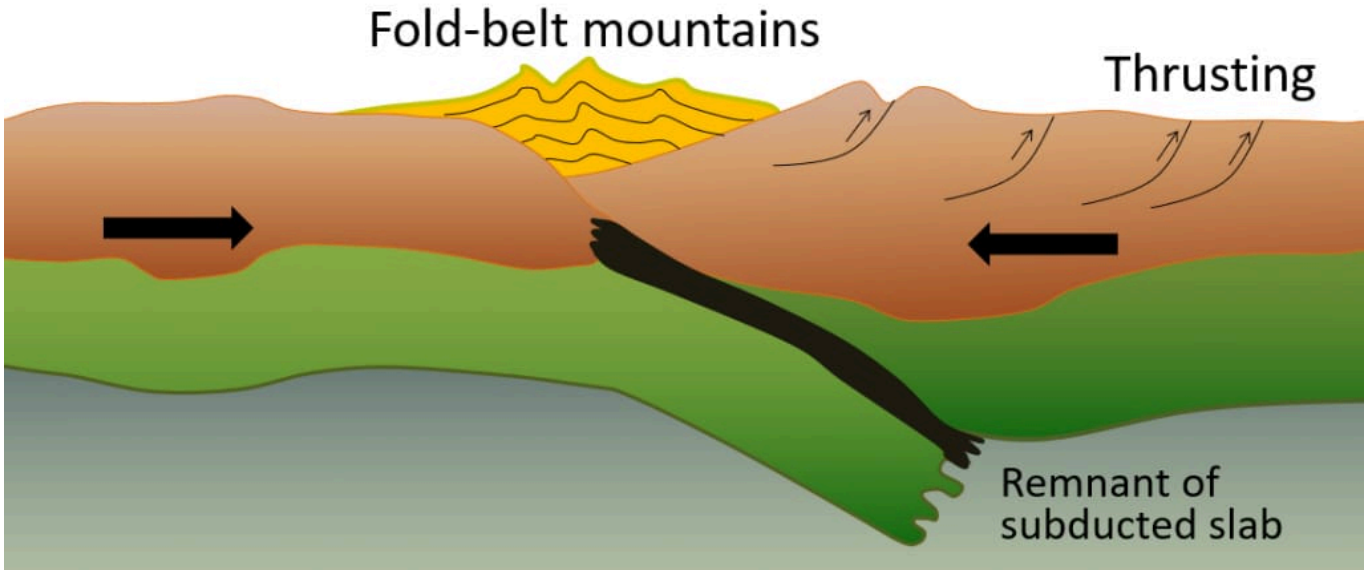


Figure 15.2.2.8 Configuration and processes of a continent-continent convergent boundary.

Examples of continent-continent convergent boundaries are the collision of the India Plate with the Eurasian Plate, creating the Himalaya Mountains, and the collision of the African Plate with the Eurasian Plate, creating the series of ranges extending from the Alps in Europe to the Zagros Mountains in Iran. The Rocky Mountains in B.C. and Alberta are also a result of continent-continent collisions.

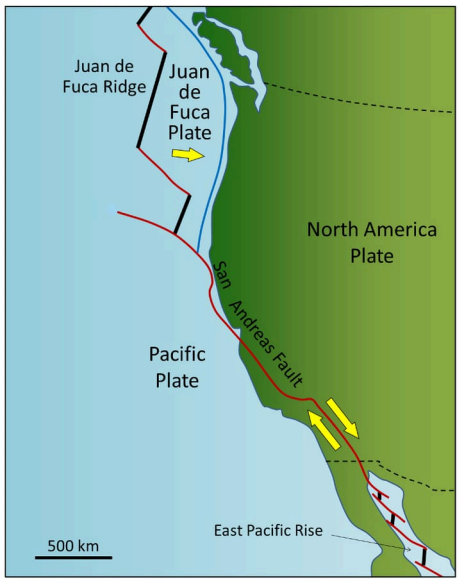


Figure 15.2.2.9 The San Andreas Fault extends from the north end of the East Pacific Rise in the Gulf of California to the southern end of the Juan de Fuca Ridge. All of the red lines on this map are transform faults.

Transform boundaries exist where one plate slides past another without production or destruction of crustal material. As explained above, most transform faults connect segments of mid-ocean ridges and are thus ocean-ocean plate boundaries (Figure 15.2.2.11). Some transform faults connect continental parts of plates. An example is the San Andreas Fault, which extends from the southern end of the Juan de Fuca Ridge to the northern end of the East Pacific Rise (ridge) in the Gulf of California (Figures 10.28 and 10.29). The part of California west of the San Andreas Fault and all of Baja California are on the Pacific Plate. Transform faults do not just connect divergent boundaries. For example, the Queen Charlotte Fault connects the north end of the Juan de Fuca Ridge, starting at the north end of Vancouver Island, to the Aleutian subduction zone.



Figure 15.2.2.10 The San Andreas Fault at Parkfield in central California. The person with the orange shirt is standing on the Pacific Plate and the person at the far side of the bridge is on the North American Plate. The bridge is designed to accommodate motion on the fault by sliding on its foundation.

Exercise 10.4 A different type of transform fault

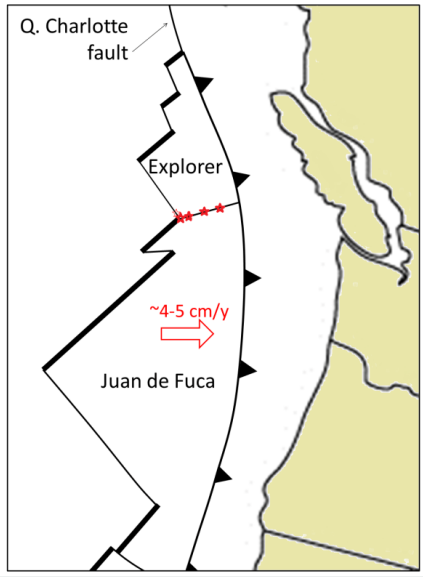


Figure 15.2.2.11

This map shows the Juan de Fuca (JDF) and Explorer Plates off the coast of Vancouver Island. We know that the JDF Plate is moving toward the North American Plate at around 4 centimeters per year to 5 centimeters per year. We think that the Explorer Plate is also moving east, but we don't know the rate, and there is evidence that it is slower than the JDF Plate.

The boundary between the two plates is the Nootka Fault, which is the location of frequent small-to-medium earthquakes (roughly up to magnitude 5), as depicted by the red stars. Explain why the Nootka Fault is a transform fault, and show the *relative* sense of motion along the fault with two small arrows.

See Appendix 3 for Exercise 10.4 answers.

As originally described by Wegener in 1915, the present continents were once all part of a supercontinent, which he termed **Pangea** (meaning *all land*). More recent studies of continental matchups and the magnetic ages of ocean-floor rocks have

enabled us to reconstruct the history of the break-up of Pangea.

Pangea began to rift apart along a line between Africa and Asia and between North America and South America at around 200 Ma. During the same period, the Atlantic Ocean began to open up between northern Africa and North America, and India broke away from Antarctica. Between 200 and 150 Ma, rifting started between South America and Africa and between North America and Europe, and India moved north toward Asia. By 80 Ma, Africa had separated from South America, most of Europe had separated from North America, and India had separated from Antarctica. By 50 Ma, Australia had separated from Antarctic, and shortly after that, India collided with Asia. To see the timing of these processes for yourself, go to time lapse of Continental Movements.

Within the past few million years, rifting has taken place in the Gulf of Aden and the Red Sea, and also within the Gulf of California. Incipient rifting has begun along the Great Rift Valley of eastern Africa, extending from Ethiopia and Djibouti on the Gulf of Aden (Red Sea) all the way south to Malawi.

Over the next 50 million years, it is likely that there will be full development of the east African rift and creation of new ocean floor. Eventually Africa will split apart. There will also be continued northerly movement of Australia and Indonesia. The western part of California (including Los Angeles and part of San Francisco) will split away from the rest of North America, and eventually sail right by the west coast of Vancouver Island, en route to Alaska. Because the oceanic crust formed by spreading on the mid-Atlantic ridge is not currently being subducted (except in the Caribbean), the Atlantic Ocean is slowly getting bigger, and the Pacific Ocean is getting smaller. If this continues without changing for another couple hundred million years, we will be back to where we started, with one supercontinent.

Pangea, which existed from about 350 to 200 Ma, was not the first supercontinent. It was preceded by Pannotia (600 to 540 Ma), by Rodinia (1,100 to 750 Ma), and by others before that.

In 1966, Tuzo Wilson proposed that there has been a continuous series of cycles of continental rifting and collision; that is, break-up of supercontinents, drifting, collision, and formation of other supercontinents. At present, North and South America, Europe, and Africa are moving with their respective portions of the Atlantic Ocean. The eastern margins of North and South America and the western margins of Europe and Africa are called **passive margins** because there is no subduction taking place along them.

This situation may not continue for too much longer, however. As the Atlantic Ocean floor gets weighed down around its margins by great thickness of continental sediments (i.e., geosynclines), it will be pushed farther and farther into the mantle, and eventually the oceanic lithosphere may break away from the continental lithosphere (Figure 15.2.2.12). A subduction zone will develop, and the oceanic plate will begin to descend under the continent. Once this happens, the continents will no longer continue to move apart because the spreading at the mid-Atlantic ridge will be taken up by subduction. If spreading along the mid-Atlantic ridge continues to be slower than spreading within the Pacific Ocean, the Atlantic Ocean will start to close up, and eventually (in a 100 million years or more) North and South America will collide with Europe and Africa.

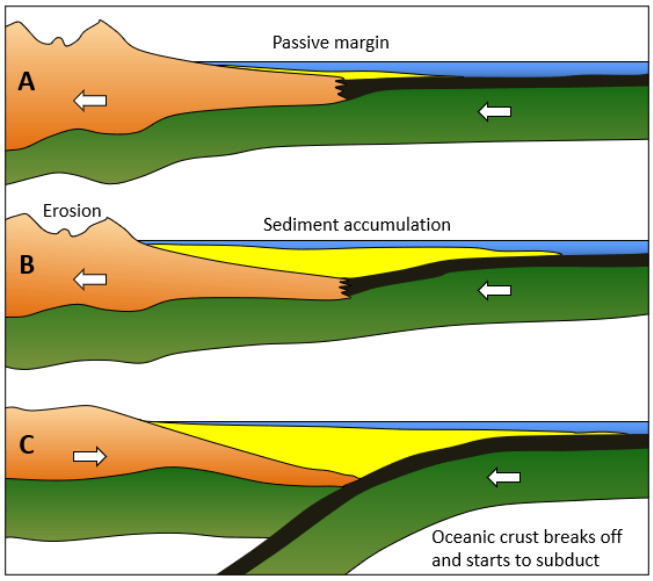


Figure 15.2.2.12 Development of a subduction zone at a passive margin. Times A, B, and C are separated by tens of millions of years. Once the oceanic crust breaks off and starts to subduct the continental crust (North America in this case) will no longer be pushed to the west and will likely start to move east because the rate of spreading in the Pacific basin is faster than that in the Atlantic basin.

There is strong evidence around the margins of the Atlantic Ocean that this process has taken place before. The roots of ancient mountain belts, which are present along the eastern margin of North America, the western margin of Europe, and the northwestern margin of Africa, show that these land masses once collided with each other to form a mountain chain, possibly as big as the Himalayas. The apparent line of collision runs between Norway and Sweden, between Scotland and England, through Ireland, through Newfoundland, and the Maritimes, through the northeastern and eastern states, and across the northern end of Florida. When rifting of Pangea started at approximately 200 Ma, the fissuring was along a different line from the line of the earlier collision. This is why some of the mountain chains formed during the earlier collision can be traced from Europe to North America and from Europe to Africa.

That the Atlantic Ocean rift may have occurred in approximately the same place during two separate events several hundred million years apart is probably no coincidence. The series of hot spots that has been identified in the Atlantic Ocean may also have existed for several hundred million years, and thus may have contributed to rifting in roughly the same place on at least two separate occasions (Figure 15.2.2.13).

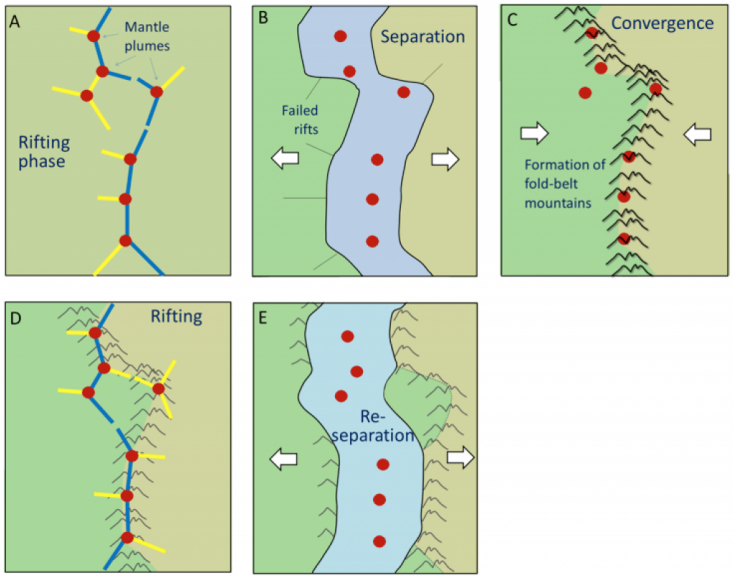


Figure 15.2.2.13 A scenario for the Wilson cycle. (A) The cycle starts with continental rifting above a series of mantle plumes. (B) The continents separate, and then (C) re-converge some time later, forming a fold-belt mountain chain. (D) Eventually rifting is repeated, possibly because of the same set of mantle plumes, but this time the rift is in a different place.

Exercise 10.5 Getting to know the plates and their boundaries

This map shows the boundaries between the major plates. Without referring to the plate map in Figure 15.2.2.1, or any other resources, write in the names of as many of the plates as you can. Start with the major plates, and then work on the smaller ones. Don't worry if you can't name them all.

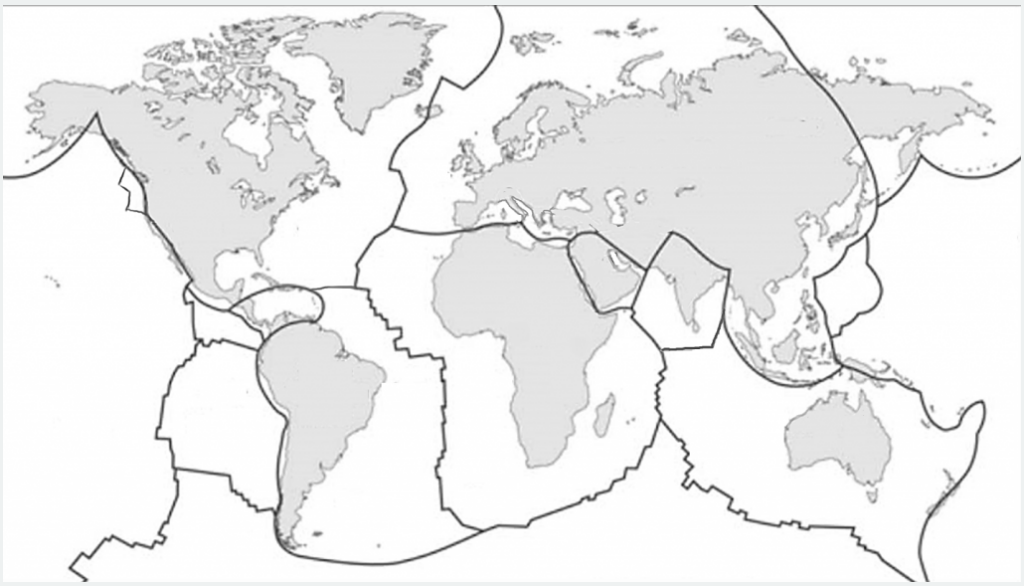


Figure 15.2.2.14

Once you've named most of the plates, draw arrows to show the general plate motions. Finally, using a highlighter or colored pencil, label as many of the boundaries as you can as divergent, convergent, or transform.

See Appendix 3 for Exercise 10.5 answers.

Image descriptions

Figure 15.2.2.1 image description: Descriptions of 15 different plates and their movements.

Plate name	Description of plate	Bordering plates (ordered from longest border to shortest)	Description of movement
Africa plate	This plate includes all of Africa and the surrounding ocean, including the eastern Atlantic Ocean, the surrounding Antarctic Ocean, and the western Indian ocean.	South America plate, Antarctic plate, Eurasia plate, North America plate, Arabia plate, India plate, Australia plate	This plate is moving north east towards the Arabia and Eurasia plates.
Antarctic plate.	This plate makes up all of Antarctica and much of the surrounding ocean.	Pacific plate, Australia plate, Africa plate, Scotia plate, Nazca plate, South America plate.	The part of the plate around the South America plate is moving northwards and a little east. The part of the plate around the Australia plate is moving southwards.
Arabia plate	This plate includes all of Saudi Arabia, and much of the Levant (up to Iraq and Syria).	Eurasia plate, Africa plate, India plate	This plate is moving north east towards the Eurasia plate.

Plate name	Description of plate	Bordering plates (ordered from longest border to shortest)	Description of movement
Australia plate	This plate includes Australia and much of the surrounding ocean. New Guinea and the northern parts of New Zealand are part of the Australia plate. The ocean area along southern Asia up to the India plate is also a part of the Australia plate.	Antarctic plate, Pacfic plate, Eurasia plate, India Plate, Africa plate.	This plate is moving north east towards the Eurasia plate and the Pacific plate.
Caribbean plate	This plate is small. It includes the central Caribbean countries and runs along the northern edge of South America.	North America plate, South America plate, Cocos plate.	N/A
Cocos plate	This plate is small. It runs along the west coast of Mexico and western Caribbean countries.	Nazca plate, Pacific plate, North America plate, Caribbean plate.	This plate is moving north east towards the Caribbean and North America plates.
Eurasia plate	This plate includes the northeastern part of the Atlantic ocean, all of Europe, all of Russia (except its most eastern part), and down through southeast Asia, including China and Indonesia.	North America plate, Africa plate, Australia plate, Arabia plate, India plate, Filipino plate.	This plate is rotating in a clockwise direction towards the Pacific plate.
Filipino plate	This plate includes the islands that make up the Philipines and north to include parts of southern Japan.	Eurasia plate, Pacific plate.	This plate is moving north west towards the Eurasia plate.
India plate	This plate includes India and the surrounding India Ocean.	Australia plate, Eurasia plate, Africa plate, Arabia plate.	This plate is moving north north east towards the Eurasia plate.
Juan de Fuca plate	This plate is small. It runs along the north western coast of the United States and the southern British Columbia coast.	Pacific plate, North America plate.	N/A
Nazca plate	This plate is in the Pacific Ocean between the Pacific plate and the South America plate.	South America plate, Pacific plate, Antarctic plate, Cocos plate	This plate is moving directly east towards the South America plate.
North America Plate	This plate includes all of North America, Greenland, the eastern most part of Russia, northern Japan, and the northwestern part of the Atlantic ocean.	Eurasia plate, Pacific plate, Africa plate, Caribbean plate, South America plate, Cocos plate, Juan de Fuca plate	This plate is rotating counter clockwise in towards the Pacific plate.
Pacific plate	This plate makes up most of the Pacific Ocean.	North America plate, Australia plate, Antarctic plate, Nazca plate, Filipino plate, Cocos plate, Juan de Fuca plate	This plate is moving northwest towards the Australia, Filipino, and Eurasia plates.
Scotia plate	This plate is small. It runs from the tip of South America eastwards to form a barrier between the Antarctic plate and the South America plate.	Antarctic plate, South America plate.	N/A
South America plate	This plate starts at the western edge of South America and stretches east into the southwestern parst of the Atlantic Ocean.	Africa plate, Nazca plate, Scotia plate, Caribbean plate, Antarctic plate, North America plate	This plate moves north and slightly west towards the Caribbean plate and the North America plate.

[Return to Figure 15.2.2.1]

Media Attributions

- Figure 15.2.2.1: “Plates tect2 en” by the USGS. Adapted by Steven Earle. Public domain.
- Figures 10.4.2, 10.4.3, 10.4.5, 10.4.6, 10.4.7, 10.4.8, 10.4.9, 10.4.10, 10.4.11, 10.4.12, 10.4.13, 10.4.14: © Steven Earle. CC BY.
- Figure 15.2.2.4: © Steven Earle. CC BY. Based on Keary and Vine, 1996, Global Tectonics (2ed), Blackwell Science Ltd., Oxford.

This page titled 15.2.2: Plate, Plate Motions, and Plate Boundary Processes is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Steven Earle (BCCampus) via source content that was edited to the style and standards of the LibreTexts platform.

- 10.4: Plate, Plate Motions, and Plate Boundary Processes by Steven Earle is licensed CC BY 4.0. Original source: https://opentextbc.ca/physicalgeology2ed.

15.2.3: Mechanisms for Plate Motion

It has been often repeated in this text and elsewhere that convection of the mantle is critical to plate tectonics, and while this is almost certainly so, other forces likely play a significant role. One side in the argument holds that the plates are only moved by the traction caused by mantle convection. The other side holds that traction plays only a minor role and that two other forces, ridge-push and slab-pull, are more important (Figure 15.2.3.1). Some argue that the real answer lies somewhere in between.

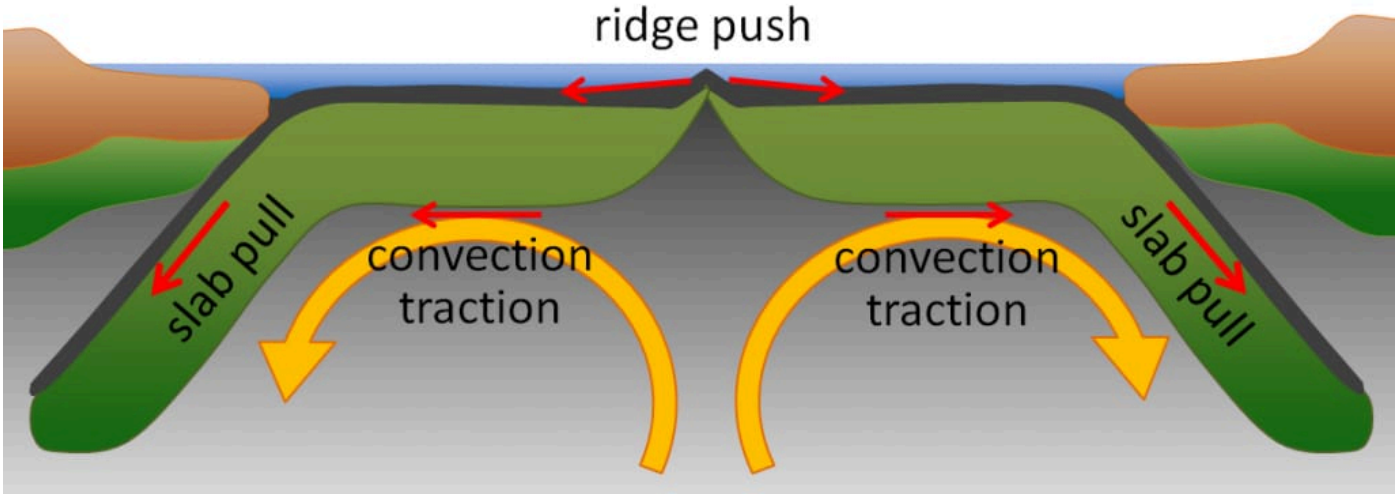


Figure 15.2.3.1 Models for plate motion mechanisms.

Kearey and Vine (1996)[1] have listed some compelling arguments in favour of the **ridge-push/slab-pull** model, as follows: (a) plates that are attached to subducting slabs (e.g., Pacific, Australian, and Nazca Plates) move the fastest, and plates that are not (e.g., North American, South American, Eurasian, and African Plates) move significantly slower; (b) in order for the traction model to apply, the mantle would have to be moving about five times faster than the plates are moving (because the coupling between the partially liquid asthenosphere and the plates is not strong), and such high rates of convection are not supported by geophysical models; and (c) although large plates have potential for much higher convection traction, plate velocity is not related to plate area.

In the ridge-push/slab-pull model, which is the one that has been adopted by most geologists working on plate-tectonic problems, the lithosphere is the upper surface of the convection cells, as is illustrated in Figure 15.2.3.2.

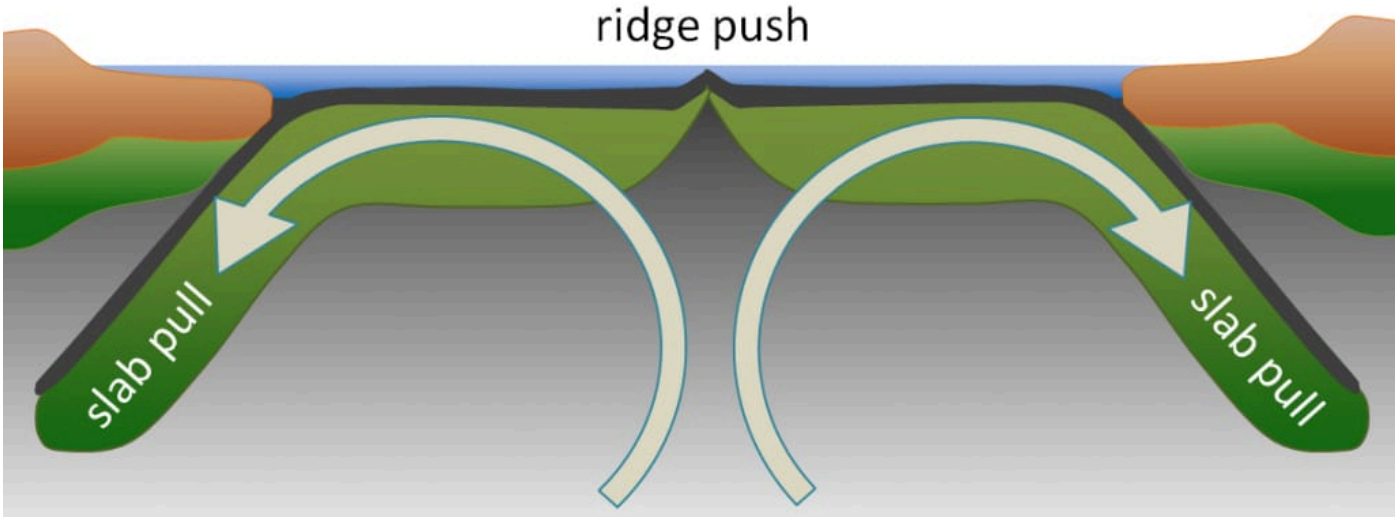


Figure 15.2.3.2 The ridge-push/slab-pull model for plate motion, in which the lithosphere is the upper surface of the convective systems.

Although ridge-push/slab-pull is the widely favoured mechanism for plate motion, it's important not to underestimate the role of mantle convection. Without convection, there would be no ridges to push from because upward convection brings hot buoyant rock to surface. Furthermore, many plates, including our own North American Plate, move along nicely—albeit slowly—without any slab-pull happening.

Image Descriptions

Figure 15.2.3.1 **image description:** In this model, there are three forces working to move the plates. Ridge-push forces cause two plates to pull apart on the surface. Slab-pull forces pull the plates down. This movement of out and down is also encouraged by convection traction, or clockwise and counterclockwise currents that are present beneath the plates. [Return to Figure 15.2.3.1]

Media Attributions

- Figures 10.5.1, 10.5.2: © Steven Earle. CC BY.

---

1. Kearey and Vine , 1996, Global Tectonics (2ed), Blackwell Science Ltd., Oxford ↩

This page titled 15.2.3: Mechanisms for Plate Motion is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Steven Earle (BCCampus) via source content that was edited to the style and standards of the LibreTexts platform.

- 10.5: Mechanisms for Plate Motion by Steven Earle is licensed CC BY 4.0. Original source: https://opentextbc.ca/physicalgeology2ed.

15.3: Earthquakes

This section examines earthquakes as the result of sudden movement within Earth's crust. It explains how earthquakes occur, how they are related to plate tectonics, and how their location and magnitude are measured. The section also considers the effects and hazards associated with earthquakes and why their impacts can vary considerably from one event to another.

This page titled 15.3: Earthquakes is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Robert Wagner.

15.3.1: What is an Earthquake?

An earthquake is the shaking caused by the **rupture** (breaking) and subsequent displacement of rocks (one body of rock moving with respect to another) beneath Earth's surface.

A body of rock that is under stress becomes deformed. When the rock can no longer withstand the deformation, it breaks and the two sides slide past each other. Most earthquakes take place near plate boundaries—but not necessarily right on a boundary—and not necessarily even on a pre-existing fault.

The engineering principle of **elastic deformation**, which can be used to understand earthquakes, is illustrated in Figure 15.3.1.1 . The stress applied to a rock—typically because of ongoing plate movement—results in strain or deformation of the rock (Figure 15.3.1.1b). Because most rock is strong (unlike loose sand, for example), it can withstand a significant amount of deformation without breaking. But every rock has a deformation limit and will rupture (break) once that limit is reached. At that point, in the case of rocks within the crust, the rock breaks and there is displacement along the **rupture surface** (Figure 15.3.1.1c). The magnitude of the earthquake depends on the extent of the area that breaks (the area of the rupture surface) and the average amount of displacement (sliding).

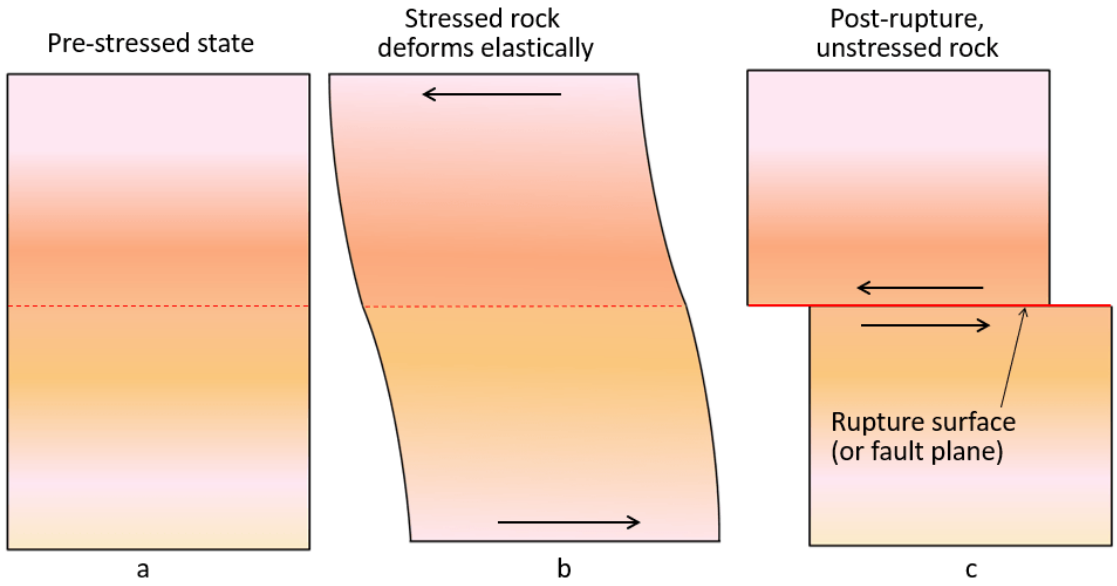


Figure 15.3.1.1 Depiction of the concept of elastic deformation and rupture. The plate boundary is shown as a dashed red line. In b the two plates are moving as shown by the arrows, but they are locked against each along the plate boundary so they are both deforming and the rocks are stressed. In c there has been a rupture along the boundary and the stress is released.

The concept of a rupture surface, which is critical to understanding earthquakes, is illustrated in Figure 15.3.1.2. An earthquake does not happen at a point, it happens over an area within a plane, although not necessarily a *flat* plane. Within the area of the rupture surface, the amount of displacement is variable (Figure 15.3.1.2), and, by definition, it decreases to zero at the edges of the rupture surface because the rock beyond that point isn't displaced at all. The extent of a rupture surface and the amount of displacement will depend on a number of factors, including the type and strength of the rock, and the degree to which it was stressed beforehand.

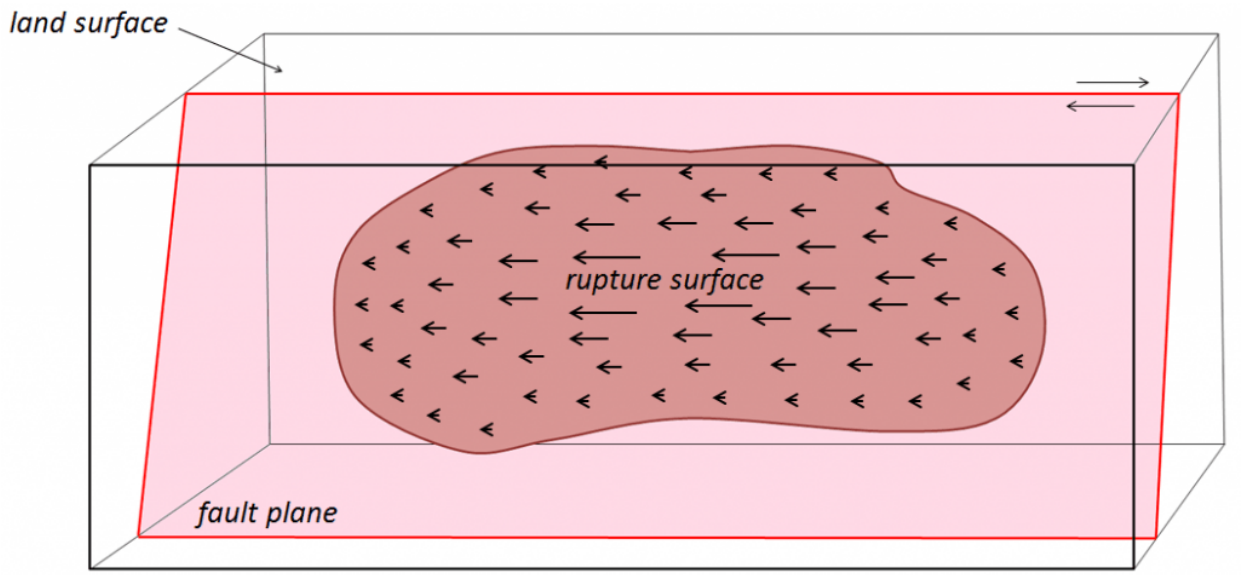


Figure 15.3.1.2 A rupture surface (dark pink), on a steeply dipping fault plane (light pink). The diagram represents a part of the crust that may be up to tens or hundreds of kilometers long. The rupture surface is the part of the fault plane along which displacement occurred. In this example, the near side of the fault is moving to the left, and the lengths of the arrows within the rupture surface represent relative amounts of displacement.

Earthquake rupture doesn't happen all at once; it starts at a single point and spreads rapidly from there. Depending on the extent of the rupture surface, the propagation of failures out from the point of initiation is typically completed within seconds to several tens of seconds (Figure 15.3.1.3). The initiation point isn't necessarily in the centre of the rupture surface; it may be close to one end, near the top, or near the bottom.

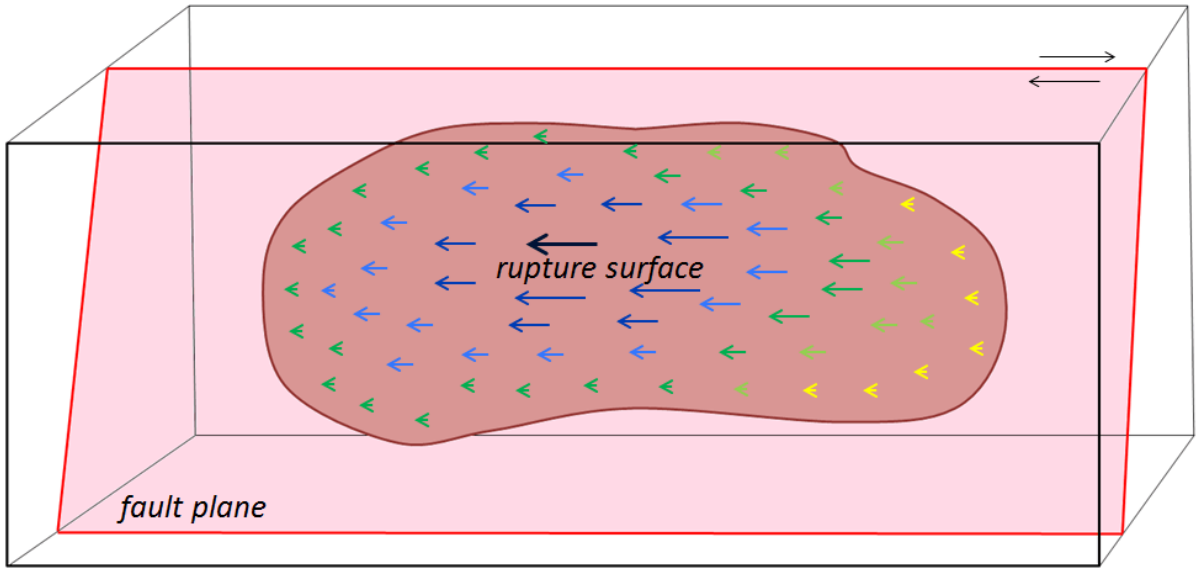


Figure 15.3.1.3 Propagation of failure on a rupture surface. In this case, the failure starts at the dark blue heavy arrow in the centre and propagates outward, reaching the left side first (green arrows) and the right side last (yellow arrows).

Figure 15.3.1.4 shows the distribution of immediate **aftershocks** associated with the 1989 Loma Prieta earthquake. Panel (b) is a section along the San Andreas Fault; this view is equivalent to what is shown in Figures 11.1.2 and 11.1.3. The area of red dots is the rupture surface; each red dot is a specific aftershock that was recorded on a seismometer. The hexagon labelled "main earthquake" represents the first or main shock. When that happened, the rock at that location broke and was displaced. That released the stress on that particular part of the fault, but it resulted in an increase of the stress on other nearby parts of the fault, and contributed to a cascade of smaller ruptures (aftershocks), in this case, over an area about 60 kilometers long and 15 kilometers wide.

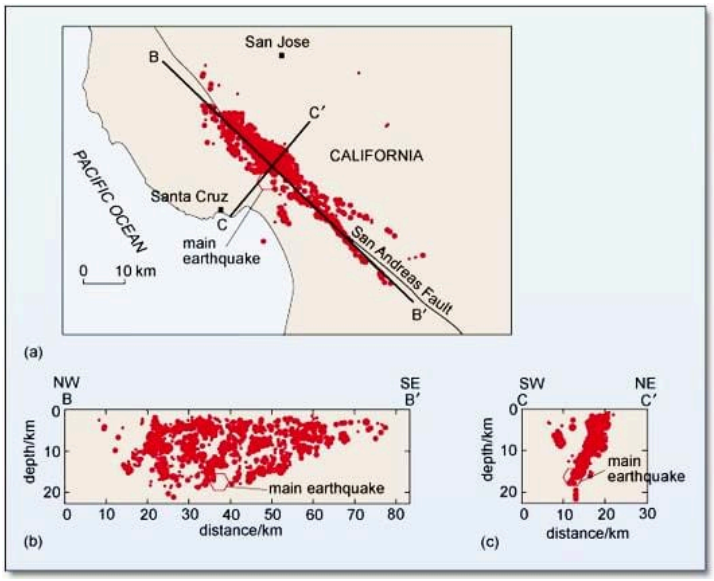


Figure 15.3.1.4 Distribution of the aftershocks of the 1989 M 6.9 Loma Prieta earthquake (a: plan view, b: section along the fault, c: section across the fault.)

So, what exactly is an aftershock then? An aftershock is an earthquake just like any other, but it is one that can be shown to have been triggered by **stress transfer** from a preceding earthquake. Within a few tens of seconds of the main Loma Prieta earthquake, there were hundreds of smaller aftershocks; their distribution defines the area of the rupture surface.

Aftershocks can be of any magnitude. Most are smaller than the earthquake that triggered them, but they can be bigger. The aftershocks shown in Figure 15.3.1.4 all happened within seconds or minutes of the main shock, but aftershocks can be delayed for hours, days, weeks, or even years. As already noted, aftershocks are related to stress transfer. For example, the main shock of the Loma Prieta earthquake triggered aftershocks in the immediate area, which triggered more in the surrounding area, eventually extending for 30 kilometers along the fault in each direction and for 15 kilometers toward the surface. But the earthquake as a whole also changed the stress on adjacent parts of the San Andreas Fault. This effect, which has been modelled for numerous earthquakes and active faults around the world, is depicted in Figure 15.3.1.5. Stress was reduced in the area of the rupture (blue), but was increased at either end of the rupture surface (red and yellow).

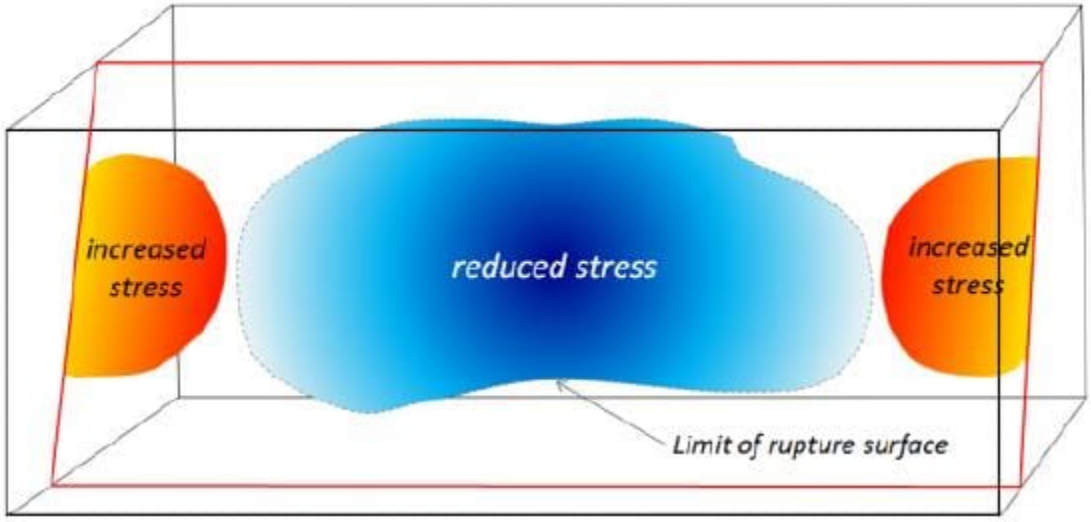


Figure 15.3.1.5 Depiction of stress changes related to an earthquake. Stress decreases in the area of the rupture surface, but increases on adjacent parts of the fault.

Stress transfer isn't necessarily restricted to the fault along which an earthquake happened. It will affect the rocks in general around the site of the earthquake and may lead to increased stress on other faults in the region. The effects of stress transfer don't necessarily show up right away. Segments of faults are typically in some state of stress, and the transfer of stress from another

area is only rarely enough to push a fault segment beyond its limits to the point of rupture. The stress that is added by stress transfer accumulates along with the ongoing buildup of stress from plate motion and eventually leads to another earthquake.

Episodic Tremor and Slip

Episodic tremor and slip (ETS) is periodic slow sliding along part of a subduction boundary. It does not produce recognizable earthquakes, but does produce seismic tremor (rapid seismic vibrations on a seismometer). It was first discovered on the Vancouver Island part of the Cascadia subduction zone by Geological Survey of Canada geologists Herb Dragert and Garry Rogers.[1]

The boundary between the subducting Juan de Fuca Plate and the North America Plate can be divided into three segments (Figure 15.3.1.6). The cold upper part of the Juan de Fuca Plate boundary is locked. The plates are stuck and don't move, except with very large earthquakes that happen *approximately* every 500 years (the last one was approximately M9 on January 26, 1700). The warm lower part of the boundary is sliding continuously because the warm rock is weaker. The central part of the boundary isn't cold enough to be stuck, but isn't warm enough to slide continuously. Instead it slips episodically, approximately every 14 months for about 2 weeks, moving a few centimeters each time.

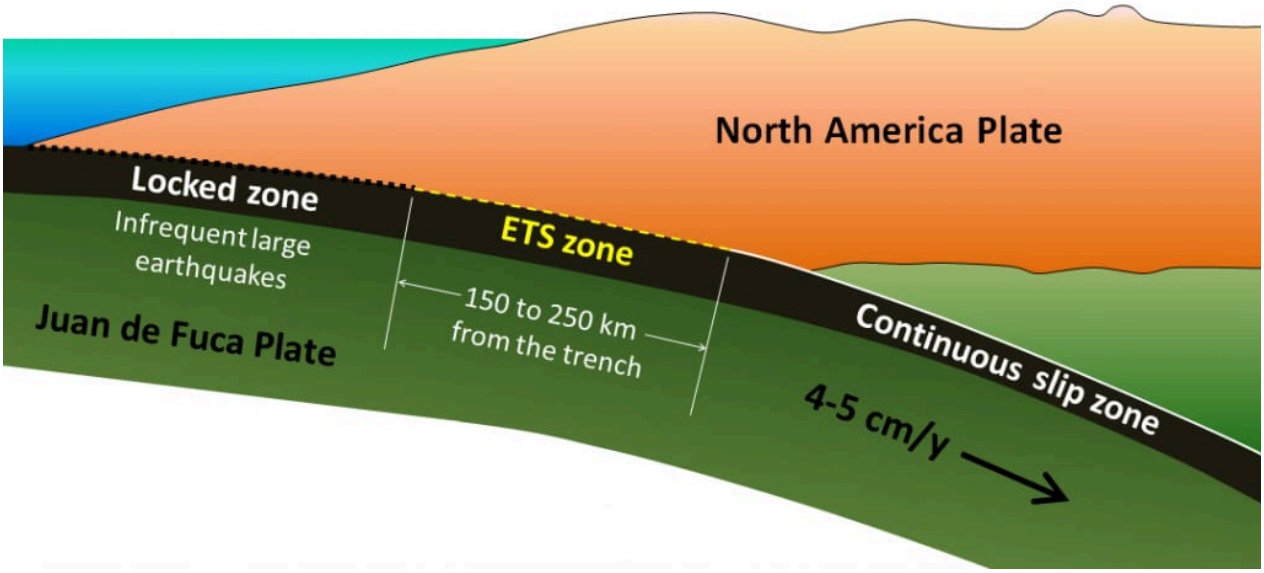


Figure 15.3.1.6 The boundary between the subducting Juan de Fuca Plate and the North America Plate is locked in the upper part, slides continuously in the lower part and slides episodically in the middle part.

You might be inclined to think that it's a good thing that there is periodic slip on this part of the plate because it releases some of the tension and reduces the risk of a large earthquake. In fact, the opposite is likely the case. The movement along the ETS part of the plate boundary acts like a medium-sized earthquake and leads to stress transfer to the adjacent locked part of the plate. Approximately every 14 months, during the two-week ETS period, there is a transfer of stress to the shallow locked part of the Cascadia subduction zone, and therefore an increased chance of a large earthquake.

Since 2003, ETS processes have also been observed on subduction zones in Mexico, New Zealand and Japan.

Media Attributions

- Figure 15.3.1.1, 11.1.2, 11.1.3, 11.1.6: © Steven Earle. CC BY.
- Figure 15.3.1.4: © Open University. CC BY-SA.
- Figure 15.3.1.5: © Steven Earle. CC BY. Based on data from 2010 Laguna Salada earthquake by Stein and Toda.

---

1. Rogers, G. and Dragert, H., 2003, Episodic tremor and slip on the Cascadia subduction zone: the chatter of silent slip, Science, V. 300, p. 1942-1943. ↩

This page titled 15.3.1: What is an Earthquake? is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Steven Earle (BCCampus) via source content that was edited to the style and standards of the LibreTexts platform.

- 11.1: What is an Earthquake? by Steven Earle is licensed CC BY 4.0. Original source: https://opentextbc.ca/physicalgeology2ed.

15.3.2: Earthquakes and Plate Tectonics

The distribution of earthquakes across the globe is shown in Figure 15.3.2.1. It is relatively easy to see the relationships between earthquakes and the plate boundaries. Along divergent boundaries like the mid-Atlantic ridge and the East Pacific Rise, earthquakes are common, but restricted to a narrow zone close to the ridge, and consistently at less than a 30 kilometre depth. Shallow earthquakes are also common along transform faults, such as the San Andreas Fault. Along subduction zones, as we saw in Chapter 10, earthquakes are very abundant, and they are increasingly deep on the landward side of the subduction zone.

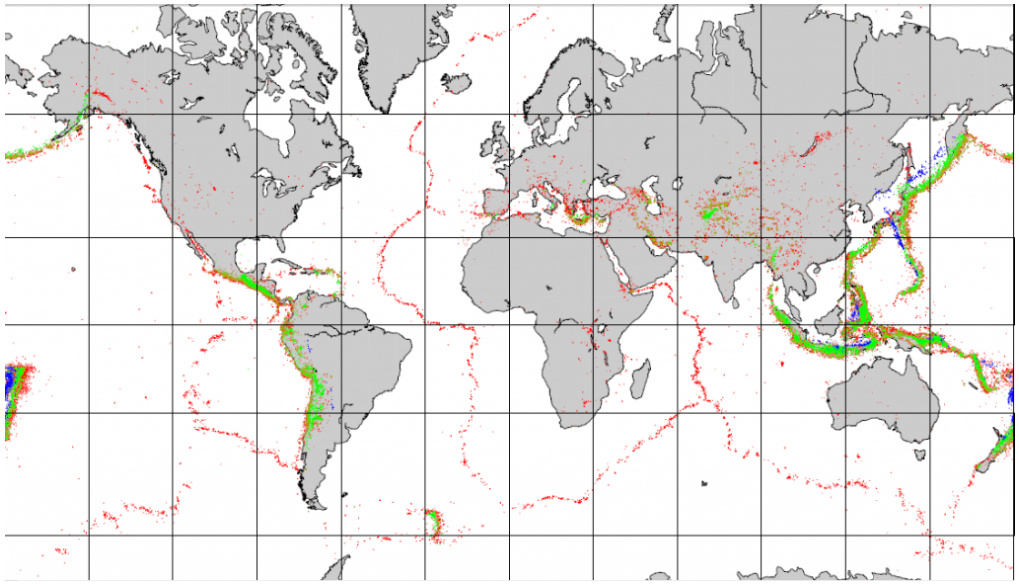


Figure 15.3.2.1 General distribution of global earthquakes of magnitude 4 and greater from 2004 to 2011, color coded by depth (red: 0 to 33 kilometers, orange 33 to 70 kilometers, green: 70 to 300 kilometers, blue: 300 to 700 kilometers).

Earthquakes are also relatively common at a few intraplate locations. Some are related to the buildup of stress due to continental rifting or the transfer of stress from other regions, and some are not well understood. Examples of intraplate earthquake regions include the Great Rift Valley area of Africa, the Tibet region of China, and the Lake Baikal area of Russia.

Earthquakes at Divergent and Transform Boundaries

Figure 15.3.2.2 provides a closer look at magnitude (M) 4 and larger earthquakes in an area of divergent boundaries in the mid-Atlantic region near the equator. Here, as we saw in Chapter 10, the segments of the mid-Atlantic ridge are offset by some long transform faults. Most of the earthquakes are located along the transform faults, rather than along the spreading segments, although there are clusters of earthquakes at some of the ridge-transform boundaries. Some earthquakes do occur on spreading ridges, but they tend to be small and infrequent because of the relatively high rock temperatures in the areas where spreading is taking place.

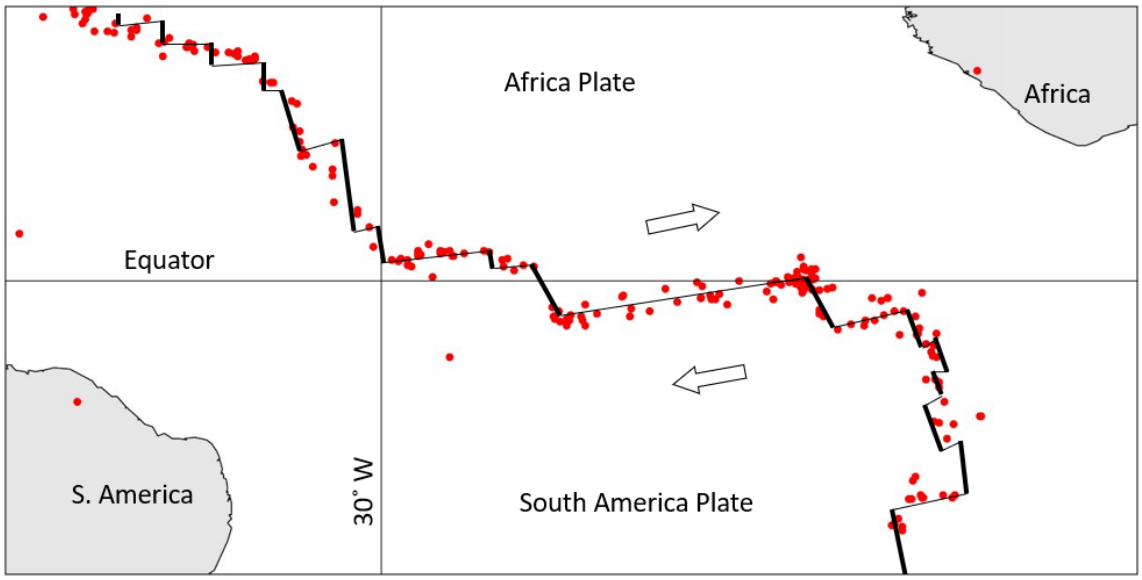


Figure 15.3.2.2 Distribution of earthquakes of M4 and greater in the area of the mid-Atlantic ridge near the equator from 1990 to 1996. All are at a depth of 0 to 33 kilometers.

Earthquakes at Convergent Boundaries

The distribution and depths of earthquakes in the Caribbean and Central America area are shown in Figure 15.3.2.3. In this region, the Cocos Plate is subducting beneath the North America and Caribbean Plates (ocean-continent convergence), and the South and North America Plates are subducting beneath the Caribbean Plate (ocean-ocean convergence). In both cases, the earthquakes get deeper with distance from the trench. In Figure 15.3.2.3, the South America Plate is shown as being subducted beneath the Caribbean Plate in the area north of Colombia, but since there is almost no earthquake activity along this zone, it is questionable whether subduction is actually taking place.

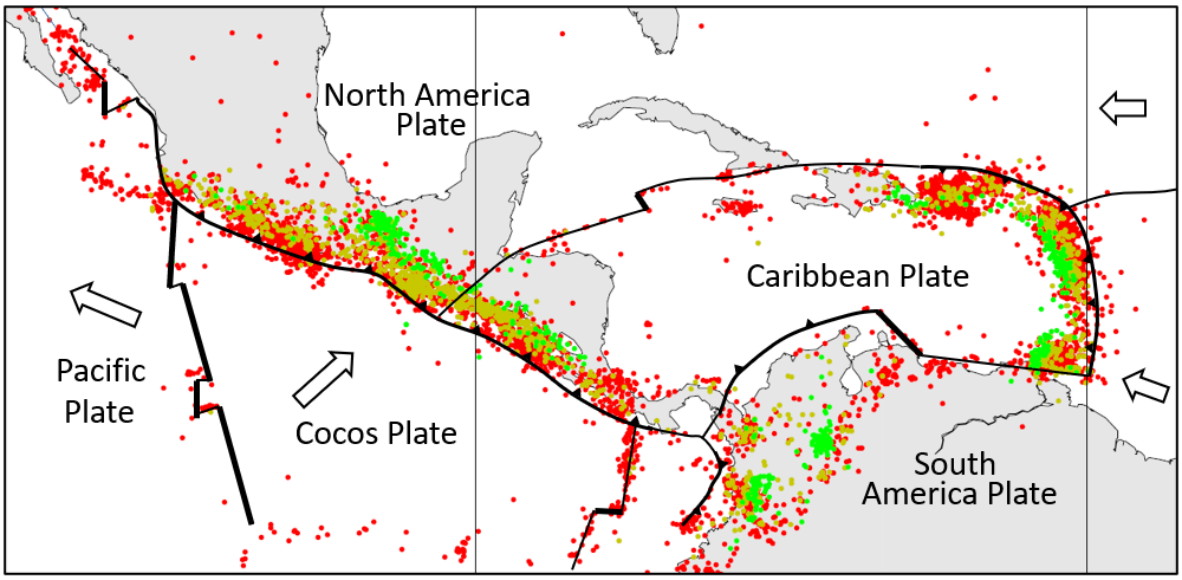


Figure 15.3.2.3 Distribution of earthquakes of M4 and greater in the Central America region from 1990 to 1996 (red: 0 to 33 kilometers, orange: 33 to 70 kilometers, green: 70 to 300 kilometers, blue: 300 to 700 kilometers) (Spreading ridges are heavy lines, subduction zones are toothed lines, and transform faults are light lines.)

There are also various divergent and transform boundaries in the area shown in Figure 15.3.2.3, and as we've seen in the mid-Atlantic area, most of these earthquakes occur along the transform faults.

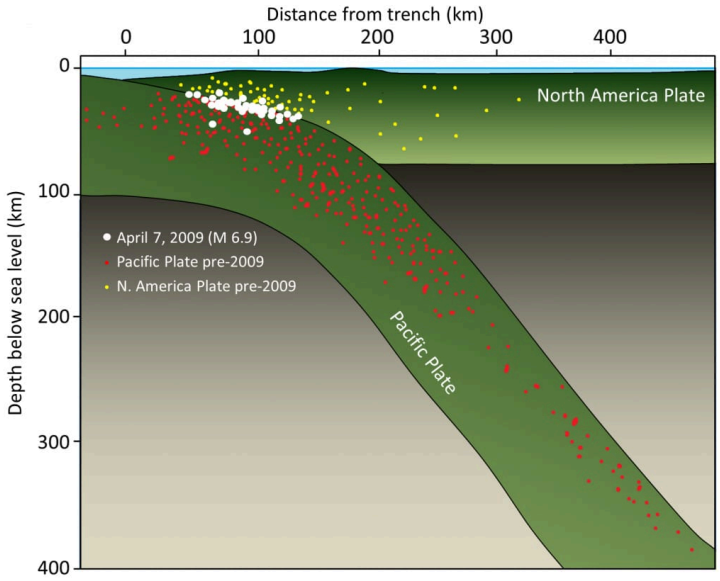


Figure 15.3.2.4 Distribution of earthquakes in the area of the Kuril Islands, Russia (just north of Japan) (White dots represent the April 2009 M6.9 earthquake. Red and yellow dots are from background seismicity over several years prior to 2009.)

The distribution of earthquakes with depth in the Kuril Islands of Russia in the northwest Pacific is shown in Figure 15.3.2.4. This is an ocean-ocean convergent boundary. The small red and yellow dots show background seismicity over a number of years, while the larger white dots are individual shocks associated with a M6.9 earthquake in April 2009. The relatively large earthquake took place on the upper part of the plate boundary between 60 kilometers and 140 kilometers inland from the trench. As we saw for the Cascadia subduction zone, this is where large subduction earthquakes are expected to occur.

In fact, all of the very large earthquakes — M9 or higher — take place at subduction boundaries because there is the potential for a greater width of rupture zone on a gently dipping boundary than on a steep transform boundary. The largest earthquakes on transform boundaries are in the order of M8.

The background seismicity at this convergent boundary, and on other similar ones, is predominantly near the upper side of the subducting plate. The frequency of earthquakes is greatest near the surface and especially around the area where large subduction quakes happen, but it extends to at least a 400 kilometre depth. There is also significant seismic activity in the overriding North America Plate, again most commonly near the region of large quakes, but also extending for a few hundred kilometers away from the plate boundary.

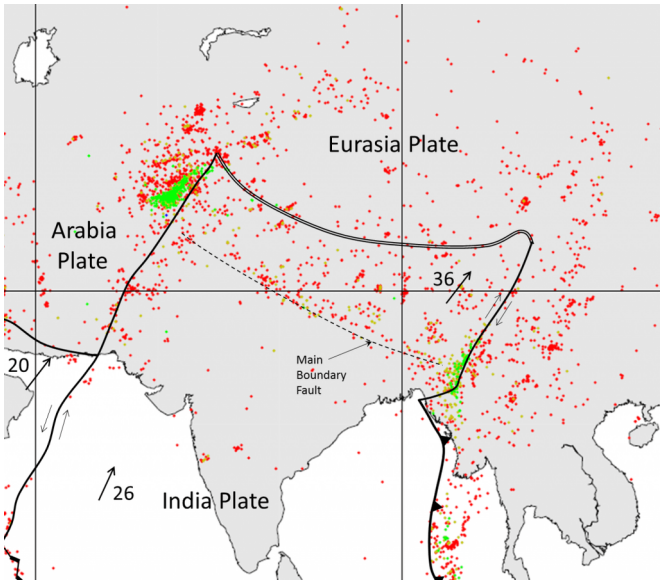


Figure 15.3.2.5 Distribution of earthquakes in the area where the India Plate is converging with the Asia Plate (data from 1990 to 1996, red: 0 to 33 kilometers, orange: 33 to 70 kilometers, green: 70 to 300 kilometers). (Spreading ridges are heavy lines, subduction zones are toothed lines, and transform faults are light lines. The double line along the northern edge of the India Plate indicates convergence, but not subduction. Plate motions are shown in millimeters per year).

The distribution of earthquakes in the area of the India-Eurasia plate boundary is shown in Figure 15.3.2.5. This is a continent-continent convergent boundary, and it is generally assumed that although the India Plate continues to move north toward the Asia Plate, there is no actual subduction taking place. There are transform faults on either side of the India Plate in this area.

The entire northern India and southern Asia region is very seismically active. Earthquakes are common in northern India, Nepal, Bhutan, Bangladesh and adjacent parts of China, and throughout Pakistan and Afghanistan. Many of the earthquakes are related to the transform faults on either side of the India Plate, and most of the others are related to the significant tectonic squeezing caused by the continued convergence of the India and Asia Plates. That squeezing has caused the Asia Plate to be thrust over top of the India Plate, building the Himalayas and the Tibet Plateau to enormous heights. Most of the earthquakes of Figure 15.3.2.5 are related to the thrust faults shown in Figure 15.3.2.6 (and to hundreds of other similar ones that cannot be shown at this scale). The southernmost thrust fault in Figure 15.3.2.6 is equivalent to the Main Boundary Fault in Figure 15.3.2.5.

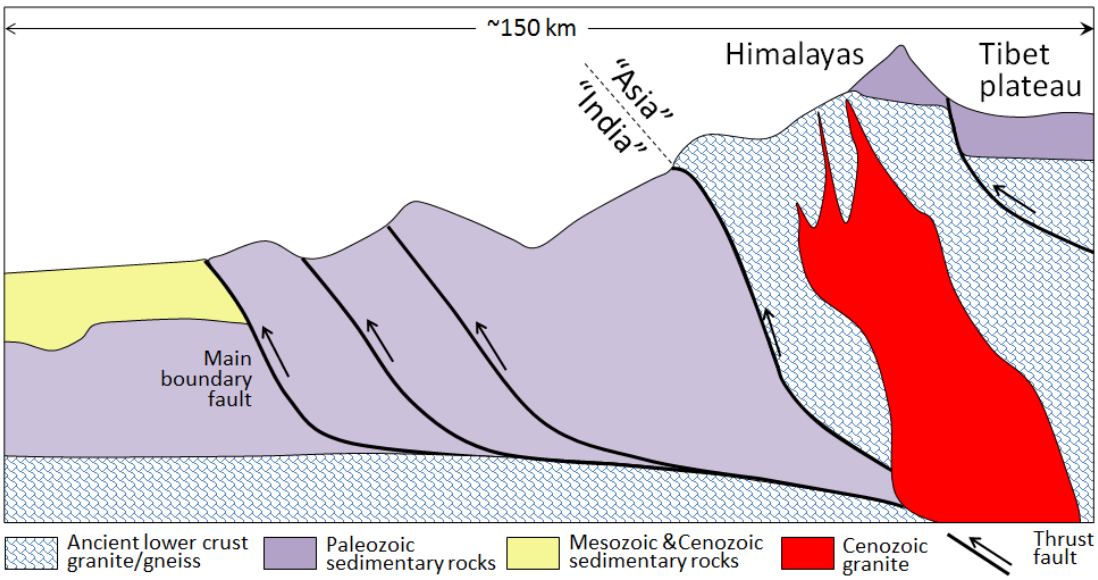


Figure 15.3.2.6 Schematic diagram of the India-Asia convergent boundary, showing examples of the types of faults along which earthquakes are focused. The devastating Nepal earthquake of May 2015 took place along one of these thrust faults.

There is a very significant concentration of both shallow and deep (greater than 70 kilometers) earthquakes in the northwestern

part of Figure 15.3.2.5. This is northern Afghanistan, and at depths of more than 70 kilometers, many of these earthquakes are within the mantle as opposed to the crust. It is interpreted that these deep earthquakes are caused by northwestward subduction of part of the India Plate beneath the Asia Plate in this area.

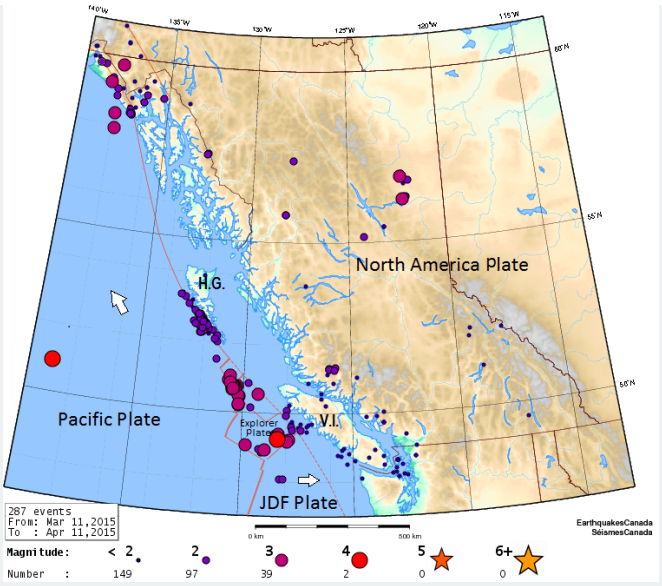


Figure 15.3.2.7

This map shows the incidence and magnitude of earthquakes in British Columbia over a one-month period in March and April 2015.

1. What is the likely origin of the earthquakes between the Juan de Fuca (JDF) and Explorer Plates?
2. The string of small earthquakes adjacent to Haida Gwaii (H.G.) coincides closely with the rupture surface of the 2012 M7.8 earthquake in that area. How might these earthquakes be related to that one?
3. Most of the earthquakes around Vancouver Island (V.I.) are relatively shallow. What is their likely origin?
4. Some of the earthquakes in B.C. are interpreted as being caused by natural gas extraction (including fracking). Which of the earthquakes here could fall into this category?

See Appendix 3 for Exercise 11.1 answers.

Image descriptions

Figure 15.3.2.7 image description: The incidence and magnitude of earthquakes in British Columbia over a one-month period in March and April 2015: There were a few dozen smaller earthquakes spread out around Vancouver Island and the sunshine coast with a magnitude of 2. Farther west along the Explorer Plate, which is between the North American plate, the Juan de Fuca Plate, and the Pacific Plate, there were quite a few earthquakes with a magnitude of 3 and at least one earthquake with a magnitude of 4. Between the North American Plate and the Pacific Plate off the south-west coast of Haida Gwaii, there was a large cluster of earthquakes with magnitudes of 2. Along the Alaskan panhandle, there was a collection of 2- and 3-magnitude earthquakes. In addition, there were two 3-magnitude earthquakes west of Fort St. John in northern British Columbia and one or two 2-magnitude earthquakes. In total, this map shows one hundred and forty-nine earthquakes with a magnitude less than 2, ninety-seven earthquakes with a magnitude of 2, thirty-nine earthquakes with a magnitude of 3, and two earthquakes with a magnitude of 4 [Return to Figure 15.3.2.7]

Media Attributions

- Figure 15.3.2.1: Global Earthquakes © Dale Sawyer, Rice University. Used with permission.
- Figure 15.3.2.2: Earthquakes Around the Mid-Atlantic Ridge © Steven Earle after Dale Sawyer, Rice University.
- Figure 15.3.2.3: Earthquakes Around the Central-American Region © Steven Earle after Dale Sawyer, Rice University.
- Figure 15.3.2.4: Earthquakes Around the Kuril Islands. © Steven Earle after Gavin Hayes, from data from the USGS [PDF].
- Figure 15.3.2.5: Earthquakes Around the India Plate © Steven Earle after Dale Sawyer, Rice University.

- Figure 15.3.2.6: India-Asia Convergent Boundary © Steven Earle based on D. Vouichard, from a United Nations University document.

This page titled 15.3.2: Earthquakes and Plate Tectonics is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Steven Earle (BCCampus) via source content that was edited to the style and standards of the LibreTexts platform.

- 11.2: Earthquakes and Plate Tectonics by Steven Earle is licensed CC BY 4.0. Original source: https://opentextbc.ca/physicalgeology2ed.

15.3.3: Measuring Earthquakes

There are two main ways to measure earthquakes. The first of these is an estimate of the energy released, and the value is referred to as **magnitude**. This is the number that is typically used by the press when a big earthquake happens. It is often referred to as "Richter magnitude," but that is a misnomer, and it should be just "magnitude." There are many ways to measure magnitude—including Charles Richter's method developed in 1935—but they are all ways to estimate the same number, which is proportional to the amount of energy released.

The other way of assessing the impact of an earthquake is to assess what people felt and how much damage was done. This is known as **intensity**. Intensity values are assigned to locations, rather than to the earthquake itself, and therefore intensity can vary widely, depending on the proximity to the earthquake and the types of materials underneath and the local conditions.

Earthquake Magnitude

Before we look more closely at magnitude we need to review what we know about body waves, and look at surface waves. Body waves are of two types, P waves, or primary or compression waves (like the compression of the coils of a spring), and S waves, or secondary or shear waves (like the flick of a rope). An example of P and S seismic wave records is shown in Figure 15.3.3.1 . The critical parameters for the measurement of magnitude are labelled, including the time interval between the arrival of the P- and S-waves—which is used to determine the distance from the earthquake to the seismic station, and the amplitude of the S-waves—which is used to estimate the magnitude of the earthquake.

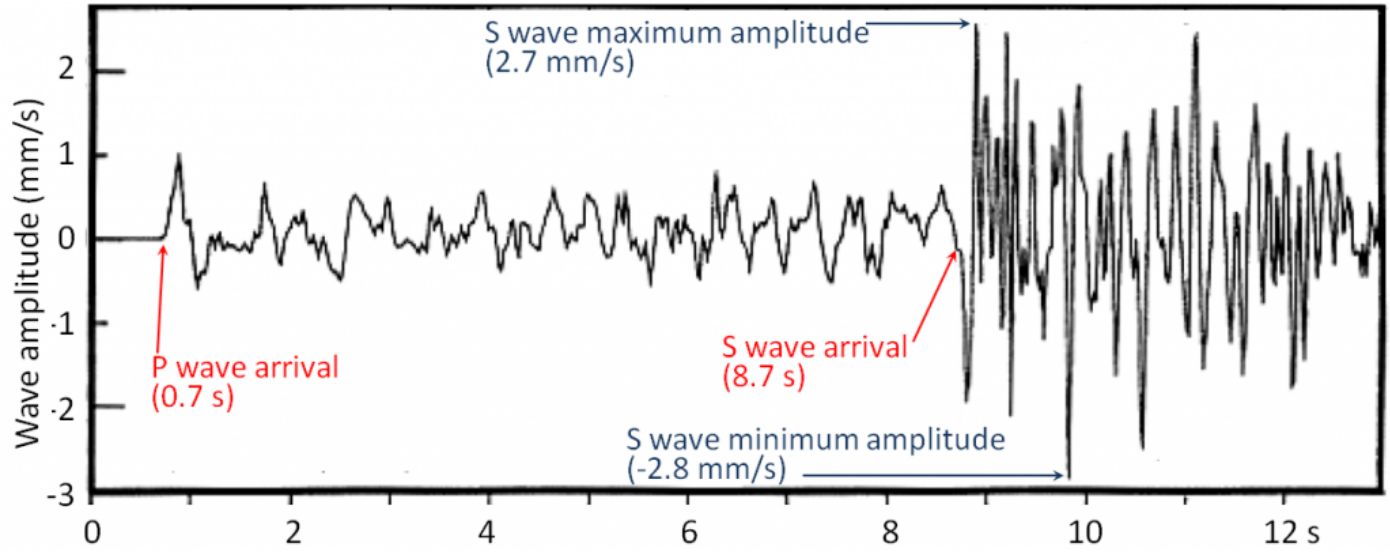


Figure 15.3.3.1 P-waves and S-waves from a small (M4) earthquake that took place near Vancouver Island in 1997.

When body waves (P or S) reach Earth's surface, some of their energy is transformed into surface waves, of which there are two main types, as illustrated in Figure 15.3.3.2. **Rayleigh waves** are characterized by vertical motion of the ground surface, like waves on water, while **Love waves** are characterized by horizontal motion. Both Rayleigh and Love waves are about 10% slower than S-waves (so they arrive later at a seismic station). Surface waves typically have greater amplitudes than body waves, and they do more damage.

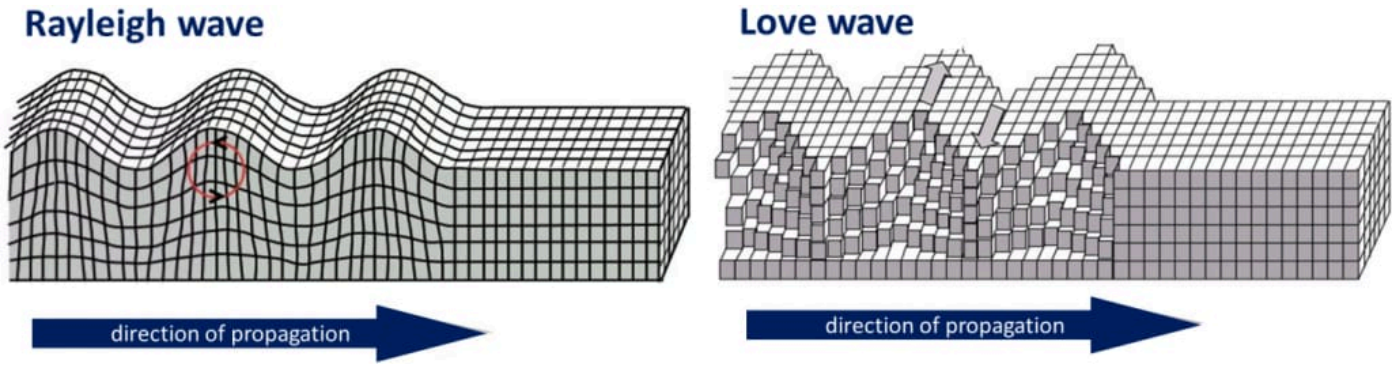


Figure 15.3.3.2 Depiction of seismic surface waves.

Other important terms for describing earthquakes are **hypocentre** (or **focus**) and **epicentre**. The hypocentre is the actual location of an individual earthquake shock at depth in the ground, and the epicentre is the point on the land surface vertically above the hypocentre (Figure 15.3.3.3).

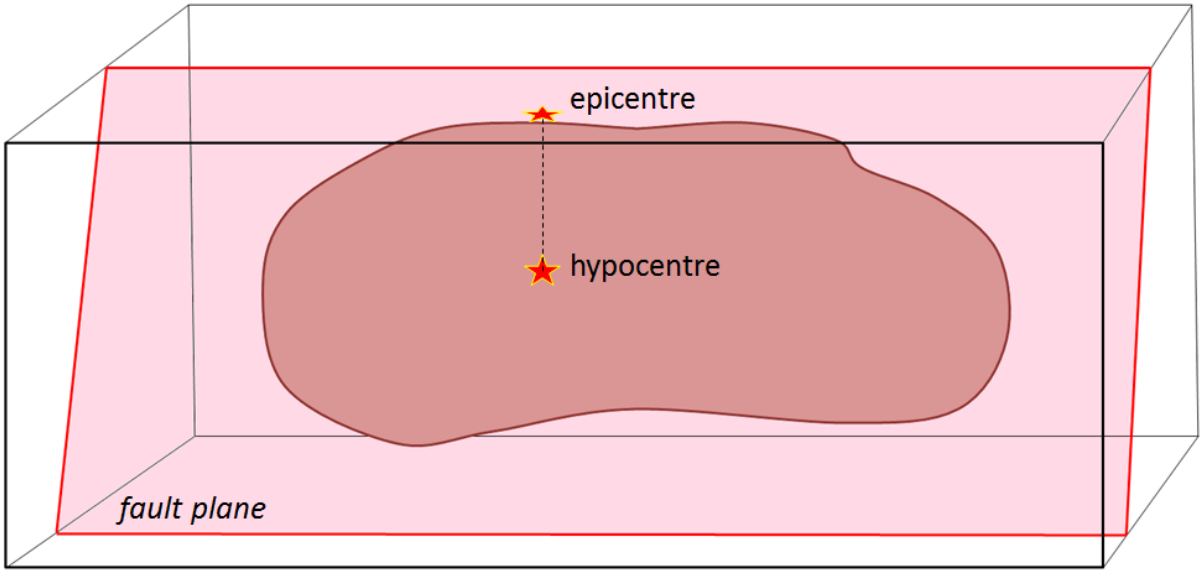


Figure 15.3.3.3 Epicentre and hypocentre.

A number of methods for estimating magnitude are listed in Table 11.1. Local magnitude (ML) was widely used until late in the 20th century, but **moment magnitude** (MW) is now more commonly used because it gives more accurate estimates (especially with larger earthquakes) and can be applied to earthquakes at any distance from a seismometer. Surface-wave magnitudes can also be applied to measure distant large earthquakes.

Because of the increasing size of cities in earthquake-prone areas (e.g., China, Japan, California) and the increasing sophistication of infrastructure, it is becoming important to have very rapid warnings and magnitude estimates of earthquakes that have already happened. This can be achieved by using P-wave data to determine magnitude because P-waves arrive first at seismic stations, in many cases several seconds ahead of the more damaging S-waves and surface waves. Operators of electrical grids, pipelines, trains, and other infrastructure can use the information to automatically shut down systems so that damage and casualties can be limited.

Table 11.1 A summary of some of the different methods for estimating earthquake magnitude.[1]

[Skip Table]

Type	M Range	Dist. Range	Comments
Local or Richter (M_L)	2 to 6	0 to 400 kilometers	The original magnitude relationship defined in 1935 by Richter and Gutenberg. It is based on the maximum amplitude of S-waves recorded on a Wood-Anderson torsion seismograph. ML values can be calculated using data from modern instruments. L stands for local because it only applies to earthquakes relatively close to the seismic station.
Moment (M_W)	Greater than 3.5	All	Based on the **seismic moment** of the earthquake, which is equal to the average amount of displacement on the fault times the fault area that slipped. It can also be estimated from seismic data if the seismometer is tuned to detect long-period body waves.
Surface wave (M_S)	5 to 8	20 to 180°	A magnitude for distant earthquakes based on the amplitude of surface waves measured at a period near 20 seconds.
P-wave	2 to 8	Local	Based on the amplitude of P-waves. This technique is being increasingly used to provide very rapid magnitude estimates so that early warnings can be sent to utility and transportation operators to shut down equipment before the larger (but slower) S-waves and surface waves arrive.

Exercise 11.2 Moment magnitude estimates from earthquake parameters

Use this moment magnitude calculation tool to estimate the moment magnitude based on the approximate length, width, and displacement values provided in the following table:

Table 11.2 Calculate Moment Magnitude Based on Length, Width, and Displacement Values

[Skip Table]

Length (kilometers)	Width (kilometers)	Displacement (meters)	Earthquake	MW?
60	15	4	The 1946 Vancouver Island earthquake	
0.4	0.2	.5	The small Vancouver Island earthquake shown in Figure 15.3.3.1	
20	8	4	The 2001 Nisqually earthquake described in Exercise 11.3	
1,100	120	10	The 2004 Indian Ocean earthquake	
30	11	4	The 2010 Haiti earthquake	

The largest recorded earthquake had a magnitude of 9.5. Could there be a 10? You can answer that question using this tool. See what numbers are needed to make MW = 10. Are they reasonable?

See Appendix 3 for Exercise 11.2 answers.

The magnitude scale is logarithmic; in fact, the amount of energy released by an earthquake of M4 is 32 times higher than that released by one of M3, and this ratio applies to all intervals in the scale. If we assign an arbitrary energy level of 1 unit to a M1 earthquake the energy for quakes up to M8 will be as shown on the following table:

Table 11.3 The energy of an earthquake increases by 32 times at each magnitude level.

Magnitude	Energy
1	1
2	32
3	1,024
4	32,768
5	1,048,576
6	33,554,432
7	1,073,741,824
8	34,359,738,368

In any given year, when there is a large earthquake on Earth (M8 or M9), the amount of energy released by that one event will likely exceed the energy released by all smaller earthquake events combined.

Earthquake Intensity

The intensity of earthquake shaking at any location is determined not only by the magnitude of the earthquake and its distance, but also by the type of underlying rock or unconsolidated materials. If buildings are present, the size and type of buildings (and their inherent natural vibrations) are also important.

Intensity scales were first used in the late 19th century, and then adapted in the early 20th century by Giuseppe Mercalli and modified later by others to form what we know call the modified Mercalli intensity scale (Table 11.4). Intensity estimates are important because they allow us to characterize parts of any region into areas that are especially prone to strong shaking versus those that are not. The key factor in this regard is the nature of the underlying geological materials, and the weaker those are, the more likely it is that there will be strong shaking. Areas underlain by strong solid bedrock tend to experience much less shaking than those underlain by unconsolidated river or lake sediments.

Table 11.4 The modified Mercalli intensity scale.

[Skip Table]

Level of intensity	Description
Not felt (1)	Not felt except by a very few under especially favourable conditions
Weak (2)	Felt only by a few persons at rest, especially on upper floors of buildings
Weak (3)	Felt quite noticeably by persons indoors, especially on upper floors of buildings; many people do not recognize it as an earthquake; standing motor cars may rock slightly; vibrations similar to the passing of a truck; duration estimated
Light (4)	Felt indoors by many, outdoors by few during the day; at night, some awakened; dishes, windows, doors disturbed; walls make cracking sound; sensation like heavy truck striking building; standing motor cars rocked noticeably
Moderate (5)	Felt by nearly everyone; many awakened; some dishes, windows broken; unstable objects overturned; pendulum clocks may stop
Strong (6)	Felt by all, many frightened; some heavy furniture moved; a few instances of fallen plaster; damage slight
Very Strong (7)	Damage negligible in buildings of good design and construction; slight to moderate in well-built ordinary structures; considerable damage in poorly built or badly designed structures; some chimneys broken

[Skip Table]

Level of intensity	Description
Severe (8)	Damage slight in specially designed structures; considerable damage in ordinary substantial buildings with partial collapse; damage great in poorly built structures; fall of chimneys, factory stacks, columns, monuments, walls; heavy furniture overturned
Violent (9)	Damage considerable in specially designed structures; well-designed frame structures thrown out of plumb; damage great in substantial buildings, with partial collapse; buildings shifted off foundations
Extreme (10)	Some well-built wooden structures destroyed; most masonry and frame structures destroyed with foundations; rails bent
Extreme (11)	Few, if any (masonry), structures remain standing; bridges destroyed; broad fissures in ground; underground pipelines completely out of service; earth slumps and land slips in soft ground; rails bent greatly
Extreme (12)	Damage total; waves seen on ground surfaces; lines of sight and level distorted; objects thrown upward into the air

An example of this effect is the 1985 M8 earthquake that struck the Michoacán region of western Mexico, southwest of Mexico City. There was relatively little damage in the area around the epicentre, but there was tremendous damage and about 5,000 deaths in heavily populated Mexico City some 350 kilometers from the epicentre. The key reason for this is that Mexico City was built largely on the unconsolidated and water-saturated sediment of former Lake Texcoco. These sediments resonate at a frequency of about two seconds, which was similar to the frequency of the body waves that reached the city. For the same reason that a powerful opera singer can break a wine glass by singing the right note, the amplitude of the seismic waves was amplified by the lake sediments. Survivors of the disaster recounted that the ground in some areas moved up and down by about 20 centimeters every two seconds for over two minutes. Damage was greatest to buildings between 5 and 15 storeys tall, because they also resonated at around two seconds, which amplified the shaking.

Exercise 11.3 Estimating intensity from personal observations

The following observations were made by residents of the Nanaimo area during the M6.8 Nisqually earthquake near Olympia, Washington in 2001. Estimate the Mercalli intensities using Table 11.4.

Table 11.5

[Skip Table]

Building Type	Floor	Shaking Felt	How long it lasted (in seconds)	Description of Motion	Intensity?
House	1	no	10	Heard a large rumble lasting not even 10 seconds, mirror swayed	
House	2	moderate	60	Candles, pictures and CDs on bookshelf moved, towels fell off racks	
House	1	no		Pots hanging over stove moved and crashed together	
House	1	weak		Rolling feeling with a sudden stop, picture fell off mantle, chair moved	
Apartment	1	weak	10	Sounded like a big truck then everything shook for a short period	
House	1	moderate	20-30	Teacups rattled but didn't fall off	
Institution	2	moderate	15	Creaking sounds, swaying movement of shelving	
House	1	moderate	15-30	Bed banging against the wall with me in it, dog barking aggressively	

See Appendix 3 for Exercise 11.3 answers.

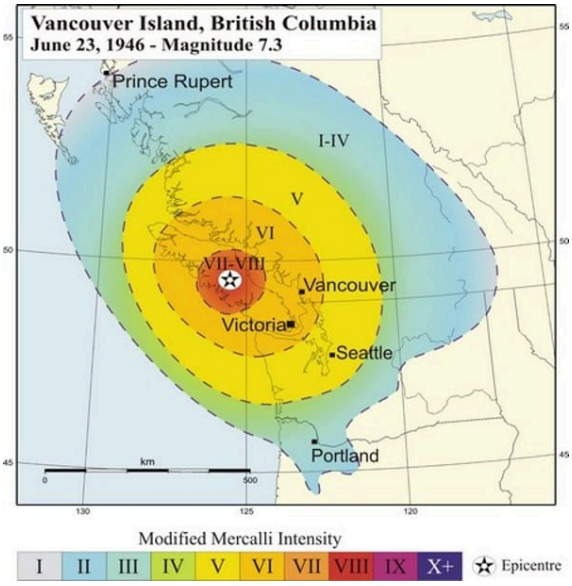


Figure 15.3.3.4 Intensity map for the 1946 M7.3 Vancouver Island earthquake. [Long Decsription]

An intensity map for the 1946 M7.3 Vancouver Island earthquake is shown in Figure 15.3.3.4. The intensity was greatest in the central island region where, in some communities, chimneys were damaged on more than 75% of buildings, some roads were made impassable, and a major rock slide occurred. The earthquake was felt as far north as Prince Rupert, as far south as Portland Oregon, and as far east as the Rockies.

Image Descriptions

Figure 15.3.3.1 **image description:** P-waves and S-waves from a small (M4) earthquake near Vancouver Island in 1997. The P-wave arrived in 0.7 seconds with an amplitude ranging from negative 0.7 millimeters per second to 1.1 millimeters per second and lasting until the arrival of the S-wave. The S-wave arrived at 8.7 seconds, with a minimum amplitude of negative 2.8 millimeters per second and a maximum amplitude of 2.7 millimeters per second. The S-wave's net amplitude gradually decreased over the next 5 seconds. [Return to Figure 15.3.3.1]

Figure 15.3.3.4 **image description:** The graduated intensity of the 1945 M7.3 Vancouver Island earthquake based on the modified Mercalli intensity scale. The area surrounding the epicentre of the earthquake which included central Vancouver Island ranged between a very strong (7) and severe (8) intensity. The next ring included the northern and southern parts of Vancouver Island, as well as a part of the main land coast including Vancouver and much of the Sunshine coast a strong (6) intensity. The next ring, which reached experienced a moderate (5) intensity, included Seattle and much of the BC interior. The outermost ring ranged between not felt (1) and light (4) intensity. It was felt as far north as Prince Rupert and the southern tip of Haida Gwaii, south eastern BC, and as far south as north western Oregon. [Return to Figure 15.3.3.4]

Media Attributions

- **Figure** 15.3.3.1**:** © Steven Earle. CC BY.
- **Figure** 15.3.3.2 **(left):** "Rayleigh Wave." Adapted by Steven Earle. Public domain.
- **Figure** 15.3.3.2 **(right):** "Love Wave" © Nicoguaro. Adapted by Steven Earle. CC BY.
- **Figure** 15.3.3.3**:** © Steven Earle. CC BY.
- **Figure** 15.3.3.4**:** "Vancouver Island, British Columbia June 23, 1946 – Magnitude 7.3," © National Resources Canada. Used under the terms allowing for non-commercial reproduction. This reproduction is a copy of an official work that is published by the Government of Canada. This reproduction has not been produced in affiliation with, or with the endorsement of the Government of Canada.

Text Attributions

- Table 11.4: The modified Mercalli intensity scale © Wikipedia. Adapted by Steven Earle. CC BY-SA.

---

1. Table 11.1 by Steven Earle. ↵

This page titled 15.3.3: Measuring Earthquakes is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Steven Earle (BCCampus) via source content that was edited to the style and standards of the LibreTexts platform.

- 11.3: Measuring Earthquakes by Steven Earle is licensed CC BY 4.0. Original source: https://opentextbc.ca/physicalgeology2ed.

15.3.4: The Impacts of Earthquakes

Some of the common impacts of earthquakes include structural damage to buildings, fires, damage to bridges, highways, pipelines and electrical transmission lines, initiation of slope failures, liquefaction, and tsunami. The types of impacts depend to a large degree on where the earthquake is located: whether it is predominantly urban or rural, densely or sparsely populated, highly developed or underdeveloped, and of course on the ability of the infrastructure to withstand shaking.

As we've seen from the example of the 1985 Mexico earthquake, the geological foundations on which structures are built can have a significant impact on earthquake shaking. When an earthquake happens, the seismic waves produced have a wide range of frequencies. The energy of the higher frequency waves tends to be absorbed by solid rock, while the lower frequency waves (with periods slower than one second) pass through the solid rock without being absorbed, but are eventually absorbed and amplified by soft sediments. It is therefore very common to see much worse earthquake damage in areas underlain by soft sediments than in areas of solid rock. A good example of this is in the Oakland area near San Francisco, where parts of a two-layer highway built on soft sediments collapsed during the 1989 Loma Prieta earthquake (Figure 15.3.4.1).



Figure 15.3.4.1 A part of the Cypress Freeway in Oakland California that collapsed during the 1989 Loma Prieta earthquake.

Building damage is also greatest in areas of soft sediments, and multi-storey buildings tend to be more seriously damaged than smaller ones. Buildings can be designed to withstand most earthquakes, and this practice is increasingly applied in earthquake-prone regions. Turkey is one such region, and even though Turkey had a relatively strong building code in the 1990s, adherence to the code was poor, as builders did whatever they could to save costs, including using inappropriate materials in concrete and reducing the amount of steel reinforcing. The result was that there were over 17,000 deaths in the 1999 M7.6 Izmit earthquake (Figure 15.3.4.2). After two devastating earthquakes that year, Turkish authorities strengthened the building code further, but the new code has been applied only in a few regions, and enforcement of the code is still weak, as revealed by the amount of damage from a M7.1 earthquake in eastern Turkey in 2011.



Figure 15.3.4.2 Buildings damaged by the 1999 earthquake in the Izmit area, Turkey.

Fires are commonly associated with earthquakes because fuel pipelines rupture and electrical lines are damaged when the ground shakes (Figure 15.3.4.3). Most of the damage in the great 1906 San Francisco earthquake was caused by massive fires in the downtown area of the city (Figure 15.3.4.4). Some 25,000 buildings were destroyed by those fires, which were fuelled by broken gas pipes. Fighting the fires was difficult because water mains had also ruptured. The risk of fires can be reduced through P-wave early warning systems if utility operators can reduce pipeline pressure and close electrical circuits.



Figure 15.3.4.3 Some of the effects of the 2011 Tohoku earthquake in the Sendai area of Japan. An oil refinery is on fire, and a vast area has been flooded by a tsunami.



Figure 15.3.4.4 Fires in San Francisco following the 1906 earthquake.



Figure 15.3.4.5 The Las Colinas debris flow at Santa Tecla (a suburb of the capital San Salvador) triggered by the January 2001 El Salvador earthquake. This is just one of many hundreds of slope failures that resulted from that earthquake. Over 500 people died in the area affected by this slide.

Earthquakes are important triggers for failures on slopes that are already weak. An example is the Las Colinas slide in the city of Santa Tecla, El Salvador, which was triggered by a M7.6 offshore earthquake in January 2001 (Figure 15.3.4.5).

Ground shaking during an earthquake can be enough to weaken rock and unconsolidated materials to the point of failure, but in many cases the shaking also contributes to a process known as **liquefaction**, in which an otherwise solid body of sediment is transformed into a liquid mass that can flow. When water-saturated sediments are shaken, the grains become rearranged to the point where they are no longer supporting one another. Instead, the water between the grains is holding them apart and the material can flow. Liquefaction can lead to the collapse of buildings and other structures that might be otherwise undamaged. A good example is the collapse of apartment buildings during the 1964 Niigata earthquake (M7.6) in Japan (Figure 15.3.4.6). Liquefaction can also contribute to slope failures and to fountains of sandy mud (sand volcanoes) in areas where there is loose saturated sand beneath a layer of more cohesive clay.



Figure 15.3.4.6 Collapsed apartment buildings in the Niigata area of Japan. The material beneath the buildings was liquefied to varying degrees by the 1964 earthquake.

Parts of the Fraser River delta are prone to liquefaction-related damage because the region is characterized by a 2 metre to 3 metre thick layer of fluvial silt and clay over top of at least 10 meters of water-saturated fluvial sand (Figure 15.3.4.7). Under these

conditions, it can be expected that seismic shaking will be amplified and that the sandy sediments will liquefy. This could lead to subsidence and tilting of buildings, and to failure and sliding of the silt and clay layer. Current building-code regulations in the Fraser delta area require that measures be taken to strengthen the ground underneath multi-storey buildings prior to construction.

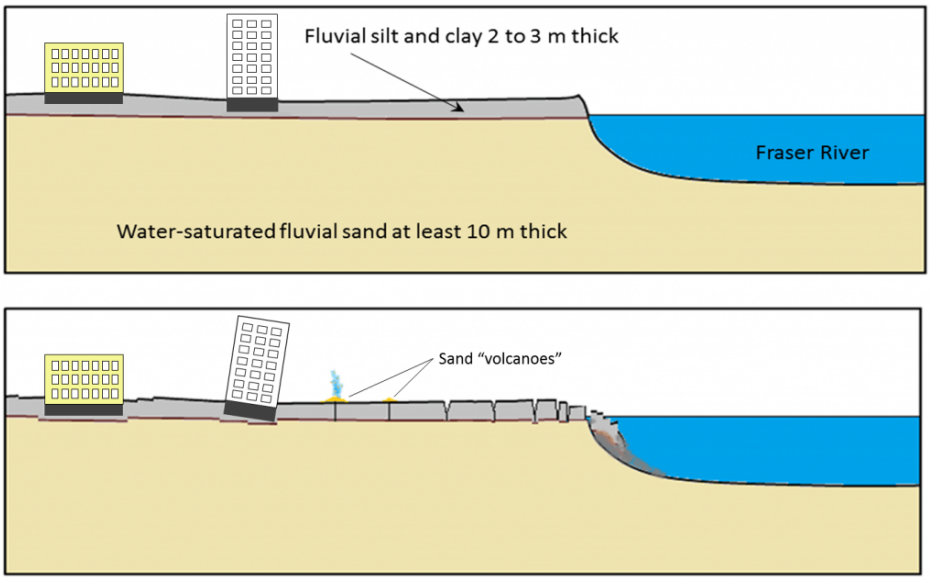


Figure 15.3.4.7 Recent unconsolidated sedimentary layers in the Fraser River delta area (top) and the potential consequences in the event of a damaging earthquake.

Exercise 11.4 Creating liquefaction and discovering the harmonic frequency

There are a few ways that you can demonstrate the process of liquefaction for yourself. The simplest is to go to a sandy beach (lake, ocean, or river) and find a place near the water's edge where the sand is wet. This is best done with your shoes off, so let's hope it's not too cold! While standing in one place on a wet part of the beach, start moving your feet up and down at a frequency of about once per second. Within a few seconds the previously firm sand will start to lose strength, and you'll gradually sink in up to your ankles.



Figure 15.3.4.8

If you can't get to a beach, or if the weather isn't cooperating, put some sand (sandbox sand will do) into a small container, saturate it with water, and then pour the excess water off. You can shake it gently to get the water to separate and then pour the excess water away, and you may have to do that more than once. Place a small rock on the surface of the sand; it should sit there

for hours without sinking in. Now, holding the container in one hand gently thump the side or the bottom with your other hand, about twice a second. The rock should gradually sink in as the sand around it becomes liquefied.

As you were moving your feet up and down or thumping the pot, it's likely that you soon discovered the most effective rate for getting the sand to liquefy; this would have been close to the natural harmonic frequency for that body of material. Stepping up and down as fast as you can (several times per second) on the wet beach would not have been effective, nor would you have achieved much by stepping once every several seconds. The body of sand vibrates most readily in response to shaking that is close to its natural harmonic frequency, and liquefaction is also most likely to occur at that frequency.

See Appendix 3 for Exercise 11.4 answers.

Earthquakes that take place beneath the ocean have the potential to generate **tsunami**. (Tsunami is the Japanese word for harbour wave. It is the same in both singular and plural.) The most likely situation for a significant tsunami is a large (M7 or greater) subduction-related earthquake. As shown in Figure 15.3.4.9, during the time between earthquakes the overriding plate becomes distorted by elastic deformation; it is squeezed laterally (Figure 15.3.4.9B) and pushed up.

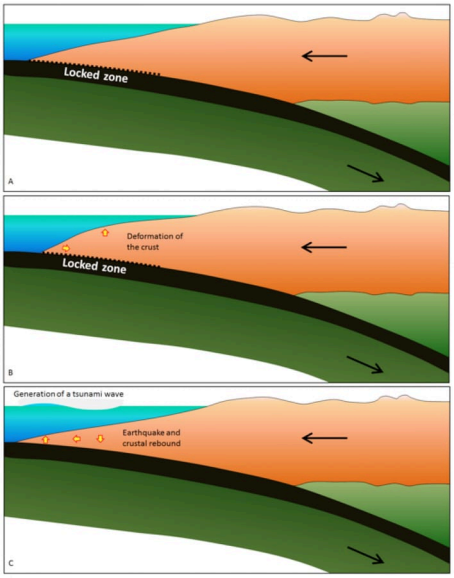


Figure 15.3.4.9 Elastic deformation and rebound of overriding plate at a subduction setting (B). The release of the locked zone during an earthquake (C) results in both uplift and subsidence on the sea floor, and this is transmitted to the water overhead, resulting in a tsunami.

When an earthquake happens (Figure 15.3.4.9C), the plate rebounds and there is both uplift and subsidence on the sea floor, in some cases by as much as several meters vertically over an area of thousands of square kilometers. This vertical motion is transmitted through the water column where it generates a series of waves that then spread across the ocean.

Subduction earthquakes with magnitude less than 7 do not typically generate significant tsunami because the amount of vertical displacement of the sea floor is minimal. Sea-floor transform earthquakes, even large ones (M7 to M8), don't typically generate tsunami either, because the motion is mostly side to side, not vertical.

Tsunami waves travel at velocities of several hundred kilometers per hour and easily make it to the far side of an ocean in about the same time as a passenger jet. The simulated one shown in Figure 15.3.4.10 is similar to that created by the 1700 Cascadia earthquake off the coast of British Columbia, Washington, and Oregon, which was recorded in Japan nine hours later.

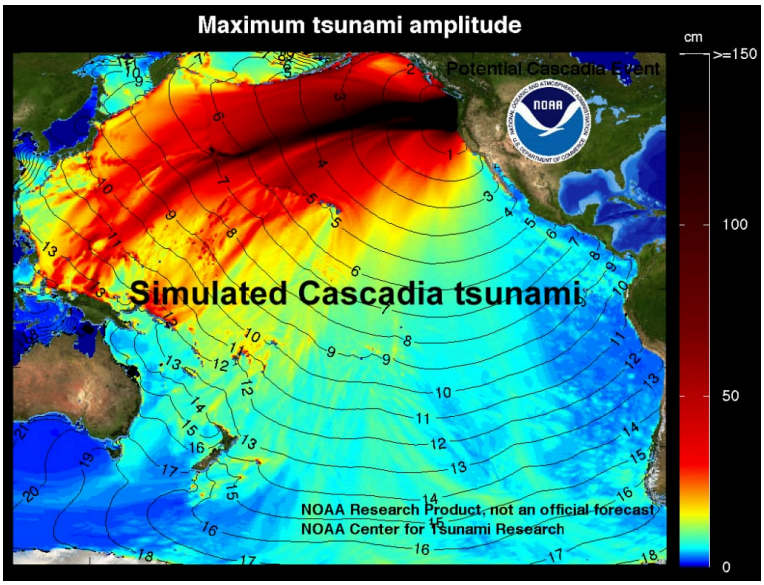


Figure 15.3.4.10 Model of the tsunami from the 1700 Cascadia earthquake (around M9) showing open-ocean wave heights (colors) and travel time contours. Tsunami wave amplitudes typically increase in shallow water.

Tsunami are discussed further in Chapter 17 under the topic of waves and coasts.

Media Attributions

- Figure 15.3.4.1: “Cypress Collapsed” by the USGS. Public domain.
- Figure 15.3.4.2: “Izmit, Turkey Earthquake” by the USGS. Public domain.
- Figure 15.3.4.3: “SH-60B helicopter flies over Sendai” by the United States Navy. Public domain.
- Figure 15.3.4.4: “San Francisco Fire 1906” from the Library of Congress. Public domain.
- Figure 15.3.4.5: “ElSalvadorslide” by the USGS. Public domain.
- Figure 15.3.4.6: “Liquefaction at Niigata.” Public domain.
- Figures 11.4.7, 11.4.8, 11.4.9: © Steven Earle. CC BY.
- Figure 15.3.4.10: “Maximum Tsunami Amplitude” from NOAA/PMEL/Center for Tsunami Research. Public domain.

This page titled 15.3.4: The Impacts of Earthquakes is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Steven Earle (BCCampus) via source content that was edited to the style and standards of the LibreTexts platform.

- 11.4: The Impacts of Earthquakes by Steven Earle is licensed CC BY 4.0. Original source: https://opentextbc.ca/physicalgeology2ed.

15.4: Volcanoes

This section explores volcanism and its relationship to plate tectonics. It examines how the composition of magma influences eruption style, introduces several major types of volcanoes, and considers where and why volcanic activity occurs. The section concludes by examining the hazards associated with volcanic eruptions and their effects on people and the environment.

This page titled 15.4: Volcanoes is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Robert Wagner.

15.4.1: Plate Tectonics and Volcanism

The relationships between plate tectonics and volcanism are shown on Figure 15.4.1.1. As summarized in Chapter 3, magma is formed at three main plate-tectonic settings: divergent boundaries (decompression melting), convergent boundaries (flux melting), and mantle plumes (decompression melting).

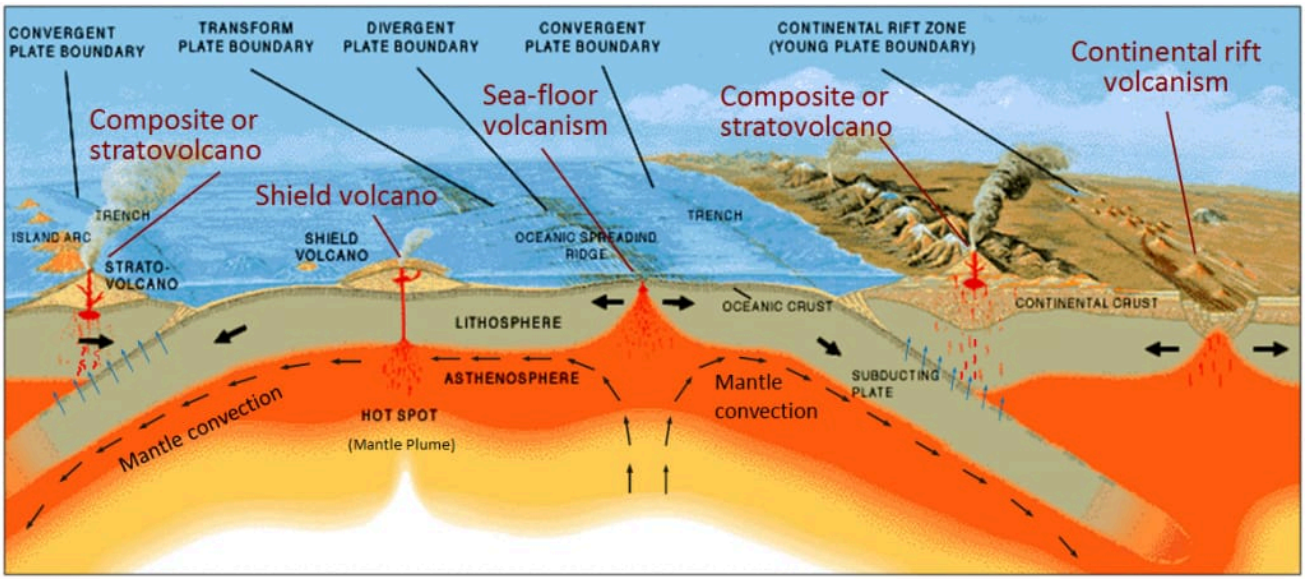


Figure 15.4.1.1 The plate-tectonic settings of common types of volcanism. Composite volcanoes form at subduction zones, either on ocean-ocean convergent boundaries (left) or ocean-continent convergent boundaries (right). Both shield volcanoes and cinder cones form in areas of continental rifting. Shield volcanoes form above mantle plumes, but can also form at other tectonic settings. Sea-floor volcanism can take place at divergent boundaries, mantle plumes and ocean-ocean-convergent boundaries.

The mantle and crustal processes that take place in areas of volcanism are illustrated in Figure 15.4.1.2. At a spreading ridge, hot mantle rock moves slowly upward by convection (centimetre/year), and within about 60 kilometers (km) of the surface, partial melting starts because of decompression. Over the triangular area shown in Figure 15.4.1.2a, about 10% of the ultramafic mantle rock melts, producing mafic magma that moves upward toward the axis of spreading (where the two plates are moving away from each other). The magma fills vertical fractures produced by the spreading and spills out onto the sea floor to form basaltic **pillows** (more on that later) and lava flows. There is spreading-ridge volcanism taking place about 200 km offshore from the west coast of Vancouver Island.

Exercise 4.1 How thick is the oceanic crust?

Figure 15.4.1.2a shows a triangular zone about 60 km thick; within this zone, approximately 10% of the mantle rock melts to form oceanic crust. Based on this information, approximately how thick do you think the resulting oceanic crust should be?

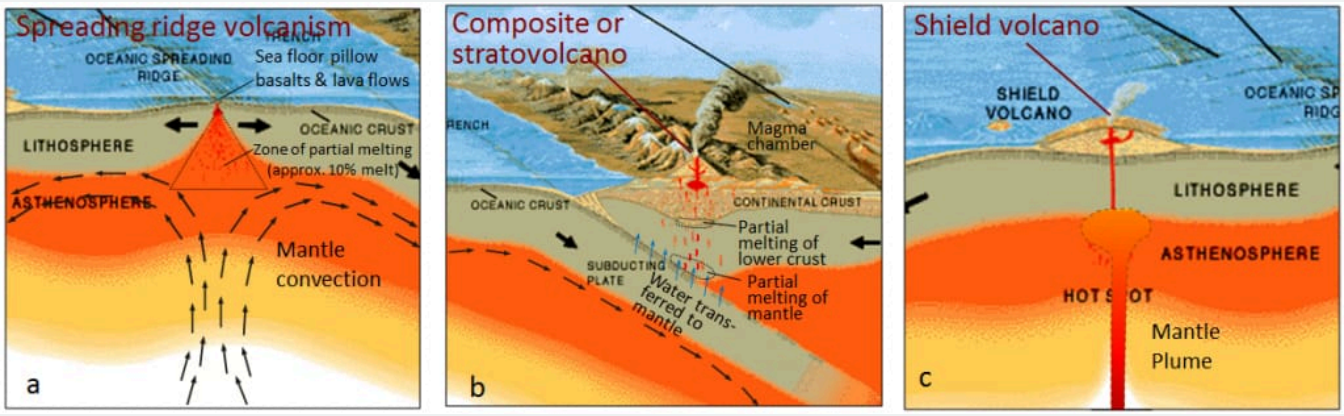


Figure 15.4.1.2 The processes that lead to volcanism in the three main volcanic settings on Earth: (a) volcanism related to plate divergence, (b) volcanism at an ocean-continent boundary (Similar processes take place at an ocean-ocean convergent boundary), and (c) volcanism related to a mantle plume.

See Appendix 3 for Exercise 4.1 answers.

At an ocean-continent convergent boundary, part of a plate that is made up of oceanic crust is subducting beneath part of another plate made up of continental crust. At an ocean-ocean convergent boundary, oceanic crust is being subducted beneath another oceanic-crust plate.[/footnote] (Figure 15.4.1.2b). In both situations the oceanic crust is heated up, and while there isn't enough heat to melt the subducting crust, there is enough heat to force the water out of some of its minerals. This released water rises into the overlying mantle where it contributes to flux melting of the mantle rock. The mafic magma produced rises through the mantle to the base of the crust. There it contributes to partial melting of crustal rock, and thus it assimilates much more felsic material. That magma, now likely intermediate in composition, continues to rise and assimilate crustal material. In the upper part of the crust, it accumulates into plutons. From time to time, the magma from the plutons rises toward surface, leading to volcanic eruptions. Mount Garibaldi (Figures 4.0.1 and 4.0.2) is an example of subduction-related volcanism.

A mantle plume is an ascending column of hot rock (not magma) that originates deep in the mantle, possibly just above the core-mantle boundary. Mantle plumes are thought to rise approximately 10 times faster than the rate of mantle convection. The ascending column may be on the order of kilometers to tens of kilometers across, but near the surface it spreads out to create a mushroom-style head that is several tens to over 100 km across. Near the base of the lithosphere (the rigid part of the mantle), the mantle plume (and possibly some of the surrounding mantle material) partially melts to form mafic magma that rises to feed volcanoes. Since most mantle plumes are beneath the oceans, the early stages of volcanism typically take place on the sea floor. Over time, islands may form like those in Hawaii.

Volcanism in northwestern B.C. (Figures 4.1.3 and 4.1.4) is related to continental rifting. This area is not at a divergent or convergent boundary, and there is no evidence of an underlying mantle plume. A likely explanation is that the crust of northwestern B.C. is being stressed by the northward movement of the Pacific Plate against the North America Plate, and the resulting crustal fracturing provides a conduit for the flow of magma from the mantle. This may, or may not, be an early stage of continental rifting, such as that found in eastern Africa.

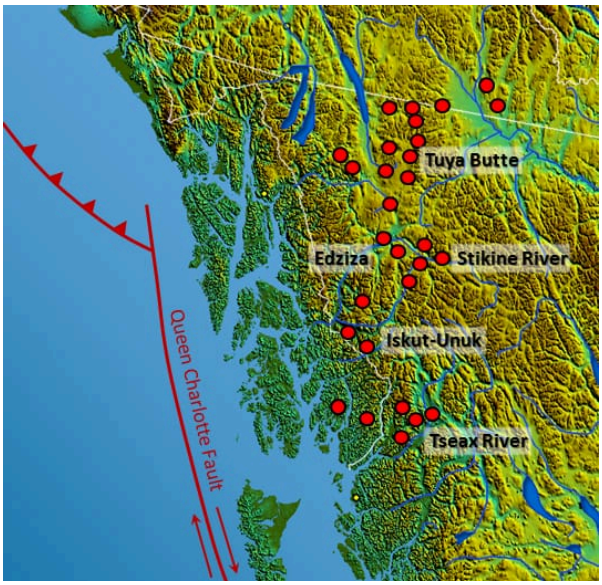


Figure 15.4.1.3 Volcanoes and volcanic fields in the Northern Cordillera Volcanic Province, B.C.



Figure 15.4.1.4 Volcanic rock at the Tseax River area, northwestern B.C.

Media Attributions

- Figure 15.4.1.1: Understanding Plate Motions by USGS. Public domain. Modified by Steven Earle.
- Figure 15.4.1.2: By USGS. Public domain. Modified by Steven Earle.
- Figure 15.4.1.3: "South-West Canada" by USGS. Public domain. Modified by Steven Earle. Volcanic locations from Edwards, B. & Russell, J. (2000). Distribution, nature, and origin of Neogene-Quaternary magmatism in the northern Cordilleran volcanic province, Canada. Geological Society of America Bulletin. pp. 1280-1293 [Steven Earle] Cordillera Volcanic Province, B.C.
- Figure 15.4.1.4: © Steven Earle. CC BY.

This page titled 15.4.1: Plate Tectonics and Volcanism is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Steven Earle (BCCampus) via source content that was edited to the style and standards of the LibreTexts platform.

- 4.1: Plate Tectonics and Volcanism by Steven Earle is licensed CC BY 4.0. Original source: https://opentextbc.ca/physicalgeology2ed.

15.4.2: Magma Composition and Eruption Style

As noted in the previous section, the types of magma produced in the various volcanic settings can differ significantly. At divergent boundaries and oceanic mantle plumes, where there is little interaction with crustal materials the magma tends to be consistently mafic. At subduction zones, where the magma ascends through significant thicknesses of crust, interaction between the magma and the crustal rock—some of which is quite felsic—leads to increases in the felsic character of the magma.

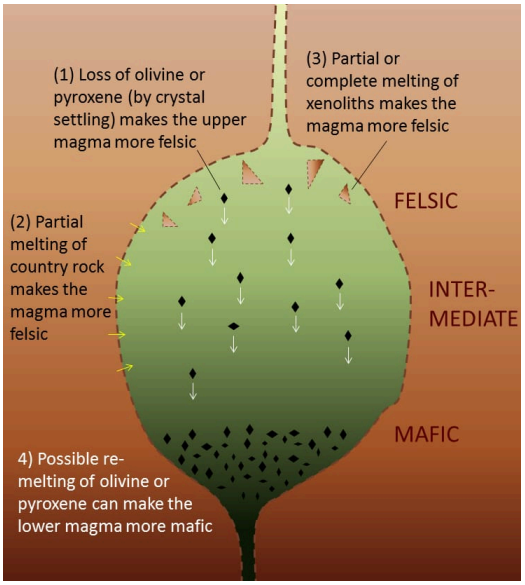


Figure 15.4.2.1 The important processes that lead to changes in the composition of magmas stored within magma chambers within relatively felsic rocks of the crust.

As illustrated in Figure 15.4.2.1, several processes can make magma that is stored in a chamber within the crust more felsic than it was to begin with, and can also contribute to development of vertical zonation from more mafic at the bottom to more felsic at the top. Partial melting of country rock and country-rock xenoliths increases the overall felsic character of the magma; first, because the country rocks tends to be more felsic than the magma, and second, because the more felsic components of the country rock melt preferentially. Settling of ferromagnesian crystals from the upper part of the magma, and possible remelting of those crystals in the lower part can both contribute to the vertical zonation from relatively mafic at the bottom to more felsic at the top.

From the perspective of volcanism there are some important differences between felsic and mafic magmas. First, as we've already discussed, felsic magmas tend to be more viscous because they have more silica, and hence more polymerization. Second, felsic magmas tend to have higher levels of volatiles; that is, components that behave as gases during volcanic eruptions. The most abundant volatile in magma is water (H_2O), followed typically by carbon dioxide (CO_2), and then by sulfur dioxide (SO_2).

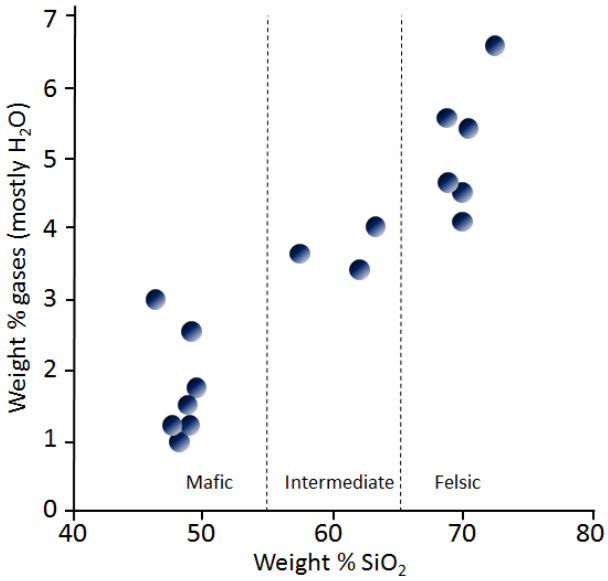


Figure 15.4.2.2 Variations in the volatile compositions of magmas as a function of silica content.

The general relationship between the SiO_2 content of magma and the amount of volatiles is shown in Figure 15.4.2.2. Although there are many exceptions to this trend, mafic magmas typically have 1% to 3% volatiles, intermediate magmas have 3% to 4% volatiles, and felsic magmas have 4% to 7% volatiles.

Differences in viscosity and volatile levels have significant implications for the nature of volcanic eruptions. When magma is deep beneath the surface and under high pressure from the surrounding rocks, the gases remain dissolved. As magma approaches the surface, the pressure exerted on it decreases. Gas bubbles start to form, and the more gas there is in the magma, the more bubbles form. If the magma is runny enough for gases to rise up through it and escape to surface, the pressure will not become excessive. Assuming that it can break through to the surface, the magma will flow out relatively gently. An eruption that involves a steady non-violent flow of magma is called **effusive**.

Exercise 4.2 Under pressure!



Figure 15.4.2.3

A good analogy for a magma chamber in the upper crust is a plastic bottle of pop on the supermarket shelf. Go to a supermarket and pick one up off the shelf (something not too dark). You'll find that the bottle is hard because it was bottled under pressure, and you should be able to see that there are no gas bubbles inside.

Buy a small bottle of pop (you don't have to drink it!) and open it. The bottle will become soft because the pressure is released, and small bubbles will start forming. If you put the lid back on and shake the bottle (best to do this outside!), you'll enhance the processes of bubble formation, and when you open the lid, the pop will come gushing out, just like an explosive volcanic eruption.

A pop bottle is a better analogue for a volcano than the old baking soda and vinegar experiment that you did in elementary school, because pop bottles—like volcanoes—come pre-charged with gas pressure. All we need to do is release the confining pressure and the gases come bubbling out, forcing the pop with them.

If the magma is felsic, and therefore too viscous for gases to escape easily, or if it has a particularly high gas content, it is likely to be under high pressure. Viscous magma doesn't flow easily, so even if there is a conduit for it to move towards surface, it may not flow out. Under these circumstances pressure will continue to build as more magma moves up from beneath and gases continue to exsolve. Eventually some part of the volcano will break and then all of that pent-up pressure will lead to an explosive eruption.

Mantle plume and spreading-ridge magmas tend to be consistently mafic, so effusive eruptions are the norm. At subduction zones, the average magma composition is likely to be close to intermediate, but as we've seen, magma chambers can become zoned and so compositions ranging from felsic to mafic are possible—even likely. Eruption styles can be correspondingly variable.

Image Descriptions

Figure 15.4.2.1 **image description:** Four processes that can affect the composition of magma stored in chambers.

1. Loss of olivine or pyroxene (by crystal settling) makes the upper magma more felsic.
2. Partial melting of country rock makes the magma more felsic.
3. Partial or complete melting of xenoliths makes the magma more felsic.
4. Possible re-melting of olivine or pyroxene can make the lower magma more mafic.

[Return to Figure 15.4.2.1]

Figure 15.4.2.2 *image description: The following table describes the range of data points adapted from the original scatter plot graph.*

Type of Magma	Weight % of gasses (mostly H_2O)	Weight % of SiO_2
Mafic	1 to 3%	47 to 50%
Intermediate	3 to 4%	57 to 64%
Felsic	4 to 7%	69 to 72%

[Return to Figure 15.4.2.2]

Media Attributions

- Figure 15.4.2.1: © Steven Earle. CC BY.
- Figure 15.4.2.2: Original image by Schminke, 2004. (Schminke, H-U., 2004, Volcanism, Springer-Verlag, Heidelberg). Modified by Steven Earle.
- Figure 15.4.2.3: Champagne uncorking photographed with a high speed air-gap flash © Niels Noordhoek. CC BY-SA.

This page titled 15.4.2: Magma Composition and Eruption Style is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Steven Earle (BCCampus) via source content that was edited to the style and standards of the LibreTexts platform.

- 4.2: Magma Composition and Eruption Style by Steven Earle is licensed CC BY 4.0. Original source: https://opentextbc.ca/physicalgeology2ed.

15.4.3: Types of Volcanoes

There are numerous types of volcanoes or volcanic sources; some of the more common ones are summarized in Table 4.1.

Table 4.1 A summary of the important types of volcanism

[Skip Table]

Type	Tectonic Setting	Size and Shape	Magma and Eruption Characteristics	Example
Cinder cone	Various; some form on the flanks of larger volcanoes	Small (10s to 100s of meters) and steep (Greater than 20°)	Most are mafic and form from the gas-rich early stages of a shield- or rift-associated eruption	Eve Cone, northern B.C.
Composite volcano	Almost all are at subduction zones	Medium size (1000s of meters high and up to 20 km across) and moderate steepness (10° to 30°)	Magma composition varies from felsic to mafic, and from explosive to effusive	Mount St. Helens
Shield volcano	Most are at mantle plumes; some are on spreading ridges	Large (up to several 1,000 meters high and up to 200 kilometers across), not steep (typically 2° to 10°)	Magma is almost always mafic, and eruptions are typically effusive, although cinder cones are common on the flanks of shield volcanoes	Kilauea, Hawaii
Large igneous provinces	Associated with “super” mantle plumes	Enormous (up to millions of square kiometers) and 100s of meters thick	Magma is always mafic and individual flows can be 10s of meters thick	Columbia River basalts
Sea-floor volcanism	Generally associated with spreading ridges but also with mantle plumes	Large areas of the sea floor associated with spreading ridges	Pillows form at typical eruption rates; lava flows develop if the rare of flow is faster	Juan de Fuca ridge
Kimberlite	Upper-mantle sourced	The remnants are typically 10s to 100s of meters across	Most appear to have had explosive eruptions forming cinder cones; the youngest one is dated at about 10 ka, and all others are at least 30 Ma	Lac de Gras Kimberlite Field, N.W.T.

The sizes and shapes of typical shield, composite, and cinder-cone volcanoes are compared in Figure 15.4.3.1, although, to be fair, Mauna Loa is the largest shield volcano on Earth; all others are smaller. Mauna Loa rises from the surrounding flat sea floor, and its diameter is in the order of 200 km. Its elevation is 4,169 m above sea level. Mount St. Helens, a composite volcano of average size, rises above the surrounding hills of the Cascade Range. Its diameter is about 6 km, and its height is 2,550 m above sea level. Cinder cones are much smaller. On this drawing, even a large cinder cone is just a dot.

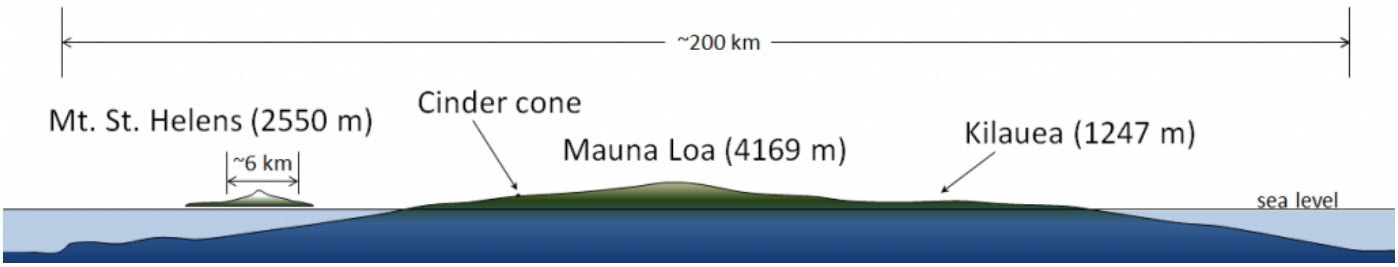


Figure 15.4.3.1 Profiles of Mauna Loa shield volcano, Mount St. Helens composite volcano, and a large cinder cone.

Cinder Cones

Cinder cones, like Eve Cone in northern B.C. (Figure 15.4.3.2), are typically only a few hundred meters in diameter, and few are

more than 200 m high. Most are made up of fragments of **vesicular** mafic rock (scoria) that were expelled as the magma boiled when it approached the surface, creating fire fountains. In many cases, these later became the sites of effusive lava flows when the gases were depleted. Most cinder cones are **monogenetic**, meaning that they formed during a single eruptive phase that might have lasted weeks or months. Because cinder cones are made up almost exclusively of loose fragments, they have very little strength. They can be easily, and relatively quickly, eroded away.

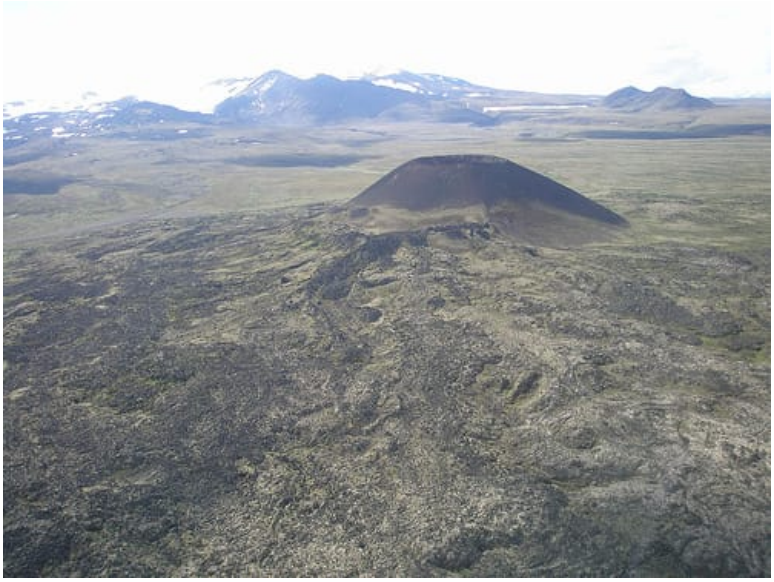


Figure 15.4.3.2 Eve Cone, situated near to Mount Edziza in northern B.C., formed approximately 700 years ago.

Composite Volcanoes

Composite volcanoes, like Mount St. Helens in Washington State (Figure 15.4.3.3), are almost all associated with subduction at convergent plate boundaries—either ocean-continent or ocean-ocean boundaries (Figure 15.4.3.2b). They can extend up to several thousand meters from the surrounding terrain, and, with slopes ranging up to 30˚ They can be up to about 20 km across. At many such volcanoes, magma is stored in a magma chamber in the upper part of the crust. For example, at Mount St. Helens, there is evidence of a magma chamber that is approximately 1 kilometre wide and extends from about 6 km to 14 km below the surface (Figure 15.4.3.4). Systematic variations in the composition of volcanism over the past several thousand years at Mount St. Helens imply that the magma chamber is zoned, from more felsic at the top to more mafic at the bottom.



Figure 15.4.3.3 The north side of Mount St. Helens in southwestern Washington State, 2003. The large 1980 eruption reduced the height of the volcano by 400 m, and a sector collapse removed a large part of the northern flank. Between 1980 and 1986 the slow eruption of more mafic and less viscous lava led to construction of a dome inside the crater.

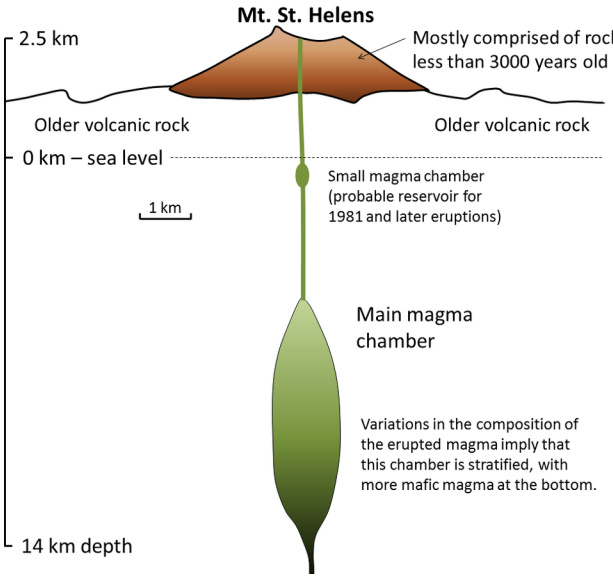


Figure 15.4.3.4 A cross-section through the upper part of the crust at Mount St. Helens showing the zoned magma chamber.

Mafic eruptions (and some intermediate eruptions), on the other hand, produce lava flows; the one shown in Figure 15.4.3.5b is thick enough (about 10 m in total) to have cooled in a **columnar jointing** pattern (Figure 15.4.3.7). Lava flows both flatten the profile of the volcano (because the lava typically flows farther than pyroclastic debris falls) and protect the fragmental deposits from erosion. Even so, composite volcanoes tend to erode quickly. Patrick Pringle, a volcanologist with the Washington State Department of Natural Resources, describes Mount St. Helens as a "pile of junk." The rock that makes up Mount St. Helens ranges in composition from rhyolite (Figure 15.4.3.5a) to basalt (Figure 15.4.3.5b); this implies that the types of past eruptions have varied widely in character. As already noted, felsic magma doesn't flow easily and doesn't allow gases to escape easily. Under these circumstances, pressure builds up until a conduit opens, and then an explosive eruption results from the gas-rich upper part of the magma chamber, producing **pyroclastic** debris, as shown on Figure 15.4.3.5a. This type of eruption can also lead to rapid melting of ice and snow on a volcano, which typically triggers large mudflows known as **lahars** (Figure 15.4.3.5a). Hot, fast-moving pyroclastic flows and lahars are the two main causes of casualties in volcanic eruptions. Pyroclastic flows killed approximately 30,000 people during the 1902 eruption of Mount. Pelée on the Caribbean island of Martinique. Most were incinerated in their homes. In 1985 a massive lahar, triggered by the eruption of Nevado del Ruiz, killed 23,000 people in the Colombian town of Armero, about 50 km from the volcano.

In a geological context, composite volcanoes tend to form relatively quickly and do not last very long. Mount St. Helens, for example, is made up of rock that is all younger than 40,000 years; most of it is younger than 3,000 years. If its volcanic activity ceases, it might erode away within a few tens of thousands of years. This is largely because of the presence of pyroclastic eruptive material, which is not strong.



Figure 15.4.3.5 Mount St. Helens volcanic deposits: (a) lahar deposits (L) and felsic pyroclastic deposits (P) and (b) a columnar basalt lava flow. The two photos were taken at locations only about 500 m apart.

Exercise 4.3 Volcanoes and Subduction

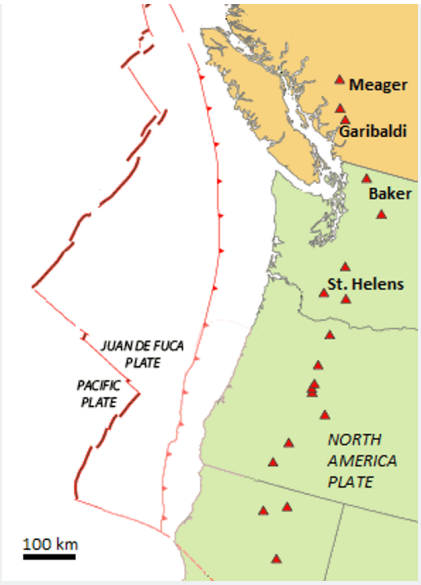


Figure 15.4.3.6

The map shown here illustrates the interactions between the North America, Juan de Fuca, and Pacific Plates off the west coast of Canada and the United States. The Juan de Fuca Plate is forming along the Juan de Fuca ridge, and is then subducted beneath the North America Plate along the red line with teeth on it ("Subduction boundary").

1. Using the scale bar in the lower left of the map, estimate the average distance between the subduction boundary and the Cascadia composite volcanoes.
2. If the subducting Juan de Fuca Plate descends 40 km for every 100 km that it moves inland, what is its likely depth in the area where volcanoes are forming?

See Appendix 3 for Exercise 4.3 answers.

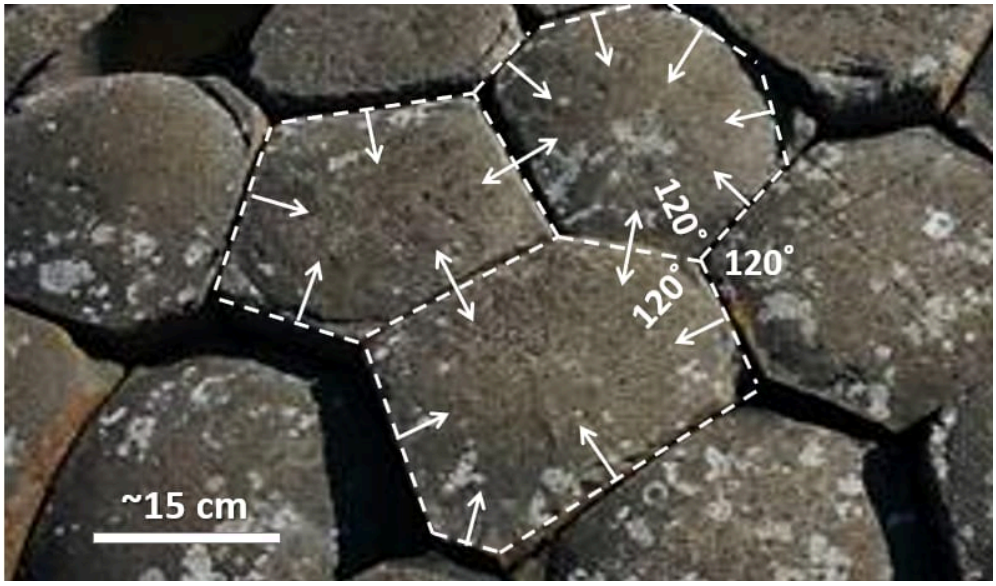


Figure 15.4.3.7 The development of columnar jointing in basalt, here seen from the top looking down. As the rock cools it shrinks, and because it is very homogenous it shrinks in a systematic way. When the rock breaks it does so with approximately 120° angles between the fracture planes. The resulting columns tend to be 6-sided but 5- and 7-sided columns also form.

Shield Volcanoes

Most **shield volcanoes** are associated with mantle plumes, although some form at divergent boundaries, either on land or on the sea floor. Because of their non-viscous mafic magma they tend to have relatively gentle slopes (2 to 10°) and the larger ones can be over 100 km in diameter. The best-known shield volcanoes are those that make up the Hawaiian Islands, and of these, the only active ones are on the big island of Hawaii. Mauna Loa, the world's largest volcano and the world's largest mountain (by volume) last erupted in 1984. Kilauea, arguably the world's most active volcano, has been erupting, virtually without interruption, since 1983. Loihi is an underwater volcano on the southeastern side of Hawaii. It is last known to have erupted in 1996, but may have erupted since then without being detected.

All of the Hawaiian volcanoes are related to the mantle plume that currently lies beneath Mauna Loa, Kilauea, and Loihi (Figure 15.4.3.8). In this area, the Pacific Plate is moving northwest at a rate of about 7 centimeters (cm) per year. This means that the earlier formed — and now extinct — volcanoes have now moved well away from the mantle plume. As shown on Figure 15.4.3.8 , there is evidence of crustal magma chambers beneath all three active Hawaiian volcanoes. At Kilauea, the magma chamber appears to be several kilometers in diameter, and is situated between 8 km and 11 km below surface.[1]

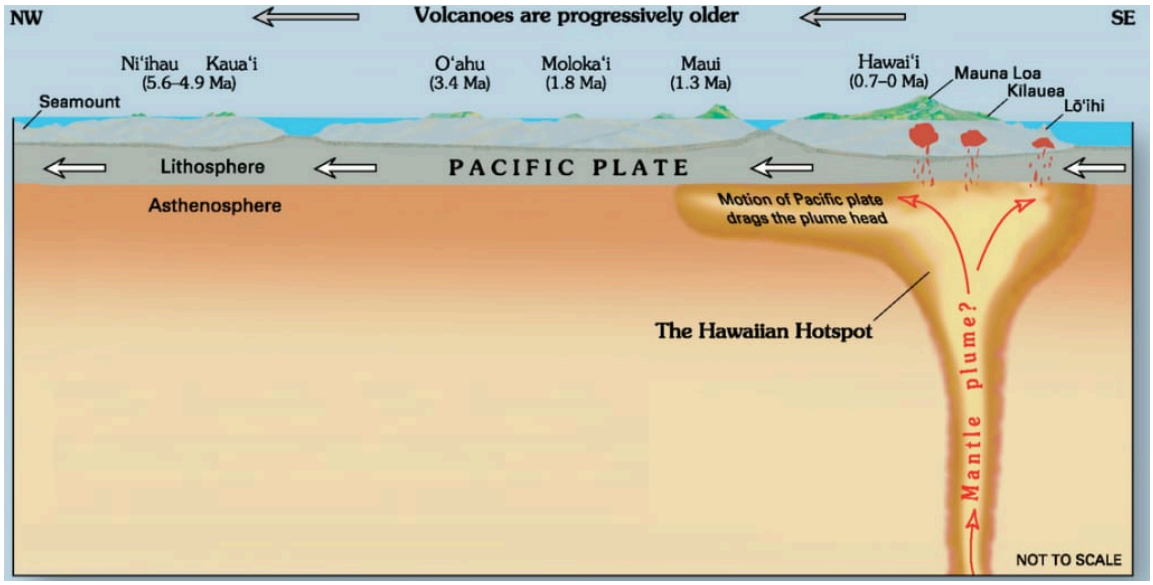


Figure 15.4.3.8 The mantle plume beneath the volcanoes of the island of Hawaii

Although it is not a prominent mountain (Figure 15.4.3.2), Kilauea volcano has a large **caldera** in its summit area (Figure 15.4.3.9). A caldera is a volcanic **crater** that is more than 2 km in diameter; this one is 4 km long and 3 km wide. It contains a smaller feature called Halema'uma'u crater, which has a total depth of over 200 m below the surrounding area. Most volcanic craters and calderas are formed above magma chambers, and the level of the crater floor is influenced by the amount of pressure exerted by the magma body. During historical times, the floors of both Kilauea caldera and Halema'uma'u crater have moved up during expansion of the magma chamber and down during deflation of the chamber.



Figure 15.4.3.9 Aerial view of the Kilauea caldera. The caldera is about 4 km across, and up to 120 m deep. It encloses a smaller and deeper crater known as Halema'uma'u.

One of the conspicuous features of Kilauea caldera is rising water vapor (the white cloud in Figure 15.4.3.9) and a strong smell of sulfur (Figure 15.4.3.10). As is typical in magmatic regions, water is the main volatile component, followed by carbon dioxide and sulfur dioxide. These, and some minor gases, originate from the magma chamber at depth and rise up through cracks in the overlying rock. This degassing of the magma is critical to the style of eruption at Kilauea, which, for most of the past 35 years, has been effusive, not explosive.



Figure 15.4.3.10 A gas-composition monitoring station (left) within the Kilauea caldera close to the edge of Halema'uma'u crater. The rising clouds are mostly composed of water vapor, but also include carbon dioxide and sulfur dioxide. Sulphur crystals (right) have formed around a gas vent in the caldera.

The Kilauea eruption that began in 1983 started with the formation of a cinder cone at Pu'u 'O'o, approximately 15 km east of the caldera (Figure 15.4.3.11). The magma feeding this eruption flowed along a major conduit system known as the East Rift, which extends for about 20 km from the caldera, first southeast and then east. Lava fountaining and construction of the Pu'u 'O'o cinder cone (Figure 15.4.3.12a) continued until 1986 at which time the flow became effusive. From 1986 to 2014, lava flowed from a gap in the southern flank of Pu'u 'O'o down the slope of Kilauea through a **lava tube** (Figure 15.4.3.12d), emerging at or near the ocean. During 2014 and 2015, the lava flowed northeast toward the community of Pahoa (see Exercise 4.4). In May of 2018 a new eruption started another 15 km east of the 2014/15 flow in the area known as Leilani Estates. The Lower East Rift Flow was active for 6 months. During that time, 35 km^2 of existing land was covered in lava and 3.5 km^2 of new land was created (Figure 15.4.3.11), about 48 km of road were covered in lava and 716 dwellings were destroyed (see USGS Overview of Kilauea Volcanoe's 2018 eruption[PDF]). Volcanic activity on the East Rift ceased in August 2018, and there has been no activity on Kilauea since then. This appears to mark the end of the eruption cycle that lasted—with only a few short interruptions—for 35 years. Kilauea will almost certainly erupt again within years or decades.

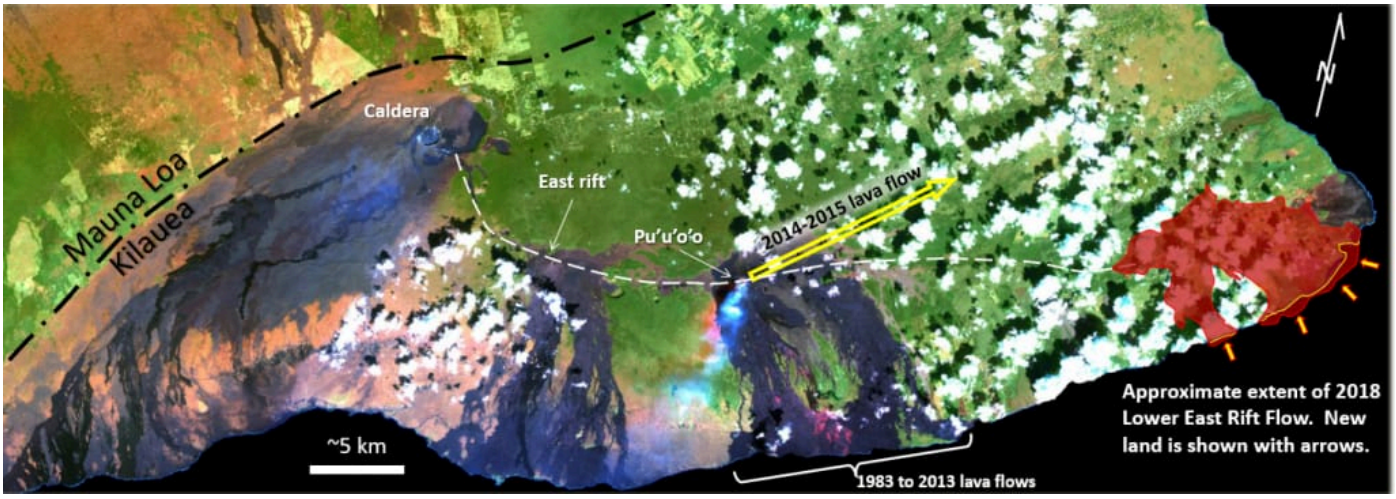


Figure 15.4.3.11 Satellite image of Kilauea volcano showing the East rift and Pu'u 'O'o, the site of the eruption that started in 1983 and continues to the present day. The 2014-2015 lava flow described in Exercise 4.4 is shown with a yellow arrow, and the area of the 2018 lava flow is shown in red. The puffy white blobs are clouds.

The two main types of textures created during effusive subaerial eruptions are pahoehoe and aa. **Pahoehoe**, ropy lava that forms as non-viscous lava, flows gently, forming a skin that gels and then wrinkles because of ongoing flow of the lava below the surface (Figure 15.4.3.12b, and "lava flow video"). **Aa**, or blocky lava, forms when magma is forced to flow faster than it is able to (down

a slope for example) (Figure 15.4.3.12c). **Tephra** (lava fragments) is produced during explosive eruptions, and accumulates in the vicinity of cinder cones.

Figure 15.4.3.12d is a view into an active lava tube on the southern edge of Kilauea. The red glow is from a stream of very hot lava (~1200°C) that has flowed underground for most of the 8 km from the Pu'u 'O'o vent. Lava tubes form naturally and readily on both shield and composite volcanoes because flowing mafic lava preferentially cools near its margins, forming solid **lava levées** that eventually close over the top of the flow. The magma within a lava tube is not exposed to the air, so it remains hot and fluid and can flow for tens of km, thus contributing to the large size and low slopes of shield volcanoes. The Hawaiian volcanoes are riddled with thousands of old lava tubes, some as long as 50 km.



Figure 15.4.3.12 Images of Kilauea volcano. (a) Pu'u'O'o cinder cone in the background with tephra in the foreground and aa lava in the middle, (b) Formation of pahoehoe on the southern edge of Kilauea, (c) Formation of aa on a steep slope on Kilauea, (d) Skylight in an active lava tube, Kilauea. Photos B & C taken in 2002 photos A & D taken in 2007.

Kilauea started forming at approximately 300 ka, while neighbouring Mauna Loa dates back to 700 ka and nearby Mauna Kea ito around 1 Ma. If volcanism continues above the Hawaii mantle plume in the same manner that it has since 85 Ma, it is likely that Kilauea will continue to erupt for at least another 500,000 years. By that time, its neighbour, Loihi, will have emerged from the sea floor, and its other neighbours, Mauna Loa and Mauna Kea, will have become significantly eroded, like their cousins, the islands to the northwest (Figure 15.4.3.8).

Exercise 4.4 Kilauea's 2014 lava flow

The U.S. Geological Survey Hawaii Volcano Observatory (HVO) map shown here, dated January 29, 2015, shows the outline of lava that started flowing northeast from Pu'u 'O'o on June 27, 2014 (the "June 27th Lava flow," a.k.a. the "East Rift Lava Flow"). The flow reached the nearest settlement, Pahoa, on October 29, 124 days later. After damaging some infrastructure west of Pahoa, the flow stopped advancing. A new outbreak occurred November 1, branching out to the north from the main flow.

What is the average rate of advance of the flow front from June 27 to October 29, 2014, in meters per day and meters per hour?

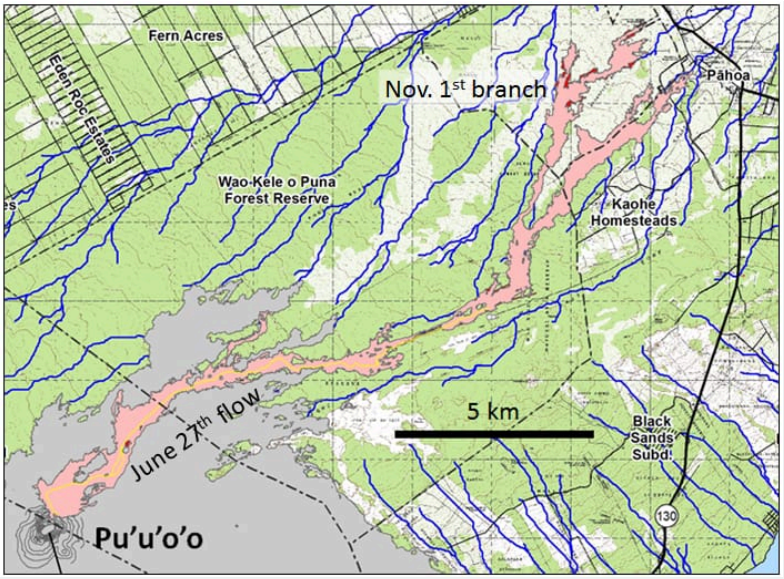


Figure 15.4.3.13

See Appendix 3 for Exercise 4.4 Answers.

Large Igneous Provinces

While the Hawaii mantle plume has produced a relatively low volume of magma for a very long time (~85 Ma), other mantle plumes are less consistent, and some generate massive volumes of magma over relatively short time periods. Although their origin is still controversial, it is thought that the volcanism leading to **large igneous provinces** (LIP) is related to very high volume but relatively short duration bursts of magma from mantle plumes. An example of an LIP is the Columbia River Basalt Group (CRGB), which extends across Washington, Oregon, and Idaho (Figure 15.4.3.14). This volcanism, which covered an area of about 160,000 square kilometers (km^2) with basaltic rock up to several hundred meters thick, took place between 17 and 14 Ma.

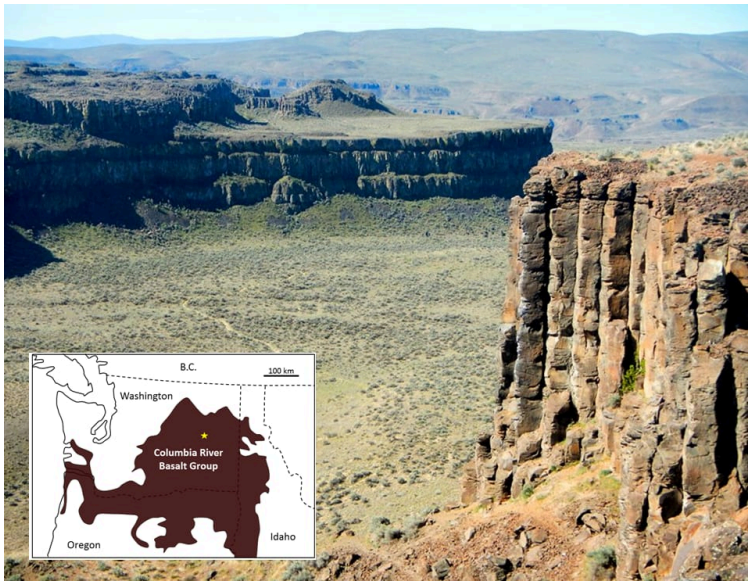


Figure 15.4.3.14 A part of the Columbia River Basalt Group at Frenchman Coulee, eastern Washington. All of the flows visible here have formed large (up to two m in diameter) columnar basalts, a result of relatively slow cooling of flows that are tens of meters thick. The inset map shows the approximate extent of the 17 to 14 Ma Columbia River Basalts, with the location of the photo shown as a star.

Most other LIP eruptions are much bigger. The Siberian Traps (also basalt), which erupted at the end of the Permian period at 250 Ma, are estimated to have produced approximately 40 times as much lava as the CRBG.

The mantle plume that is assumed to be responsible for the CRBG is now situated beneath the Yellowstone area, where it leads to felsic volcanism. Over the past 2 Ma three very large explosive eruptions at Yellowstone have yielded approximately 900 cubic kilometers (km^3) of felsic magma, about 900 times the volume of the 1980 eruption of Mount St. Helens, but only 5% of the volume of mafic magma in the CRBG.

Sea-Floor Volcanism

Some LIP eruptions occur on the sea floor, the largest known being the one that created the Ontong Java plateau in the western Pacific Ocean at around 122 Ma. But most sea-floor volcanism originates at divergent boundaries and involves relatively low-volume eruptions. Under these conditions, hot lava that oozes out into the cold seawater quickly cools on the outside and then behaves a little like toothpaste. The resulting blobs of lava are known as **pillows**, and they tend to form piles around a sea-floor lava vent (Figure 15.4.3.15). In terms of area, there is very likely more pillow basalt on the sea floor than any other type of rock on Earth.

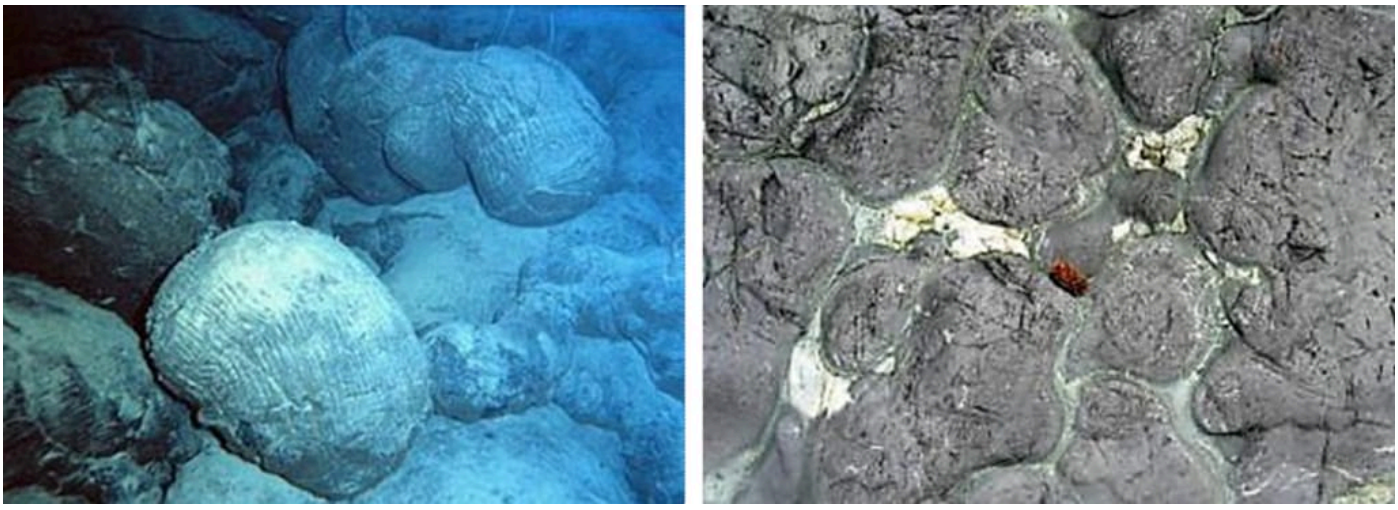


Figure 15.4.3.15 (Left) Modern sea-floor pillows in the south Pacific. (Right) Ancient sea-floor pillow basalts. Eroded 40 to 50 Ma pillows on the shore of Vancouver Island, near to Sooke. The pillows are 30 to 40 cm in diameter.

Kimberlites

While all of the volcanism discussed so far is thought to originate from partial melting in the upper mantle or within the crust, there is a special class of volcanoes called **kimberlites** that have their origins much deeper in the mantle, at depths of 150 km to 450 km. During a kimberlite eruption, material from this depth may make its way to surface quickly (hours to days) with little interaction with the surrounding rocks. As a result, kimberlite eruptive material is representative of mantle compositions: it is ultramafic.

Kimberlite eruptions that originate at depths greater than 200 km, within areas beneath old thick crust (**shields**), traverse the region of stability of diamond in the mantle, and in some cases, bring diamond-bearing material to the surface. All of the diamond deposits on Earth are assumed to have formed in this way; an example is the rich Ekati Mine in the Northwest Territories (Figure 15.4.3.16).



Figure 15.4.3.16 Ekati diamond mine, Northwest Territories, part of the Lac de Gras kimberlite field

The kimberlites at Ekati erupted between 45 and 60 Ma. Many kimberlites are older, some much older. There have been no kimberlite eruptions in historic times. The youngest known kimberlites are in the Igwisi Hills in Tanzania and are only about 10,000 years old. The next youngest known are dated to about 30 Ma.

The Smithsonian Institution maintains a comprehensive catalogue of the world's volcanoes, with information and eruptive history for nearly 2700 volcanic sites. If you spend some time looking around that site you'll discover the frequency of eruptions at different volcanoes is enormously variable, although we can make some generalizations. Focusing just on shield volcanoes and composite volcanoes some of the data are as follows:

Table 4.2 Eruptions of composite and shield volcanoes

Composite volcanoes	**Shield volcanoes**
Avachinsky (Russia): 5 eruptions over the past 7000 years	Fernandina (Galapagos): 31 eruptions over the past 1000 years
Pinatubo (Philippines): 4 eruptions over the past 9000 years	Kilauea (Hawaii): 62 eruptions over the past 250 years
Adams (Oregon, USA): 6 eruptions over the past 7000 years	Nyamuragira (Congo): 48 eruptions over the past 154 years

Based only on these numbers it is evident that, in general, shield volcanoes are much more active than composite volcanoes, but there are many exceptions to this trend. Some composite volcanoes are nearly as active as the shield volcanoes listed here, and some shield volcanoes that are still considered to be "active" are almost as inactive as the composite volcanoes listed here.

Image Descriptions

Figure 15.4.3.4 image description: Mount St. Helens rises over 2.5 kilometers above sea lever and consists mostly of rock less than 3,000 years old. Underneath the mountain is older volcanic rock. Just below sea level is a small magma chamber, which is a probable reservoir for 1981 and later eruptions. Down 5 to 14 kilometers below sea level is the main magma chamber. Variations in the composition of the erupted magma imply this chamber is stratified, with more magma at the bottom. [Return to Figure 15.4.3.4]

Figure 15.4.3.5 image description: Image (A) shows a cliff wall with grey/brown and orange horizontal layers. The sides look soft like they would be easily worn away. The grey/brown layers are lahar deposits and the orange layers are felsic pyroclastic deposits.

Image (B) shows a columnar basalt lava flow which looks like a rocky, stone cliff with vertical layers. [Return to Figure 15.4.3.5]

Figure 15.4.3.13 **image description:** The U.S. Geological Survey Hawaii Volcano Observatory (HVO) map, dated January 29, 2015, shows the outline of lava that started flowing northeast from Pu'u 'O'o on June 27, 2004 (the "June 27th Lava flow," a.k.a. the "East Rift Lava Flow"). The flow reached the nearest settlement, Pahoa, on October 29, after covering a distance of 20 km in 124 days. After damaging some infrastructure west of Pahoa, the flow stopped advancing. A new outbreak occurred November 1, branching out to the north from the main flow about 6 km southwest of Pahoa. [Return to Figure 15.4.3.13]

Figure 15.4.3.14 **image description:** The Columbia River Basalt Group covers most of south eastern Washington state and stretches along the borders between Washington, Idaho, and Oregon. The columnar basalts shown in the photo are in in eastern Washington. They rise up out of a flat valley as tall cliffs. [Return to Figure 15.4.3.14]

Media Attributions

- Figure 15.4.3.1: © Steven Earle. CC BY.
- Figure 15.4.3.2: Eve Cone © nass5518. CC BY.
- Figure 15.4.3.3: © Steven Earle. CC BY.
- Figure 15.4.3.4: Original image © Pringle, 1993. Modified by Steve Earle.
- Figure 15.4.3.5: © Steven Earle. CC BY.
- Figure 15.4.3.6: © Steven Earle. CC BY.

- Figure 15.4.3.7: © Steven Earle. CC BY.

- Figure 15.4.3.8: "Hawaii hotspot cross-sectional diagram" by USGS. Public domain.
- Figure 15.4.3.9: "Kilauea ali 2012 01 28" by NASA. Public domain.
- Figure 15.4.3.10: © Steven Earle. CC BY.
- Figure 15.4.3.11: Island of Hawai'i – Landsat mosaic by NOAA. Public domain. Modified by Steven Earle.
- Figure 15.4.3.12: © Steven Earle. CC BY.
- Figure 15.4.3.13: Image from USGS. Public domain.
- Figure 15.4.3.14: © Steven Earle. CC BY.
- Figure 15.4.3.15 (Left): Pillow Basalt Crop by NOAA. Public domain.
- Figure 15.4.3.15 (Right): © Steven Earle. CC BY.
- Figure 15.4.3.16: "Ekati mine" © Jason Pineau. CC BY-SA.

---

1. Lin, G, Amelung, F, Lavallee, Y, and Okubo, P, 2014, Seismic evidence for a crustal magma reservoir beneath the upper east rift zone of Kilauea volcano, Hawaii. Geology. V. ↵

This page titled 15.4.3: Types of Volcanoes is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Steven Earle (BCCampus) via source content that was edited to the style and standards of the LibreTexts platform.

- 4.3: Types of Volcanoes by Steven Earle is licensed CC BY 4.0. Original source: https://opentextbc.ca/physicalgeology2ed.

15.4.4: Volcanic Hazards

There are two classes of volcanic hazards, direct and indirect. Direct hazards are forces that directly kill or injure people, or destroy property or wildlife habitat. Indirect hazards are volcanism-induced environmental changes that lead to distress, famine, or habitat degradation. It is estimated that indirect effects of volcanism have accounted for approximately 8 million deaths during historical times, while direct effects have accounted for fewer than 200,000, or 2.5% of the total. Some of the more important types of volcanic hazards are summarized in Table 4.3.

Table 4.3 A summary of the important volcanic hazards

[Skip Table]

Type	Description	Risk
Tephra emissions	Small particles of volcanic rock emitted into the atmosphere	• Respiration problems for some individuals • Significant climate cooling leading to crop failure and famine • Damage to aircraft
Gas emissions	The emission of gases before, during, and after an eruption	• Climate cooling leading to crop failure and famine • In some cases, widespread poisoning
Pyroclastic density current	A very hot (several 100°C) mixture of gases and volcanic fragments (tephra) that flows rapidly (up to 100s of kilometers per hour (km/h)) down the side of a volcano	Extreme hazard — destroys anything in the way
Pyroclastic fall	Vertical fall of tephra in the area surrounding an eruption	• Thick tephra coverage of areas close to the eruption (1 km to 10s of kilometers) • Collapsed roofs
Lahar	A flow of mud and debris down a channel leading away from a volcano, triggered either by an eruption or a severe rain event	Severe risk of destruction for anything within the channel—lahar mud flows can move at 10s of km/h
Sector collapse/ debris avalanche	The failure of part of a volcano, either due to an eruption or for some other reason, leading to the failure of a large portion of the volcano	Severe risk of destruction for anything in the path of the debris avalanche
Lava flow	The flow of lava away from a volcanic vent	People and infrastructure at risk, but lava flows tend to be slow (less than km/h) and are relatively easy to avoid

Volcanic Gas and Tephra Emissions

Large volumes of **tephra** (rock fragments, mostly **pumice**) and gases are emitted during major **plinian eruptions** (large explosive eruptions with hot gas and tephra columns extending into the stratosphere) at composite volcanoes, and a large volume of gas is released during some very high-volume effusive eruptions. One of the major effects is cooling of the climate by 1° to 2°C for several months to a few years because the dust particles and tiny droplets and particles of sulfur compounds block the sun. The last significant event of this type was in 1991 and 1992 following the large eruption of Mount Pinatubo in the Philippines. A

temperature decrease of 1° to 2°C may not seem like very much, but that is the global average amount of cooling, and cooling was much more severe in some regions and at some times.

Over an eight-month period in 1783 and 1784, a massive effusive eruption took place at the Laki volcano in Iceland. Although there was relatively little volcanic ash involved, a massive amount of sulfur dioxide was released into the atmosphere, along with a significant volume of hydrofluoric acid (HF). The sulphate **aerosols** that formed in the atmosphere led to dramatic cooling in the northern hemisphere. There were serious crop failures in Europe and North America, and a total of 6 million people are estimated to have died from famine and respiratory complications. In Iceland, poisoning from the HF resulted in the death of 80% of sheep, 50% of cattle, and the ensuing famine, along with HF poisoning, resulted in more than 10,000 human deaths—about 25% of the population.

Volcanic ash can also have serious implications for aircraft because it can destroy jet engines. For example, over 5 million airline passengers had their travel disrupted by the 2010 Eyjafjallajökull volcanic eruption in Iceland.

Pyroclastic Density Currents

In a typical explosive eruption at a composite volcano, the tephra and gases are ejected with explosive force and are hot enough to be forced high up into the atmosphere. As the eruption proceeds, and the amount of gas in the rising magma starts to decrease, parts will become heavier than air, and they can then flow downward along the flanks of the volcano (Figure 15.4.4.1). As they descend, they cool more and flow faster, reaching speeds up to several hundred kilometers per hour. A **pyroclastic density current** (PDC) consists of tephra ranging in size from boulders to microscopic shards of glass (made up of the edges and junctions of the bubbles of shattered pumice), plus gases (dominated by water vapor, but also including other gases). The temperature of this material can be as high as 1000°C. Among the most famous PDCs are the one that destroyed Pompeii in the year 79 CE, killing an estimated 18,000 people, and the one that destroyed the town of St. Pierre, Martinique, in 1902, killing an estimated 30,000.

The buoyant upper parts of pyroclastic density currents can flow over water, in some cases for several kilometers. The 1902 St. Pierre PDC flowed out into the city's harbour and destroyed several wooden ships anchored there.



Figure 15.4.4.1 The plinian eruption of Mount Mayon, Philippines. in 1984. Although most of the eruption column is ascending into the atmosphere, there are pyroclastic density currents flowing down the sides of the volcano in several places.

Pyroclastic Fall

Most of the tephra from an explosive eruption ascends high into the atmosphere, and some of it is distributed around Earth by high-altitude winds. The larger components (larger than 0.1 mm) tend to fall relatively close to the volcano, and the amount produced by large eruptions can cause serious damage and casualties. The large 1991 eruption of Mount Pinatubo in the Philippines resulted in the accumulation of tens of centimeters of ash in fields and on rooftops in the surrounding populated

region. Heavy typhoon rains that hit the island at the same time added to the weight of the tephra, leading to the collapse of thousands of roofs and to at least 300 of the 700 deaths attributed to the eruption.

Lahar

A **lahar** is any mudflow or **debris flow** that is related to a volcano. Most are caused by melting snow and ice during an eruption, as was the case with the lahar that destroyed the Colombian town of Armero in 1985 (described earlier). Lahars can also happen when there is no volcanic eruption, and one of the reasons is that, as we've seen, composite volcanoes tend to be weak and easily eroded.

In October 1998, category 5 hurricane Mitch slammed into the coast of central America. Damage was extensive and 19,000 people died, not so much because of high winds but because of intense rainfall—some regions received almost 2 m of rain over a few days! Mudflows and debris flows occurred in many areas, especially in Honduras and Nicaragua. An example is at the Casita Volcano in Nicaragua, where the heavy rains weakened rock and volcanic debris on the upper slopes, resulting in a debris flow that rapidly built in volume as it raced down the steep slope, and then ripped through the towns of El Porvenir and Rolando Rodriguez killing more than 2,000 people (Figure 15.4.4.2). El Porvenir and Rolando Rodriguez were new towns that had been built without planning approval in an area that was known to be at risk of lahars.



Figure 15.4.4.2 Part of the path of the lahar from Casita Volcano, October 30, 1998.

Sector Collapse and Debris Avalanche

In the context of volcanoes, **sector collapse** or flank collapse is the catastrophic failure of a significant part of an existing volcano, creating a large debris avalanche. This hazard was first recognized with the failure of the north side of Mount St. Helens immediately prior to the large eruption on May 18, 1980. In the weeks before the eruption a large bulge had formed on the side of the volcano, the result of magma transfer from depth into a satellite magma body within the mountain itself. Early on the morning of May 18, a moderate earthquake struck nearby; this is thought to have destabilized the bulge, leading to Earth's largest ever observed slope failure. The failure of this part of the volcano exposed the underlying satellite magma chamber, causing it to explode sideways, which then exposed the conduit leading to the magma chamber below. The resulting plinian eruption—with a 24 kilometre high eruption column—lasted for nine hours.

In August 2010, a massive part of the flank of B.C.'s Mount Meager gave way and about 48 million cubic meters (m^3) of rock rushed down the valley, one of the largest slope failures in Canada in historical times (Figure 15.4.4.3). More than 25 slope failures have taken place at Mount Meager in the past 8,000 years, some of them more than 10 times larger than the 2010 failure.



Figure 15.4.4.3 The August 2010 Mount Meager rock avalanche, showing where the slide originated (arrow, 4 km upstream), its path down a steep narrow valley, and the debris field (and the stream that eventually cut through it) in the foreground.

Lava Flows

As we saw in Exercise 4.4, lava flows at volcanoes like Kilauea do not advance very quickly, and in most cases, people can get out of the way. Of course, it is more difficult to move infrastructure, and so buildings and roads are typically the main casualties of lava flows.

Exercise 4.5 Volcanic hazards in Squamish

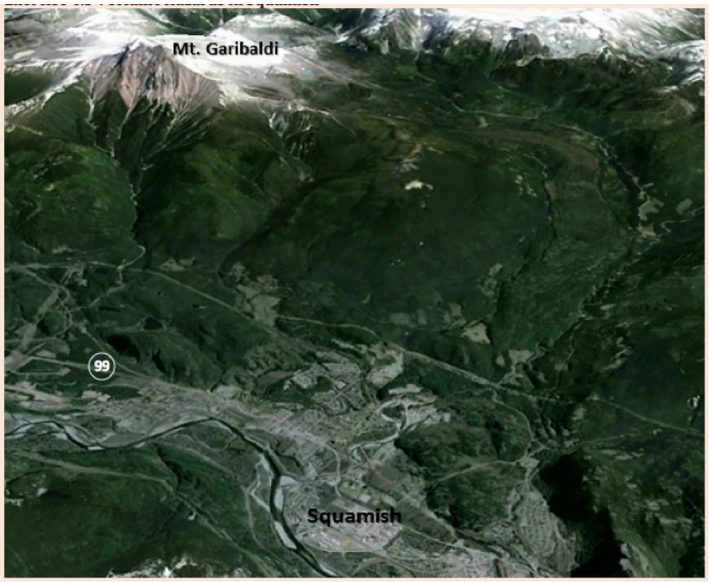


Figure 15.4.4.4

The town of Squamish is situated approximately 10 km from Mount Garibaldi, as shown in the photo. If Mount Garibaldi were to erupt, which of the following hazards could be an issue for people in and around Squamish? Explain why or why not.

1. Tephra emission.
2. Gas emission.
3. Pyroclastic denisty current.

4. Pyroclastic fall.
5. Lahar.
6. Sector collapse.
7. Lava flow.

See Appendix 3 for Exercise 4.5 answers.

Media Attributions

- Figure 15.4.4.1: "Pyroclastic flows at Mayon Volcano" by USGS. Public domain.
- Figure 15.4.4.2: "Casita Volcano" by USGS. Public domain.
- Figure 15.4.4.3: August 2010 Mt. Meager landslide, BC © Mika McKinnon. All rights reserved. Used with permission.
- Figure 15.4.4.4: Screenshot from Google Earth. © Google. Modified by Steve Earle. Used with permission.

This page titled 15.4.4: Volcanic Hazards is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Steven Earle (BCCampus) via source content that was edited to the style and standards of the LibreTexts platform.

- 4.4: Volcanic Hazards by Steven Earle is licensed CC BY 4.0. Original source: https://opentextbc.ca/physicalgeology2ed.

16: WEATHERING AND SOILS

16.1: Mechanical Weathering
16.2: Chemical Weathering
16.3: The Products of Weathering and Erosion
16.4: Weathering and the Formation of Soil

Chapter 16: Weathering and Soils - Introduction

Earth's surface is constantly being changed. While the previous chapter examined processes driven largely by energy from within Earth, such as plate tectonics, earthquakes, and volcanism, this chapter turns to processes operating at and near the surface. Once rocks are exposed to the atmosphere, water, and living organisms, they begin to break down and change.

Weathering is the breakdown or alteration of rock at or near Earth's surface. Mechanical weathering physically breaks rock into smaller pieces without changing the minerals themselves, while chemical weathering changes minerals through chemical reactions. These processes often work together: breaking a rock into smaller pieces exposes additional surface area where chemical reactions can occur.

The products of weathering do not necessarily remain where they form. **Erosion** removes and transports weathered material through processes involving water, wind, ice, and gravity. Over time, weathering and erosion help reshape landscapes while also supplying the sediment that becomes incorporated into other parts of the geological system.

Weathering also plays an essential role in the development of **soil**. Soil is more than broken rock; it develops through interactions among mineral material, organic matter, water, air, and living organisms. Its characteristics reflect the rock from which it develops as well as climate, topography, biological activity, and the amount of time available for soil formation. By examining weathering and soils, we can see another way in which the atmosphere, hydrosphere, biosphere, and geosphere continually interact at Earth's surface.

This page titled 16: Weathering and Soils is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Robert Wagner.

16.1: Mechanical Weathering

Intrusive igneous rocks form at depths of several hundreds of meters to several tens of kilometers. Sediments are turned into sedimentary rocks only when they are buried by other sediments to depths in excess of several hundreds of meters. Most metamorphic rocks are formed at depths of kilometers to tens of kilometers. Weathering cannot even begin until these rocks are uplifted through various processes of mountain building—most of which are related to plate tectonics—and the overlying material has been eroded away and the rock is exposed as an **outcrop**.[1]

The most important agents of mechanical weathering are:

- The decrease in pressure that results from removal of overlying rock
- Freezing and thawing of water in cracks in the rock
- Formation of salt crystals within the rock
- Cracking from plant roots and removal of material by burrowing animals

When a mass of rock is exposed by weathering and removal of the overlying rock, there is a decrease in the confining pressure on the rock, and the rock expands. This unloading promotes cracking of the rock, known as **exfoliation**, as shown in the granitic rock in Figure 16.1.1., which, in places, is peeling off like the layers of an onion.



Figure 16.1.1 Exfoliation fractures in granitic rock exposed on the side of the Coquihalla Highway north of Hope, B.C.



Figure 16.1.2 Exfoliation of slate at a road cut in the Columbia Mountains west of Golden, B.C.

Granitic rock tends to exfoliate parallel to the exposed surface because the rock is typically homogenous, and it doesn't have predetermined planes along which it must fracture. Sedimentary and metamorphic rocks, on the other hand, tend to exfoliate along predetermined planes (Figure 16.1.2).

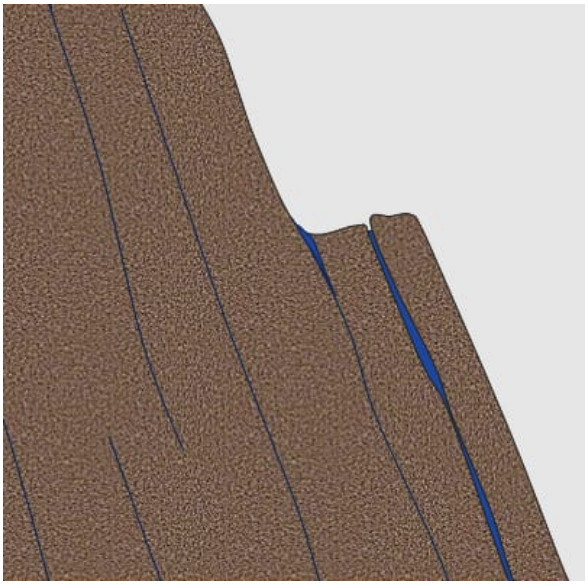


Figure 16.1.3 The process of frost wedging on a steep slope. Water gets into fractures and then freezes, expanding the fracture a little. When the water thaws it seeps a little farther into the expanded crack. The process is repeated many times, and eventually a piece of rock will be wedged away.

Frost wedging is the process by which water seeps into cracks in a rock, expands on freezing, and thus enlarges the cracks (Figure 16.1.3). The effectiveness of frost wedging is related to the frequency of freezing and thawing. Frost wedging is most effective in a climate like Canada's. In warm areas where freezing is infrequent, in very cold areas where thawing is infrequent, or in very dry areas, where there is little water to seep into cracks, the role of frost wedging is limited. If you are ever hiking in the mountains you might hear the effects of frost wedging when the Sun warms a steep rocky slope and the fragments of rock that were pried away from the surface by freezing the night before are released as that ice melts.

In many parts of Canada, the transition between freezing nighttime temperatures and thawing daytime temperatures is frequent — tens to hundreds of times a year. Even in warm coastal areas of southern B.C., freezing and thawing transitions are common at higher elevations. A common feature in areas of effective frost wedging is a **talus slope**—a fan-shaped deposit of fragments removed by frost wedging from the steep rocky slopes above (Figure 16.1.4).



Figure 16.1.4 An area with very effective frost-wedging near Keremeos, B.C. The fragments that have been wedged away from the cliffs above have accumulated in a talus deposit at the base of the slope. The rocks in this area have quite varied colors, and those are reflected in the colors of the talus.

A related process, frost heaving, takes place within unconsolidated materials on gentle slopes. In this case, water in the soil freezes and expands, pushing the overlying material up. Frost heaving is responsible for winter damage to roads all over North America.

When salt water seeps into rocks and then evaporates on a hot sunny day, salt crystals grow within cracks and pores in the rock. The growth of these crystals exerts pressure on the rock and can push grains apart, causing the rock to weaken and break. There are many examples of this on the rocky shorelines of Vancouver Island and the Gulf Islands, where sandstone outcrops are common and salty seawater is readily available (Figure 16.1.5). Salt weathering can also occur away from the coast, because most environments have some salt in them.



Figure 16.1.5 Honeycomb weathering of sandstone on Gabriola Island, B.C. The holes are caused by crystallization of salt within rock pores, and the seemingly regular pattern is related to the original roughness of the surface. It's a positive-feedback process because the holes collect salt water at high tide, and so the effect is accentuated around existing holes. This type of weathering is most pronounced on south-facing sunny exposures

The effects of plants and animals are significant in mechanical weathering. Roots can force their way into even the tiniest cracks, and then they exert tremendous pressure on the rocks as they grow, widening the cracks and breaking the rock (Figure 16.1.6).

Although animals do not normally burrow through solid rock, they can excavate and remove huge volumes of soil, and thus expose the rock to weathering by other mechanisms.



Figure 16.1.6 Conifers growing on granitic rocks at The Lions, near Vancouver, B.C.

Mechanical weathering is greatly facilitated by erosion, which is the removal of weathering products, allowing for the exposure of more rock for weathering. A good example of this is shown in Figure 16.1.4. On the steep rock faces at the top of the cliff, rock fragments have been broken off by ice wedging, and then removed by gravity. This is a form of mass wasting, which is discussed in more detail in Chapter 15. Other important agents of erosion that also have the effect of removing the products of weathering include water in streams (Chapter 13), glacial ice (Chapter 16), and waves on the coasts (Chapter 17).

Exercise 5.1 Mechanical weathering

This photo shows granitic rock at the top of Stawamus Chief near Squamish, B.C. Identify the mechanical weathering processes that you can see taking place, or you think probably take place at this location.

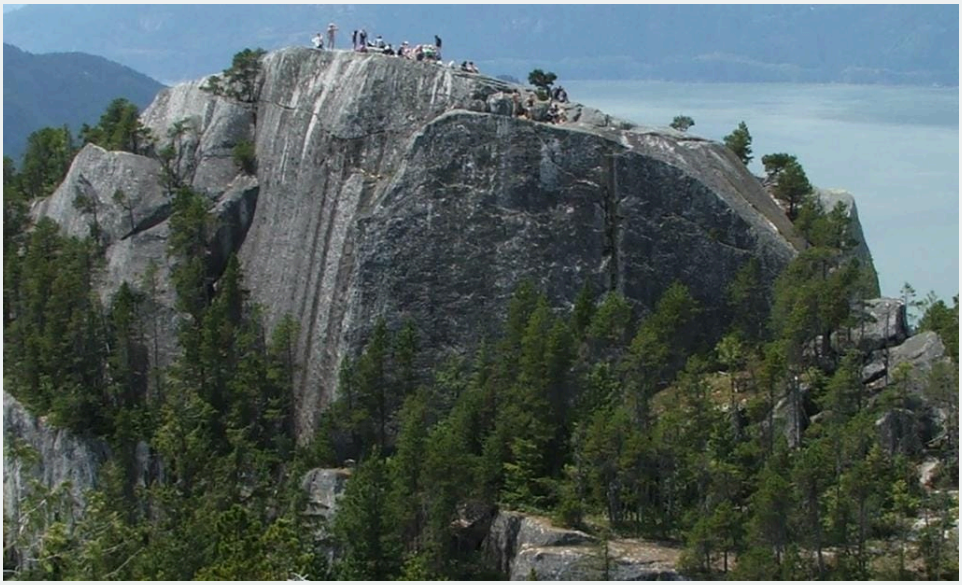


Figure 16.1.7

See Appendix 3 for Exercise 5.1 answers.

Media Attributions

- Figures 5.1.1, 5.1.2, 5.1.3, 5.1.4, 5.1.5, 5.1.6, 5.1.7: © Steven Earle. CC BY.

---

1. To a geologist, an outcrop is an exposure of bedrock, the solid rock of the crust. ↵

This page titled 16.1: Mechanical Weathering is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Steven Earle (BCCampus) via source content that was edited to the style and standards of the LibreTexts platform.

- 5.1: Mechanical Weathering by Steven Earle is licensed CC BY 4.0. Original source: https://opentextbc.ca/physicalgeology2ed.

16.2: Chemical Weathering

Chemical weathering results from chemical changes to minerals that become unstable when they are exposed to surface conditions. The kinds of changes that take place are highly specific to the mineral and the environmental conditions. Some minerals, like quartz, are virtually unaffected by chemical weathering, while others, like feldspar, are easily altered. In general, the degree of chemical weathering is greatest in warm and wet climates, and least in cold and dry climates. The important characteristics of surface conditions that lead to chemical weathering are the presence of water (in the air and on the ground surface), the abundance of oxygen, and the presence of carbon dioxide, which produces weak carbonic acid when combined with water. That process, which is fundamental to most chemical weathering, can be shown as follows:

$$H_2O + CO_2 \leftrightarrow H_2CO_3 \text{ then } H_2CO_3 \leftrightarrow H^+ + HCO_3^-$$

water + carbon dioxide ↔ carbonic acid then carbonic acid ↔ dissolved hydrogen ions + dissolved bicarbonate ions

Yikes! Chemical formulas

> Lots of people seize up when they are asked to read chemical or mathematical formulas. It's OK, you don't necessarily have to! If you don't like the formulas just read the text underneath them. In time you may get used to reading the formulas.
>
> The double-ended arrow "↔ " indicates that the reaction can go either way, but for our purposes these reactions are going towards the right.

Here we have water (e.g., as rain) plus carbon dioxide in the atmosphere, combining to create carbonic acid. Then carbonic acid dissociates (comes apart) to form hydrogen and bicarbonate ions. The amount of CO_2 in the air is enough to make weak carbonic acid. There is typically much more CO_2 in the soil, so water that percolates through the soil can become more acidic. In either case, this acidic water is a critical to chemical weathering.

In some types of chemical weathering the original mineral becomes altered to a different mineral. For example, feldspar is altered—by **hydrolysis**—to form **clay minerals** plus some ions in solution. In other cases the minerals dissolve completely, and their components go into solution. For example, calcite ($CaCO_3$) is soluble in acidic solutions.

The hydrolysis of feldspar can be written like this:

$$CaAl_2Si_2O_8 + H_2CO_3 + ½O_2 \leftrightarrow Al_2Si_2O_5(OH)_4 + Ca^{2+} + CO_3^{2-}$$

plagioclase feldspar + carbonic acid ↔ kaolinite + dissolved calcium ions + dissolved carbonate ions

This reaction shows calcium-bearing plagioclase feldspar, but similar reactions could also be written for sodium or potassium feldspars. In this case, we end up with the mineral kaolinite, along with calcium and carbonate ions in solution. Those ions can eventually combine (probably in the ocean) to form the mineral calcite. The hydrolysis of feldspar to clay is illustrated in Figure 16.2.1, which shows two images of the same granitic rock, a recently broken fresh surface on the left and a clay-altered weathered surface on the right. Other silicate minerals can also go through hydrolysis, although the end results will be a little different. For example, pyroxene can be converted to the clay minerals chlorite or smectite, and olivine can be converted to the clay mineral serpentine.

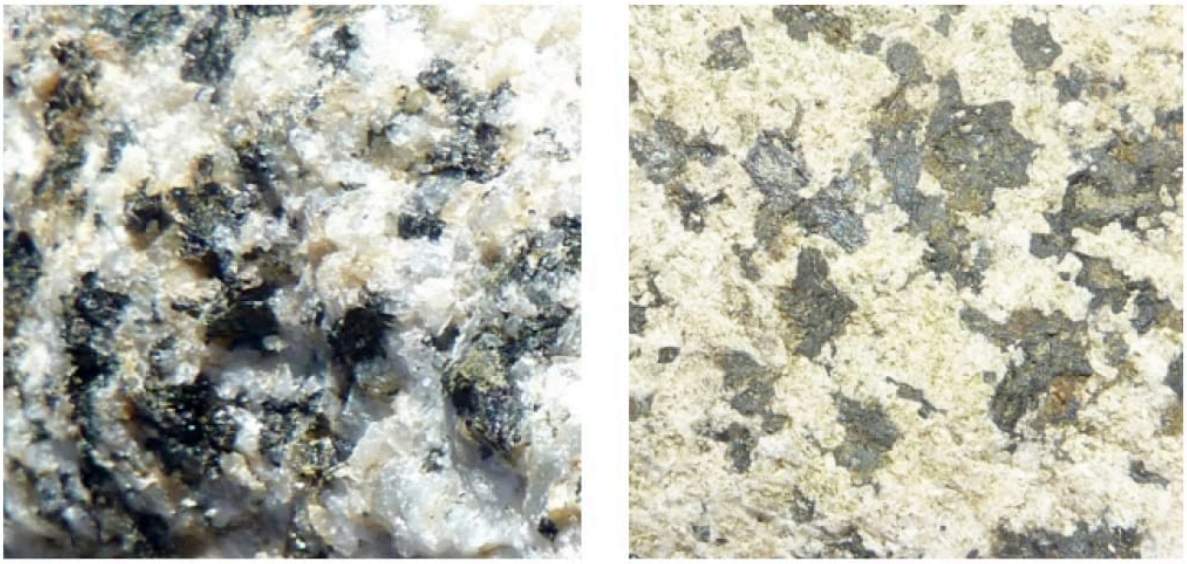


Figure 16.2.1 Unweathered (left) and weathered (right) surfaces of the same piece of granitic rock. On the unweathered surfaces the feldspars are still fresh and glassy-looking. On the weathered surface much of the feldspar has been altered to the chalky-looking clay mineral kaolinite.

Oxidation is another very important chemical weathering process. The oxidation of the iron in a ferromagnesian silicate starts with the dissolution of the iron. For olivine, the process looks like this, where olivine in the presence of carbonic acid is converted to dissolved iron, carbonate, and silicic acid:

$$Fe_2SiO_4 + 4H_2CO_3 \leftrightarrow 2Fe^{2+} + 4HCO_3^- + H_4SiO_4$$

olivine + (carbonic acid) ↔ dissolved iron ions + dissolved carbonate ions + dissolved silicic acid

But in the presence of oxygen and carbonic acid, the dissolved iron is then quickly converted to the mineral hematite:

$$2Fe^{2+} + 4HCO_3^- + ½\ O_2 + 2H_2O \leftrightarrow Fe_2O_3 + 4H_2CO_3$$

dissolved iron ions + dissolved bicarbonate ions + oxygen + water ↔ hematite + carbonic acid

The equation shown here is for olivine, but it could apply to almost any other ferromagnesian silicate, including pyroxene, amphibole, or biotite. Iron in the sulfide minerals (e.g., pyrite) can also be oxidized in this way. And the mineral hematite is not the only possible end result, as there is a wide range of iron oxide minerals that can form in this way. The results of this process are illustrated in Figure 16.2.2, which shows a granitic rock in which some of the biotite and amphibole have been altered to form the iron oxide mineral limonite.



Figure 16.2.2 A granitic rock containing biotite and amphibole which have been altered near to the rock's surface to limonite, which is a mixture of iron oxide minerals.

A special type of oxidation takes place in areas where the rocks have elevated levels of sulfide minerals, especially pyrite (FeS_2). Pyrite reacts with water and oxygen to form sulfuric acid, as follows:

$$2FeS_2 + 7O_2 + 2H_2O \leftrightarrow 2Fe^{2+} H_2SO_4 + 2H^+$$

pyrite + oxygen + water ↔ dissloved iron ions + sulfuric acid + dissolved hydrogen ions

The runoff from areas where this process is taking place is known as **acid rock drainage** (ARD), and even a rock with 1% or 2% pyrite can produce significant ARD. Some of the worst examples of ARD are at metal mine sites, especially where pyrite-bearing rock and waste material have been mined from deep underground and then piled up and left exposed to water and oxygen. One example of that is the Mt. Washington Mine near Courtenay on Vancouver Island (Figure 16.2.3), but there are many similar sites across Canada and around the world.



Figure 16.2.3 Exposed oxidizing and acid generating rocks and mine waste at the abandoned Mt. Washington Mine, B.C. (left), and an example of acid drainage downstream from the mine site (right).

At many ARD sites, the pH of the runoff water is less than 4 (very acidic). Under these conditions, metals such as copper, zinc, and lead are quite soluble, and this can lead to toxicity for aquatic and other organisms. For many years, the river downstream from the Mt. Washington Mine had so much dissolved copper in it that it was toxic to salmon. Remediation work has since been carried out at the mine and the situation has improved.

The hydrolysis of feldspar and other silicate minerals and the oxidation of iron in ferromagnesian silicates all serve to create rocks that are softer and weaker than they were to begin with, and thus more susceptible to mechanical weathering.

The weathering reactions that we've discussed so far involved the transformation of one mineral to another mineral (e.g., feldspar to clay), and the release of some ions in solution (e.g., Ca^{2+} or Fe^{2+}). Some weathering processes involve the complete dissolution of a mineral. Calcite, for example, will dissolve in weak acid, to produce calcium and bicarbonate ions. The equation is as follows:

$$\mathbf{CaCO_3 + H^+ + HCO_3^- \leftrightarrow Ca^{2+} + 2HCO_3^-}$$

calcite + dissolved hydrogen ions + dissolved bicarbonate ions ↔ dissolved calcium ions + dissolved bicarbonate ions

Calcite is the major component of limestone (typically more than 95%), and under surface conditions, limestone can dissolve completely, as shown in Figure 16.2.4. Limestone also dissolves at relatively shallow depths underground, forming limestone caves. This is discussed in more detail in Chapter 14, where we look at **groundwater**.



Figure 16.2.4 A limestone outcrop on Quadra Island, B.C. The limestone, which is primarily made up of the mineral calcite, has been dissolved to different degrees in different areas because of compositional differences. The buff-colored bands are chert, which stands out because it is not soluble.

Exercise 5.2 Chemical Weathering

The main processes of chemical weathering are hydrolysis, oxidation, and dissolution. Indicate which process is primarily involved during each of the following chemical weathering changes:

1. Pyrite to hematite
2. Calcite to calcium and bicarbonate ions
3. Feldspar to clay
4. Olivine to serpentine

See Appendix 3 for Exercise 5.2 answers.

Media Attributions

- Figure 16.2.1, 5.2.2, 5.2.3, 5.2.4: © Steven Earle. CC BY.

This page titled 16.2: Chemical Weathering is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Steven Earle (BCCampus) via source content that was edited to the style and standards of the LibreTexts platform.

- 5.2: Chemical Weathering by Steven Earle is licensed CC BY 4.0. Original source: https://opentextbc.ca/physicalgeology2ed.

16.3: The Products of Weathering and Erosion

The products of weathering and erosion are the unconsolidated materials that we find around us on slopes, beneath, beside and on top of glaciers, in stream valleys, on beaches, and in deserts. The nature of these materials—their composition, size, degree of sorting, and degree of rounding—is determined by the type of rock that is being weathered, the nature of the weathering, the erosion and transportation processes, and the climate.

In addition to these solid sediments, the other important products of weathering are several different types of ions in solution.

A summary of the weathering products of some of the common minerals present in rocks is provided in Table 5.1. In addition to the weathering products listed in the table, most of the larger fragments—larger than sand grains—that make up sediments will be pieces of rock as opposed to individual minerals.

Table 5.1 A list of the typical weathering products of some of the minerals in common rocks

Common Mineral	Typical Weathering Products
Quartz	Quartz as sand grains
Feldspar	Clay minerals plus potassium, sodium, and calcium in solution
Biotite and amphibole	Chlorite plus iron and magnesium in solution
Pyroxene and olivine	Serpentine plus iron and magnesium in solution
Calcite	Calcium and carbonate in solution
Pyrite	Iron oxide minerals plus iron in solution and sulfuric acid

Some examples of the products of weathering are shown in Figure 16.3.1. They range widely in size and shape depending on the processes involved in their transportation. If and when deposits like these are turned into sedimentary rocks, the textures of those rocks will vary significantly. Importantly, when we describe sedimentary rocks that formed millions of years in the past, we can use those properties to make inferences about the conditions that existed during their formation.



Boulders in a talus deposit at Keremeos. All are angular fragments from the same rock source.

Pebbles on a beach in Victoria. All are rounded fragments of rock from different sources.

Sand from a beach at Gabriola. Most are angular quartz grains, some are fragments of rock.

Sand from a dune in Utah. All are rounded quartz grains.

Figure 16.3.1 Products of weathering and erosion formed under different conditions.

We'll talk more about the nature and interpretation of sediments and sedimentary rocks in Chapter 6, but it's worth considering here why the sand-sized sediments shown in Figure 16.3.1 are so strongly dominated by the mineral quartz, even though quartz makes up less than 20% of Earth's crust. The explanation is that quartz is highly resistant to the types of weathering that occur at Earth's surface. It is not affected by weak acids or the presence of oxygen. This makes it unique among the minerals that are

common in igneous rocks. Quartz is also very hard, and doesn't have cleavage, so it is resistant to mechanical erosion.

So when a rock like granite is subject to chemical weathering the feldspar and the ferromagnesian silicates get converted to clays and dissolved ions such as: Ca^{2+}, Na^{+}, K^{+}, Fe^{2+}, Mg^{2+}, and H_4SiO_4, but the quartz is resistant to those processes and remains intact. The clay gradually gets eroded away, then the rock breaks apart leaving lots of grains of quartz. In other words, quartz, clay minerals, and dissolved ions are the most common products of weathering. Quartz and some of the clay minerals tend to form sedimentary deposits on and at the edges of continents, while the rest of the clay minerals and the dissolved ions tend to be washed out into the oceans to form sediments on the sea floor.

Exercise 5.3 Describing the weathering origins of sand

In the left side of the following table, a number of different sands are pictured and described. Describe some of the important weathering processes that might have led to the development of these sands.

See Appendix 3 for Exercise 5.3 answers.

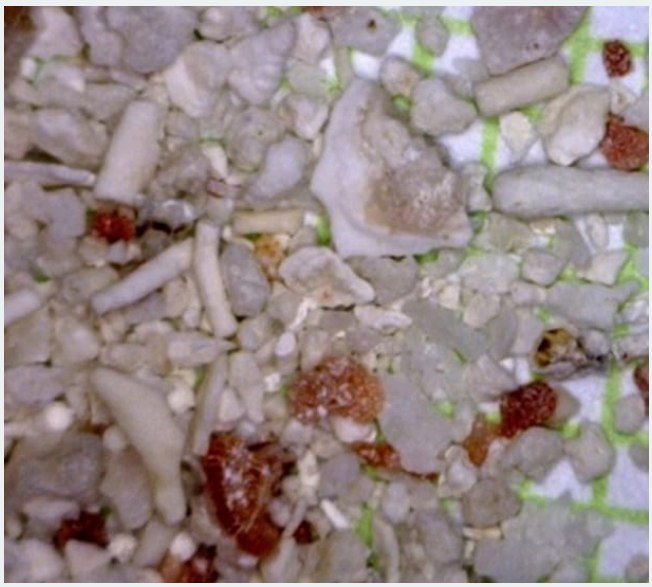
Image	Description and Location
	Fragments of coral, algae, and urchin from a shallow water area (roughly 2 meters deep) near a reef in Belize. The grain diameters are between 0.1 and 1 millimeters.
	Angular quartz and rock fragments from a glacial stream deposit near Osoyoos, B.C. The grain diameters are between 0.25 and 0.5 milimeters.

Image	Description and Location
	Rounded grains of olivine (green) and volcanic glass (black) from a beach on the big island of Hawaii. The grains are approximately 1 millimetre across.

Image Descriptions

Figure 16.3.1 image description: Examples of weathering and erosion.

1. Boulders in a talus deposit at Keremeos. All are angular fragments from the same rock source.
2. Pebbles on a beach in Victoria. All are rounded fragments of rock from different sources.
3. Sand from a beach at Gabriola Island. most are angular quartz grains, some are sand-sized fragments of rock.
4. Sand from a dune in Utah. All are rounded quartz grains.

[Return to Figure 16.3.1]

Media Attributions

- Figure 16.3.1: © Steven Earle. CC BY.

This page titled 16.3: The Products of Weathering and Erosion is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Steven Earle (BCCampus) via source content that was edited to the style and standards of the LibreTexts platform.

- 5.3: The Products of Weathering and Erosion by Steven Earle is licensed CC BY 4.0. Original source: https://opentextbc.ca/physicalgeology2ed.

16.4: Weathering and the Formation of Soil

Weathering is a key part of the process of soil formation, and soil is critical to our existence on Earth. In other words, we owe our existence to weathering, and we need to take care of our soil!

Many people refer to any loose material on Earth's surface as soil, but to geologists (and geology students) soil is the material that includes organic matter, lies within the top few tens of centimeters of the surface, and is important in sustaining plant growth.

Soil is a complex mixture of minerals (approximately 45%), organic matter (approximately 5%), and empty space (approximately 50%, filled to varying degrees with air and water). The mineral content of soils is variable, but is dominated by clay minerals and quartz, along with minor amounts of feldspar and small fragments of rock. The types of weathering that take place within a region have a major influence on soil composition and texture. For example, in a warm climate, where chemical weathering dominates, soils tend to be richer in clay. Soil scientists describe soil texture in terms of the relative proportions of sand, silt, and clay, as shown in Figure 16.4.1. The sand and silt components in this diagram are dominated by quartz, with lesser amounts of feldspar and rock fragments, while the clay component is dominated by the clay minerals.

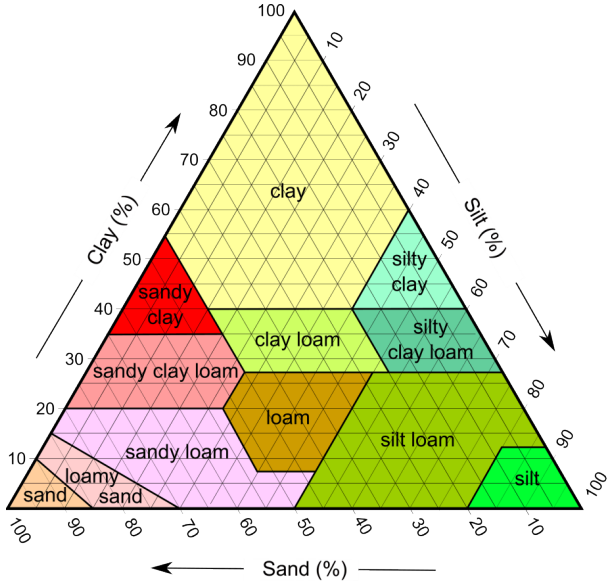


Figure 16.4.1 This diagram applies only to the mineral component of soils, and the names are textural descriptions, not soil classes.

Soil forms through accumulation and decay of organic matter and through the mechanical and chemical weathering processes described above. The factors that affect the nature of soil and the rate of its formation include climate (especially average temperature and precipitation amounts, and the consequent types and intensity of vegetation), the type of parent material, the slope of the surface, and the amount of time available.

Climate

Soils develop because of the weathering of materials on Earth's surface, including the mechanical breakup of rocks, and the chemical weathering of minerals. Soil development is facilitated by the downward percolation of water. Soil forms most readily under temperate to tropical conditions (not cold) and where precipitation amounts are moderate (not dry, but not too wet). Chemical weathering reactions (especially the formation of clay minerals) and biochemical reactions proceed fastest under warm conditions, and plant growth is enhanced in warm climates. Too much water (e.g., in rainforests) can lead to the leaching of important chemical nutrients and hence to acidic soils. In humid and poorly drained regions, swampy conditions may prevail, producing soil that is dominated by organic matter. Too little water (e.g., in deserts and semi-deserts), results in very limited downward chemical transportation and the accumulation of salts and carbonate minerals (e.g., calcite) from upward-moving water. Soils in dry regions also suffer from a lack of organic material (Figure 16.4.2).



Figure 16.4.2 Poorly developed soil on wind-blown silt (loess) in an arid part of northeastern Washington State. The thickness shown is about 1 m, and the "soil" is just the upper 2 or 3 cm.

Parent Material

Soil parent materials can include all different types of bedrock and any type of unconsolidated sediments, such as glacial deposits and stream deposits. Soils are described as **residual soils** if they develop on bedrock, and transported soils if they develop on transported material such as glacial sediments. Other sources may use the term "transported soil" to imply that the soil itself has been transported, but in this text "transported soil" is soil that is developed on transported materials, like the very thin soil shown in Figure 16.4.2. When referring to such soil, it is better to be specific and say "soil developed on unconsolidated material," because that distinguishes it from soil developed on bedrock.

Quartz-rich parent material, such as granite, sandstone, or loose sand, leads to the development of sandy soils. Quartz-poor material, such as shale or basalt, generates soils with little sand.

Parent materials provide important nutrients to residual soils. For example, a minor constituent of granitic rocks is the calcium-phosphate mineral apatite ($Ca_5(PO_4)_3(F,Cl,OH)$), which is a source of the important soil nutrient phosphorus. Basaltic parent material tends to generate very fertile soils because it also provides phosphorus, along with significant amounts of iron, magnesium, and calcium.

Some unconsolidated materials, such as river-flood deposits, make for especially good soils because they tend to be rich in clay minerals. Clay minerals have large surface areas with negative charges that are attractive to positively charged elements like calcium, magnesium, iron, and potassium—important nutrients for plant growth.

Slope

Soil can only develop where surface materials remain in place and are not frequently moved away by mass wasting. Soils cannot develop where the rate of soil formation is less than the rate of erosion, so steep slopes tend to have little or no soil.

Time

Even under ideal conditions, soil takes thousands of years to develop. Virtually all of southern Canada was still glaciated up until 14 ka, and most of the central and northern parts of B.C., the prairies, Ontario, and Quebec were still glaciated at 12 ka. Glaciers still dominated the central and northern parts of Canada until around 10 ka, and so, at that time, conditions were still not ideal for soil development even in the southern regions. Therefore, soils in Canada, and especially in central and northern Canada, are relatively young and not well developed.

The same applies to soils that are forming on newly created surfaces, such as recent deltas or sand bars, or in areas of mass wasting.

Soil Horizons

The process of soil formation generally involves the downward movement of clay, water, and dissolved ions, and a common result of that is the development of chemically and texturally different layers known as **soil horizons**. The typically developed soil horizons, as illustrated in Figure 16.4.3, are:

- O — the layer of organic matter
- A — the layer of partially decayed organic matter mixed with mineral material
- E— the eluviated (leached) layer from which some of the clay and iron have been removed to create a pale layer that may be sandier than the other layers
- B — the layer of accumulation of clay, iron, and other elements from the overlying soil
- C — the layer of incomplete weathering

Although rare in Canada, another type of layer that develops in hot arid regions is known as **caliche** (pronounced *ca-lee-chee*). It forms from the downward (or in some cases upward) movement of calcium ions, and the precipitation of calcite within the soil. When well developed, caliche cements the surrounding material together to form a layer that has the consistency of concrete.

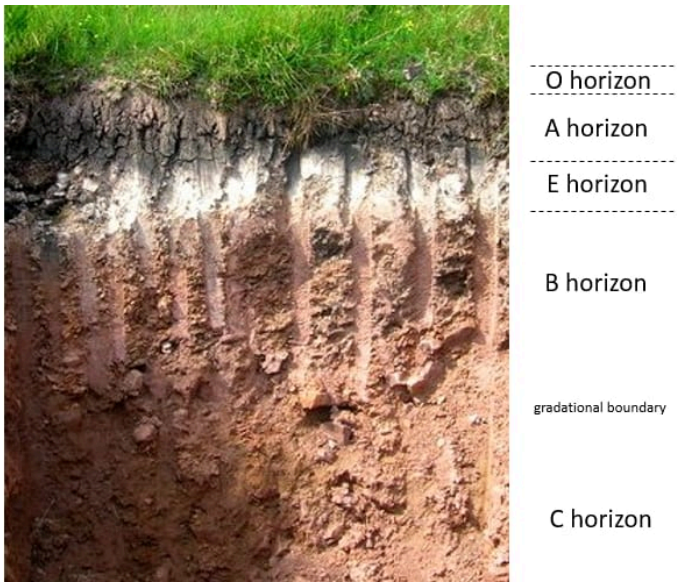


Figure 16.4.3 Soil horizons in a podsol from a site in northeastern Scotland.

Like all geological materials, soil is subject to erosion, although under natural conditions on gentle slopes, the rate of soil formation either balances or exceeds the rate of erosion. Human practices, especially those related to forestry and agriculture, have significantly upset this balance.

Soils are held in place by vegetation. When vegetation is removed, either through cutting trees or routinely harvesting crops and tilling the soil, that protection is either temporarily or permanently lost. The primary agents of the erosion of unprotected soil are water and wind.

Water erosion is accentuated on sloped surfaces because fast-flowing water obviously has greater eroding power than slow-flowing or still water (Figure 16.4.4). Raindrops can disaggregate exposed soil particles, putting the finer material (e.g., clays) into suspension in the water. **Sheetwash—**unchannelled flow across a surface—carries suspended material away, and channels erode right through the soil layer, removing both fine and coarse material.



Figure 16.4.4 Soil erosion by rain and channelled runoff on a field in Alberta.

Wind erosion is exacerbated by the removal of trees that act as wind breaks and by agricultural practices that leave bare soil exposed (Figure 16.4.5).

Tillage is also a factor in soil erosion, especially on slopes, because each time the soil is lifted by a cultivator, it is moved a few centimeters down the slope.



Figure 16.4.5 Soil erosion by wind in Alberta.

Media Attributions

- Figure 16.4.1: "SoilTexture USDA" © Mikenorton. CC BY-SA.
- Figure 16.4.2: © Steven Earle. CC BY.
- Figure 16.4.3: Podzol © Ailith Stewart. CC BY-SA.
- Figure 16.4.4: Image © Alberta Agriculture and Rural Development. All rights reserved. Used with permission.
- Figure 16.4.5: Image © Alberta Agriculture and Rural Development. All rights reserved. Used with permission.

This page titled 16.4: Weathering and the Formation of Soil is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Steven Earle (BCCampus) via source content that was edited to the style and standards of the LibreTexts platform.

- 5.4: Weathering and the Formation of Soil by Steven Earle is licensed CC BY 4.0. Original source: https://opentextbc.ca/physicalgeology2ed.

17: WATER AND LANDSCAPES

17.1: Topographic Maps

17.2: Surface Water

17.2.1: The Hydrological Cycle

17.2.2: Drainage Basins

17.2.3: Stream Erosion and Deposition

17.2.4: Stream Types

17.2.5: Flooding

17.3: Groundwater

17.3.1: Groundwater and Aquifers

17.3.2: Groundwater Flow

17.3.3: Groundwater Extraction

17.3.4: Groundwater Quality

17.4: Mass Wasting

17.4.1: Factors That Control Slope Stability

17.4.2: Classification of Mass Wasting

17.4.3: Preventing, Delaying, Monitoring, and Mitigating Mass Wasting

This page titled 17: Water and Landscapes is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Robert Wagner.

17.1: Topographic Maps

Almost all of the area of the United States has been represented on topographic maps at various scales. For many years there were two standard map scales:

- Maps, called fifteen-minute quadrangle maps, that spanned 15' of longitude and 15' of latitude at a scale of 1:62,500, which is exactly one inch to the mile.
- Maps, called seven-and-a-half-minute quadrangle maps, that spanned 7-1/2' of longitude and 7-1/2' of latitude at a scale of 1:24,00.

In recent times those quadrangle maps have been supplanted by seven-and-a-half-minute quadrangle maps that are about the same size at the old 1:24,000 maps but at a slightly different, but more rational scale of 1:25,000. In addition, there is a series of maps with a scale of 1:250,000, called *quarter-million map*s. These latter, newer, maps use the metric system for map distances and contour intervals.

The 1:25,000 maps are what geologists typically use as a base for mapping the geology of a local area. The resulting maps are called ***geologic quadrangle maps***. But often the detail of mapping that the geologist desires calls for a larger scale: that is, the map scale is such that a smaller land area is represented on a given sheet of paper. Detailed geologic maps are often at scales of between 1:10,000 and 1:5,000. *The scale of map you work with depends on the level of detail you want to show on the map.*

Topographic Contours

Many maps of land areas have series of curved lines, called contours, that represent the topography of the area. Such a map is called a ***topographic map***. A ***contour*** is *a horizontal curve that's the locus of all points on the map with the same elevation* (Figure 3-35). A good way to understand the concept of contours is to choose a series of elevations and imagine passing a corresponding series of horizontal planes through the landscape (Figure 3-36). The contours are formed by the intersection of these planes with the land surface.

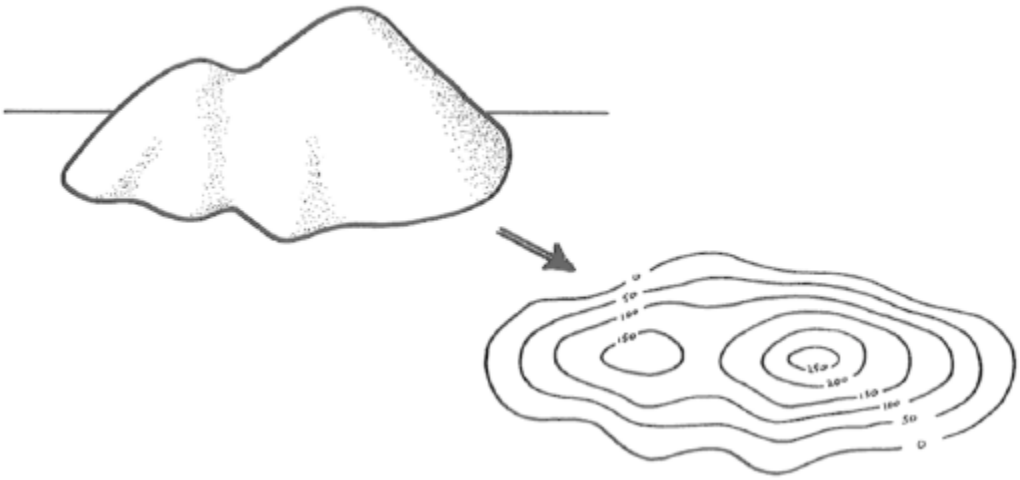


Figure 3-35. Topographic contour lines.

Adjacent contours touch one another only where the land surface is vertical, and they cross one another only where one part of the land surface hangs over another part (and in such overhanging areas, by convention the contours are dashed or dotted to indicate the overhanging relationship, or just omitted altogether).

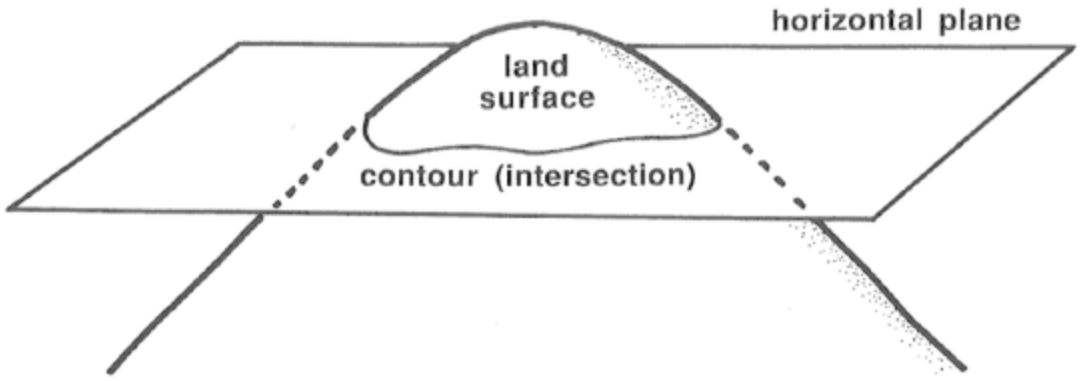


Figure 3-36. Viewing contours as the loci of horizontal planes with the land surface.

Topographic maps are useful to both technical people (geologists; environmental engineers) and to everyday people, like hikers. A geologist often uses a topographic map as a base on which to plot geology in making a geologic map.

Interpreting Contour Lines on Topographic Maps

To be a sophisticated user of a topographic map, you have to be able to *visualize the topography or "lay of the land" by examining the contour lines*. Or, even if you can't do a good job of visualizing the topography, you at least have to understand how to interpret it from the contour lines. Here are some basic pointers. The best way to learn how to visualize topography from contour lines is to practice with a real map—and you will be doing that in the take-home exercise.

For a map at a given scale, *the more closely spaced the contour lines, the steeper to slope of the surface*: closely spaced contour lines mean steep slopes, and widely spaced contour lines mean gentle slopes (Figure 3-37).

You can easily compute the local slope of the land surface, by:

1. measuring the distance between two points on the map,
2. using the map scale to find the real distance on the land surface,
3. finding the difference in elevation between the two points from the contour lines (at each point you usually have to interpolating between adjacent contour lines),
4. forming a right triangle with one leg the horizontal distance and one leg the vertical distance, and
5. finding the slope angle by trigonometry.

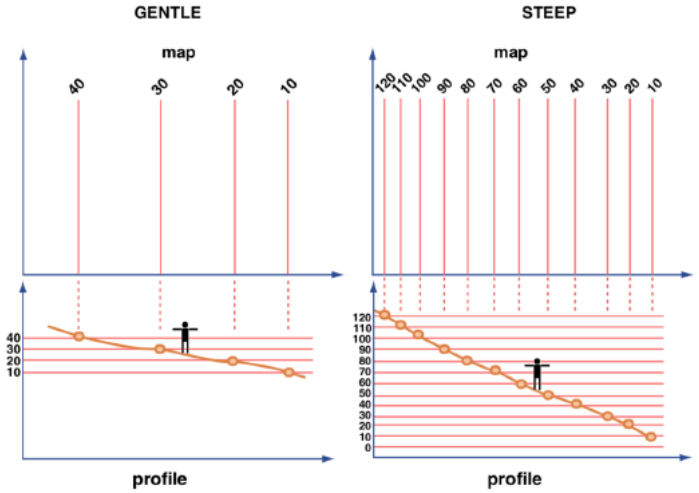


Figure 3-37. Contours on gently sloping surfaces vs. contours on steeply sloping surfaces.

I suppose I need not insult your intelligence by pointing out (Figure 3-38) that hills or ridges are located where the contour lines on either side of the hill or ridge increase in elevation toward the top of the hill, and valleys are located where the contour lines on either side of the valley decrease in elevation toward the bottom of the valley.

Valleys usually slope in one direction or another because they are occupied by streams, which, as you know, flow downhill. Such a valley shows up on the topographic map as a series of V-shaped kinks in the successive contour lines. Those “V”s point *up* the valley (Figure 3-39A). Conversely, ridges commonly have crests that slope downward in one direction or the other, and the contour lines in that case form “V”s that point *down* the ridge or spur (Figure 3-39B).

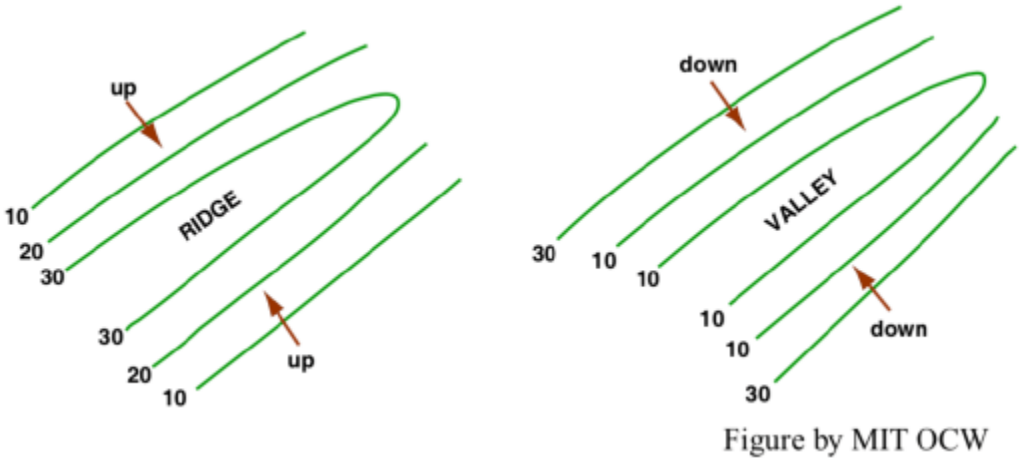


Figure 3-38. Contour lines for ridges and valleys.

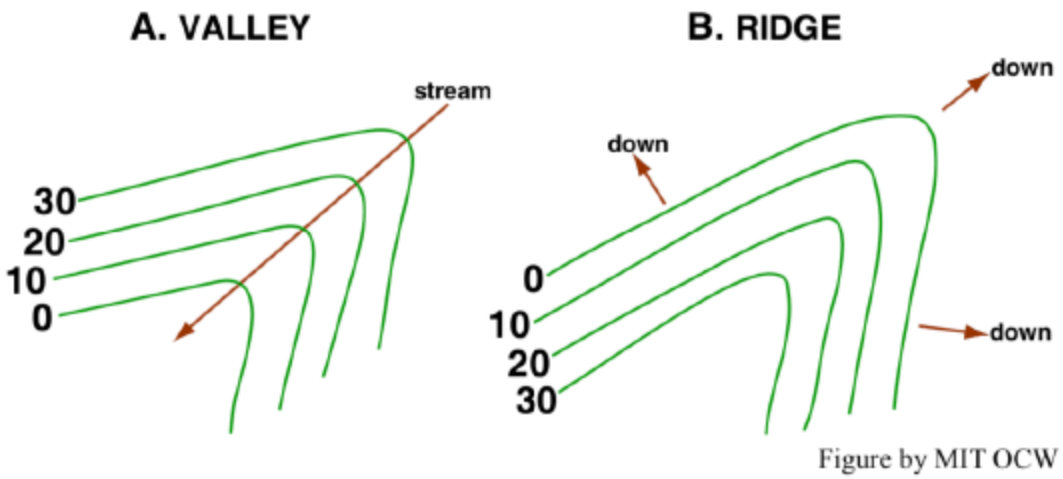


Figure 3-39: A) The “V”s of the contour lines point up valleys. B) The “V”s of the contour lines point down ridges.

How do you know the elevation of a hilltop from a contour map? The answer is: you don’t, really, unless that information is supplied separately by being printed on the map! That’s because the tippity-top in general lies on the “incomplete” contour interval above the uppermost closed contour. But at least you can bracket the elevation of the summit to be within one contour interval (Figure 3-40).

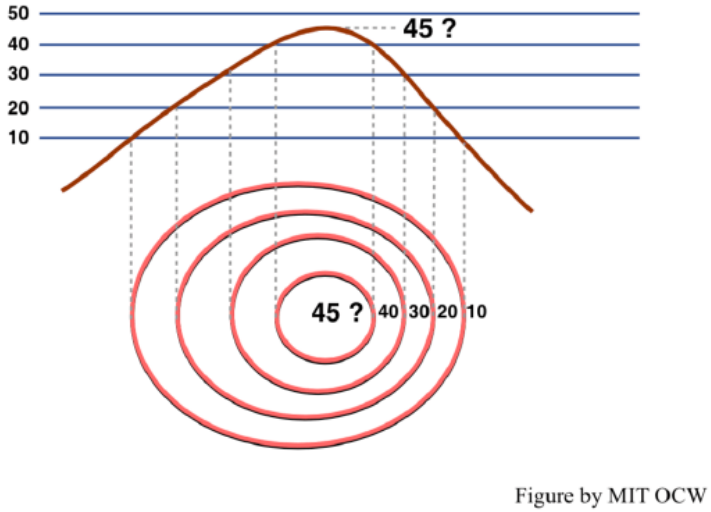


Figure 3-40. Estimating the elevation of a hilltop from a contour map.

Closed depressions on the land surface are the opposite of hilltops, but in most areas they are uncommon or even nonexistent. Closed contours for which the elevation decreases inward are denoted in a special way, by putting little perpendicular tick marks on the *downslope* side of the closed contour (Figure 3-41).

Overhangs (also not common) are also treated in a special way: the obscured segments of contour lines are either ignored or drawn as dashed curves rather than as solid curves (Figure 3-42).

Now think in terms of riding saddles or curvy potato chips, or, for those mathematically inclined, hyperbolic paraboloids (Figure 3-11). In all of these cases, one is dealing with a curved surface such that the intersection between the surface and a vertical plane is convex up in one direction and concave up in a direction approximately at right angles to the first. Such features on topographic maps are called ***saddles*** or ***passes***. Saddles are common along ridges where stream valleys indent the ridge from two opposite sides.

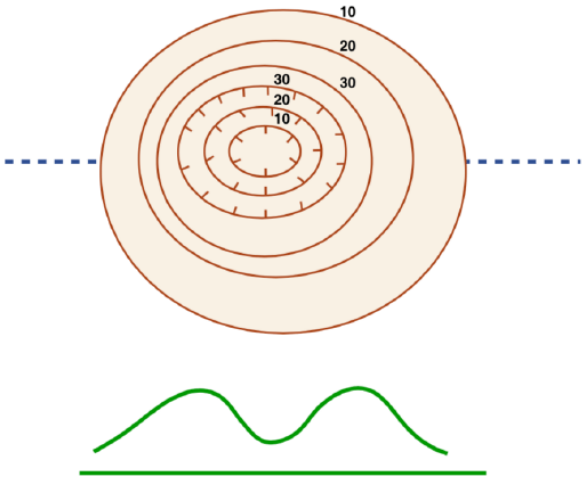


Figure by MIT OCW

Figure 3-41. How contour lines represent depressions.

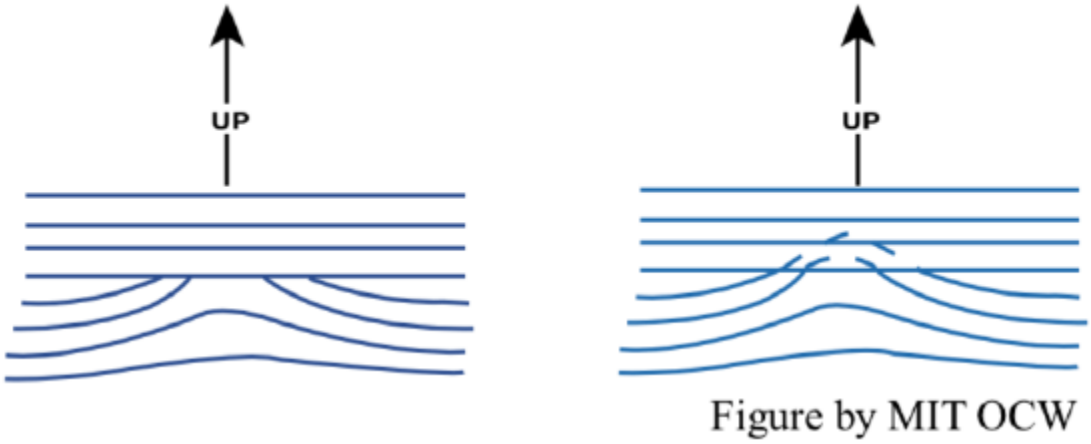


Figure by MIT OCW

Figure 3-42. How contour lines represent overhangs.

Locating Yourself on the Map

When you are working outdoors with a topographic map, the map is likely to be much more useful when you are able to *locate where you are on the map*. That's a skill we won't be able to practice in this course, unfortunately. Below are some considerations for you to keep in mind, in case you are called upon to work outdoors with a topographic map.

One way of locating yourself is simply to *choose to be somewhere that corresponds to a readily identifiable feature on the map*, like a road intersection, a stream confluence, or a mountaintop. But that's overly restrictive: suppose you need to be somewhere not near such a feature?

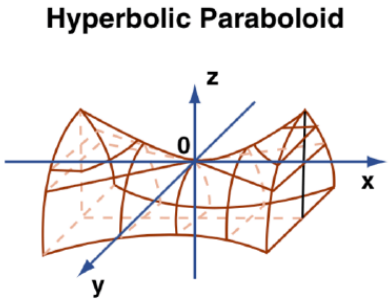


Figure by MIT OCW

Figure 3-43. A hyperbolic paraboloid (a kind of "saddle") (From Burington, 1948.)

One technique, which is useful in arid and semiarid regions with substantial relief (by ***relief*** I mean *local differences in elevation from place to place*) is to look back and forth from map to land, several times, to relate the topography shown on the map to the topography you can see on the ground. It's a knack one acquires, usually readily and without much difficulty, by practice.

Note

I have not been able to develop a laboratory exercise to make that technique real to students in a classroom, although for some years I have intended to put into practice an idea that might be useful in that regard. Perhaps one of you might be moved to pursue the idea. I would like to build a small table-top model of some small area with rugged mountain-and-valley topography by blowing up the topographic map of the area, cutting successive layers of a sheetlike material like the foam-core sheets used to mount prints, by tracing the outline of successive contour lines on successive sheets of the foam core, then stacking the sheets together, gluing them one by one, and covering the entire "terraced" mass with a moldable material like wallboard joint compound, and finally sanding the surface smooth. With such a model, one could do two instructive things:

1. mark particular points on the model and ask students to locate the points on the map, and
2. mark particular points on the map and ask students to show the corresponding location on the model.

It would be an extremely valuable teaching tool. If anyone, perhaps with some experience along that line, is interested in working on such a model, let me know.

The other technique, which is necessary in areas with low relief and/or heavy, view-obstructing vegetation, is *surveying*. You start from a known point on the map and run a survey line to the point of interest. This can be as crude as a compass-and-pace traverse or as sophisticated as a professional survey.

Visualizing the Lay of the Land

In working with topographic maps, one skill you need to develop is the ability to visualize the lay of the land from a topographic contour map. This involves picturing, in your mind's eye, what the map area would really look like, in a view from a low-flying airplane on a clear day, just from the information provided by the topographic contours. To make full use of a topographic map, you need to be able to do that.

Visualizing topography from looking at the topographic contours on the map is a skill you have to develop for yourself, by thought and lots of practice. Some people can do it much more easily than others. It's an exercise in using your "mind's eye", which is a useful skill in earth science, as well as in everyday life. (Think ahead toward when you might be wanting to remodel your home: can you picture the shape of your new kitchen, and how everything will fit together in it in three dimensions?)

To develop your skill in visualization, I will have a take-home exercise for you in which you will have to build a model of the earth's surface, out of modeling clay, just by looking at a topographic map of the area.

Topographic Profiles

It is often useful to obtain a vertical cross-section view of the land surface along a line that extends from one point on the land to another point. It's easy to construct such a profile if you have a topographic map of the area already available. Here's how to proceed:

1. Pinpoint the two points at the end of the desired profile directly on your topographic map, and draw a faint pencil line between them.
2. Lay out, on a blank sheet of paper, a corresponding faint pencil line near the top of the paper than extends from the starting point on the land surface to the ending point. Establish a conversion ratio between the scale of the map and the scale of your cross-section line; then, as you pick points off the map, you can easily plot them along your cross-section line.
3. For a large number of points along the profile line on the topographic map that happen to fall on topographic contour lines, pick off the elevations, and transfer them to your cross-section line on the sheet of paper.
4. Connect the elevation points on your cross-section with a smooth curve.

This page titled 17.1: Topographic Maps is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by John Southard (MIT OpenCourseware) via source content that was edited to the style and standards of the LibreTexts platform.

- 3.3: Topographic Maps by John Southard is licensed CC BY-NC-SA 4.0. Original source: https://ocw.mit.edu/courses/12-090-the-environment-of-the-earths-surface-spring-2007/.

17.2: Surface Water

This section examines how water moves across Earth's surface and shapes the landscape. It follows water through the hydrologic cycle and into drainage basins and stream systems, then explores how flowing water erodes, transports, and deposits sediment. Different stream patterns and processes are considered along with flooding and its effects on people and landscapes.

This page titled 17.2: Surface Water is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Robert Wagner.

17.2.1: The Hydrological Cycle

Water is constantly on the move. It is evaporated from the oceans, lakes, streams, the surface of the land, and plants (transpiration) by solar energy (Figure 17.2.1.1). It is moved through the atmosphere by winds and condenses to form clouds of water droplets or ice crystals. In response to the pull of gravity it comes back down as rain or snow and then flows through streams, into lakes, and eventually back to the oceans. Water on the surface and in streams and lakes infiltrates the ground to become groundwater. Groundwater slowly moves through the rock and surficial materials. Some groundwater returns to other streams and lakes, and some goes directly back to the oceans.

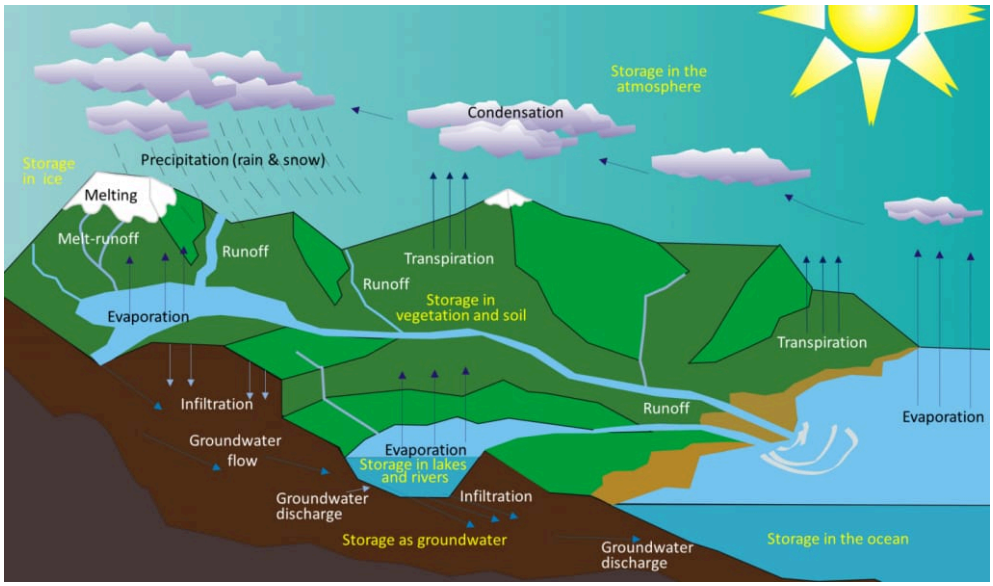


Figure 17.2.1.1 The various components of the water cycle. Black or white text indicates the movement or transfer of water from one reservoir to another. Yellow text indicates the storage of water.



Figure 17.2.1.2 Representation of the Earth's water. The 1 litre jug is filled with salty sea water (97%). The ice-cube is glacial ice (2%). The 2 teaspoons represent groundwater (1%), and the three drops represent all of the fresh water in lakes, streams, and wetlands, plus all of the water in the atmosphere.

Even while it's moving around, water is stored in various reservoirs. The largest, by far, is the oceans, accounting for 97% of the volume. Of course, that water is salty. The remaining 3% is fresh water. Two-thirds of our fresh water is stored in ice and one-third is stored in the ground. The remaining fresh water—about 0.03% of the total—is stored in lakes, streams, vegetation, and

the atmosphere. To put that in perspective, let's imagine putting all of Earth's water into a 1 litre jug (Figure 17.2.1.2). We start by almost filling the jug with 970 millilitres of water and 34 grams of salt. Then we add one regular-sized (roughly 20 millilitres) ice cube (representing glacial ice) and two teaspoons (roughly 10 millilitres) of groundwater. All of the water that we see around us in lakes and streams and up in the sky can be represented by adding three more drops from an eyedropper.

Although the proportion of Earth's water that is in the atmosphere is tiny, the actual volume is huge. At any given time, there is the equivalent of approximately 13,000 cubic kilometers (km^3) of water in the air in the form of water vapor and water droplets in clouds. Water is evaporated from the oceans, vegetation, and lakes at a rate of 1,580 km^3 per day, and just about exactly the same volume falls as rain and snow every day—over both the oceans and land. The precipitation that falls on land goes back to the ocean in the form of stream flow (117 km^3/day) and groundwater flow (6 km^3/day). Most of the rest of this chapter is about that 117 km^3/day of streamflow. The average discharge of the Fraser River into the ocean is approximately 0.31 km^3/day, or 0.26% of the total flow of all rivers.

Exercise 13.1 How long does water stay in the atmosphere and oceans?

The residence time of a water molecule in the atmosphere (or any of the other reservoirs) can be estimated by dividing the amount that is there by the rate at which it is transferred in and out. For the atmosphere, we know that the reservoir size is 13,000 km^3, and the rate of flux is 1,580 km^3 per day. That's the amount that goes into or comes out of the atmosphere in a day. If we divide 13,000 by 1,580, we get 8.22 days. This means that, *on average*, a molecule of water stays in the atmosphere for just over eight days. "Average" needs to be emphasized here because obviously some molecules stay in the air for only a few hours, while others may stay up there for weeks.

The volume of the oceans is 1,338,000,000 km^3 and the flux rate is approximately the same (1,580 km^3 per day). What is the average residence time of a water molecule in the ocean?

See Appendix 3 for Exercise 13.1 answers.

Media Attributions

- Figure 17.2.1.1: "Water Cycle Blank" © Ingwik. Adapted by Steven Earle. CC BY-SA.
- Figure 17.2.1.2: © Steven Earle. CC BY. Based on data from The Water Cycle: Freshwater Storage.

This page titled 17.2.1: The Hydrological Cycle is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Steven Earle (BCCampus) via source content that was edited to the style and standards of the LibreTexts platform.

- 13.1: The Hydrological Cycle by Steven Earle is licensed CC BY 4.0. Original source: https://opentextbc.ca/physicalgeology2ed.

17.2.2: Drainage Basins

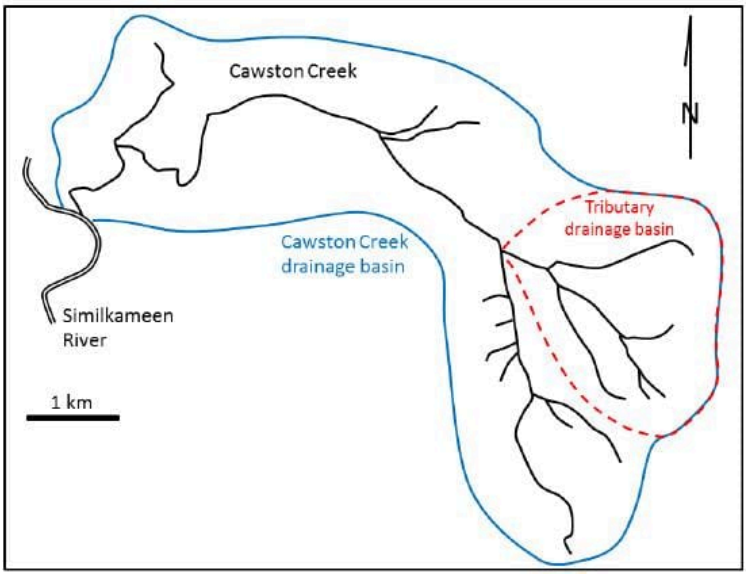


Figure 17.2.2.1 Cawston Creek near Keremeos, B.C. The blue line shows the extent of the drainage basin. The dashed red line is the drainage basin of one of its tributaries.

A **stream** is a body of flowing surface water of any size, ranging from a tiny trickle to a mighty river. The area from which the water flows to form a stream is known as its **drainage basin**. All of the precipitation (rain or snow) that falls within a drainage basin eventually flows into its stream, unless some of that water is able to cross into an adjacent drainage basin via groundwater flow. An example of a drainage basin is shown in Figure 17.2.2.1.

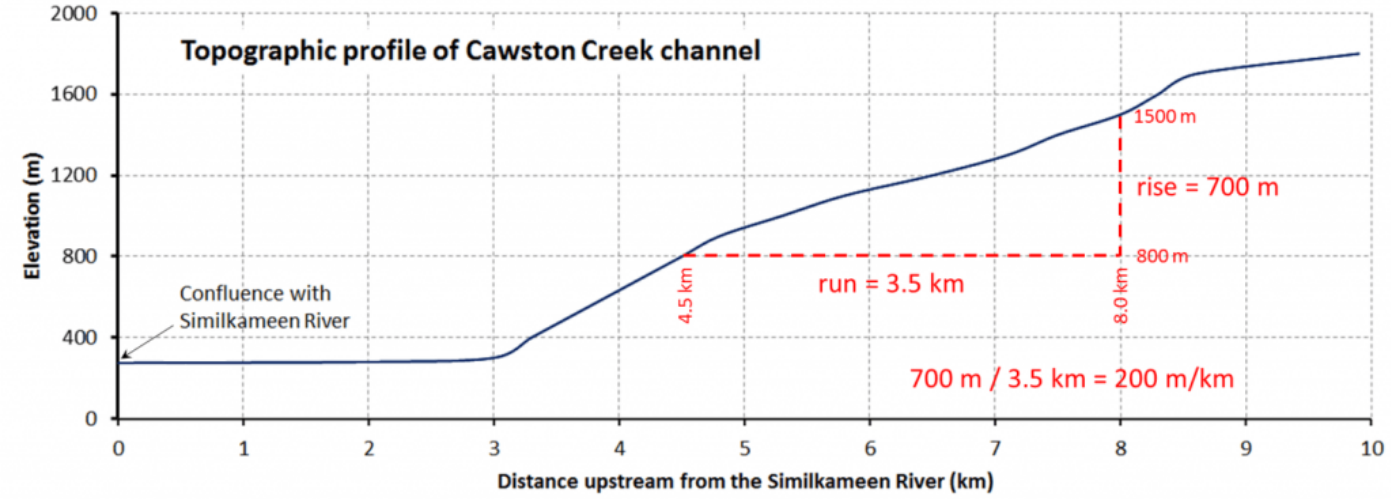


Figure 17.2.2.2 Profile of the main stem of Cawston Creek near Keremeos, B.C. The maximum elevation of the drainage basin is about 1,840 meters, near Mount Kobau. The base level is 275 meters, at the Similkameen River. As shown, the gradient of the stream can be determined by dividing the change in elevation between any two points (rise) by the distance between those two points (run).

Cawston Creek is a typical small drainage basin (approximately 25 square kilometers) within a very steep glaciated valley. As shown in Figure 17.2.2.2, the upper and middle parts of the creek have steep **gradients** (averaging about 200 meters per kilometre but ranging from 100 to 350 meters per kilometre), and the lower part, within the valley of the Similkameen River, is relatively flat (less than 5 meters per kilometre). The shape of the valley has been controlled first by tectonic uplift (related to plate convergence), then by pre-glacial stream erosion and mass wasting, then by several episodes of glacial erosion, and finally by post-glacial stream erosion. The lowest elevation of Cawston Creek (275 meters at the Similkameen River) is its **base level**. Cawston Creek cannot erode below that level unless the Similkameen River erodes deeper into its flood plain (the area that is

inundated during a flood).

Metro Vancouver's water supply comes from three large drainage basins on the north shore of Burrard Inlet, as shown in Figure 17.2.2.3. This map illustrates the concept of a drainage basin divide. The boundary between two drainage basins is the height of land between them. A drop of water falling on the boundary between the Capilano and Seymour drainage basins (a.k.a., watersheds), for example, could flow into either one of them.



Figure 17.2.2.3 The three drainage basins that are used for the Metro Vancouver water supply.

The pattern of tributaries within a drainage basin depends largely on the type of rock beneath, and on structures within that rock (folds, fractures, faults, etc.). The three main types of drainage patterns are illustrated in Figure 17.2.2.4. **Dendritic** patterns, which are by far the most common, develop in areas where the rock (or unconsolidated material) beneath the stream has no particular fabric or structure and can be eroded equally easily in all directions. Examples would be granite, gneiss, volcanic rock, and sedimentary rock that has not been folded. Most areas of British Columbia have dendritic patterns, as do most areas of the prairies and the Canadian Shield. **Trellis** drainage patterns typically develop where sedimentary rocks have been folded or tilted and then eroded to varying degrees depending on their strength. The Rocky Mountains of B.C. and Alberta are a good example of this, and many of the drainage systems within the Rockies have trellis patterns. **Rectangular** patterns develop in areas that have very little topography and a system of bedding planes, fractures, or faults that form a rectangular network. Rectangular drainage patterns are rare in Canada.

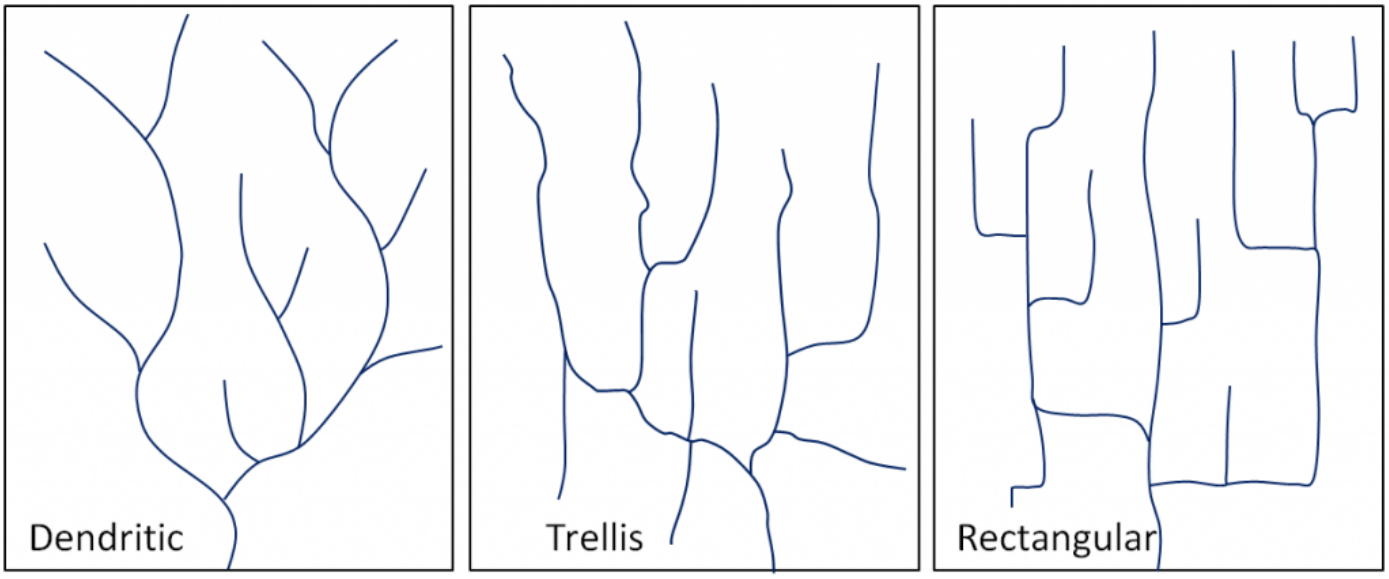


Figure 17.2.2.4 Typical dendritic, trellis, and rectangular stream drainage patterns.

In many parts of Canada, especially relatively flat areas with thick glacial sediments, and throughout much of Canadian Shield in eastern and central Canada, drainage patterns are chaotic, or what is known as **deranged** (Figure 17.2.2.5, left). Lakes and wetlands are common in this type of environment.

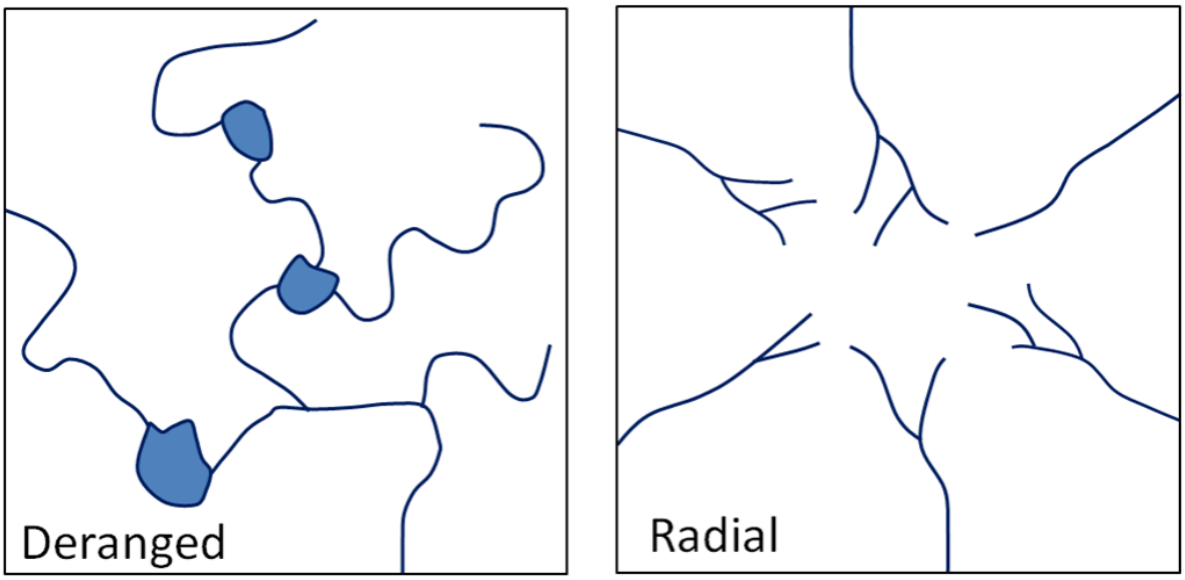


Figure 17.2.2.5 Left: a typical deranged pattern; right: a typical radial drainage pattern developed around a mountain or hill.

A fourth type of drainage pattern, which is not specific to a drainage basin, is known as **radial** (Figure 17.2.2.5, right). Radial patterns form around isolated mountains (such as volcanoes) or hills, and the individual streams typically have dendritic drainage patterns.

Over geological time, a stream will erode its drainage basin into a smooth profile similar to that shown in Figure 17.2.2.6. If we compare this with an ungraded stream like Cawston Creek (Figure 17.2.2.2), we can see that graded streams are steepest in their headwaters and their gradient gradually decreases toward their mouths. Ungraded streams have steep sections at various points, and typically have rapids and waterfalls at numerous locations along their lengths.

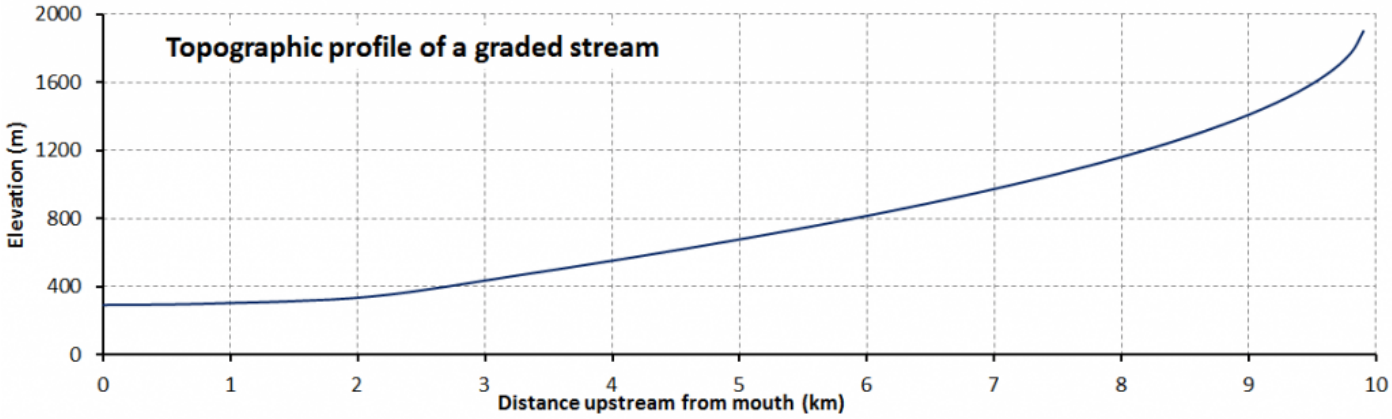


Figure 17.2.2.6 The topographic profile of a typical graded stream.

A graded stream can become ungraded if there is renewed tectonic uplift, or if there is a change in the base level, either because of tectonic uplift or some other reason. As stated earlier, the base level of Cawston Creek is defined by the level of the Similkameen River, but this can change, and has done so in the past. Figure 17.2.2.7 shows the valley of the Similkameen River in the Keremeos area. The river channel is just beyond the row of trees. The green field in the distance is underlain by material eroded from the hills behind and deposited by a small creek (not Cawston Creek) adjacent to the Similkameen River when its level was higher than it is now. Sometime in the past several centuries, the Similkameen River eroded down through these deposits (forming the steep bank on the other side of the river), and the base level of the small creek was lowered by about 10 meters. Over the next few centuries, this creek will seek to become graded again by eroding down through its own alluvial fan.

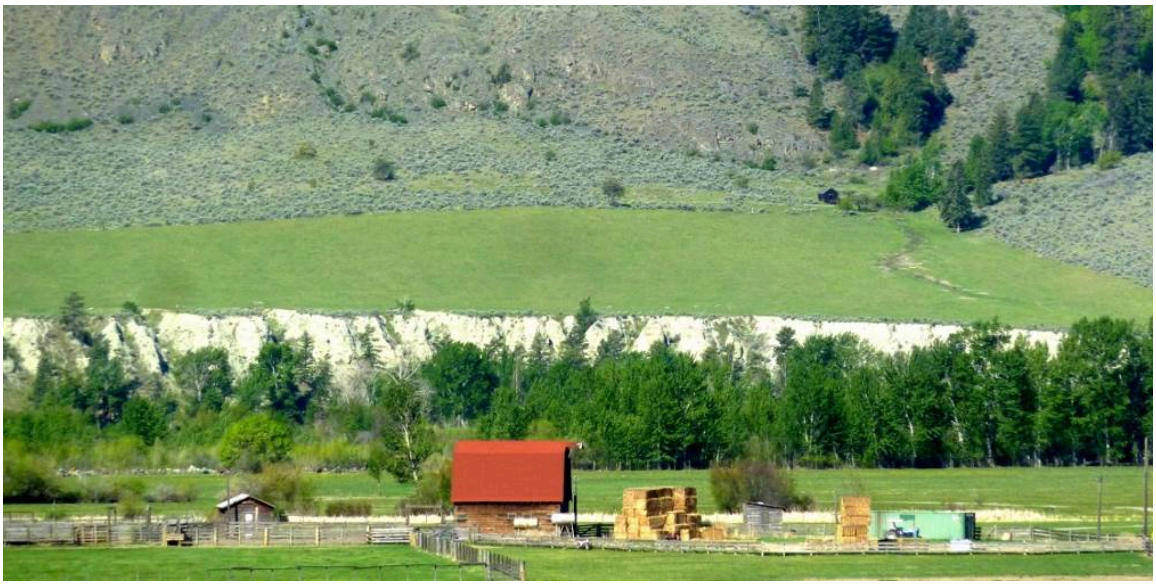


Figure 17.2.2.7 An example of a change in the base level of a small stream that flows into the Similkameen river near Keremeos. The previous base level was near the top of the sandy bank. The current base level is the river (hidden behind the trees) which has eroded down into the sandy deposit.

Another example of a change in base level can be seen along the Juan de Fuca Trail on southwestern Vancouver Island. As shown in Figure 17.2.2.8, many of the small streams along this part of the coast flow into the ocean as waterfalls. It is evident that the land in this area has risen by about 5 meters in the past few thousand years, probably in response to deglaciation. The streams that used to flow directly into the ocean now have a lot of down-cutting to do to become regraded.



Figure 17.2.2.8 Two streams with a lowered base level on the Juan de Fuca Trail, southwestern Vancouver Island.

The ocean is the ultimate base level, but lakes and other rivers act as base levels for many smaller streams. We can create an artificial base level on a stream by constructing a dam, as illustrated in Exercise 13.2.

Exercise 13.2 The effect of a dam on base level



Figure 17.2.2.9 Revelstoke Dam and Revelstoke Lake on the Columbia River at Revelstoke, BC.

When a dam is built on a stream, a reservoir (artificial lake) forms behind the dam. This temporarily (for many decades at least) creates a new base level for the part of the stream above the reservoir.

How does the formation of a reservoir affect the stream where it enters the reservoir, and what happens to the sediment it was carrying?

The water leaving the dam has no sediment in it. How does this affect the stream below the dam?

See Appendix 3 for Exercise 13.2 answers.

Sediments accumulate within the flood plain of a stream, and then, if the base level falls, or if there is less sediment to deposit, the stream may cut down through those existing sediments to form terraces. A terrace on the Similkameen River is shown in Figure 17.2.2.7 and some on the Fraser River are shown in Figure 17.2.2.10. The Fraser River photo shows at least two levels of terraces.



Figure 17.2.2.10 Terraces on the Fraser River at High Bar.

In the late 19th century, American geologist William Davis proposed that streams and the surrounding terrain develop in a cycle of erosion (Figure 17.2.2.11). Following tectonic uplift, streams erode quickly, developing deep V-shaped valleys that tend to follow relatively straight paths. Gradients are high, and profiles are ungraded. Rapids and waterfalls are common. During the mature stage, streams erode wider valleys and start to deposit thick sediment layers. Gradients are slowly reduced and grading increases. In old age, streams are surrounded by rolling hills, and they occupy wide sediment-filled valleys. Meandering patterns are common.

Davis's work was done long before the idea of plate tectonics, and he was not familiar with the impacts of glacial erosion on streams and their environments. While some parts of his theory are out of date, it is still a useful way to understand streams and their evolution.

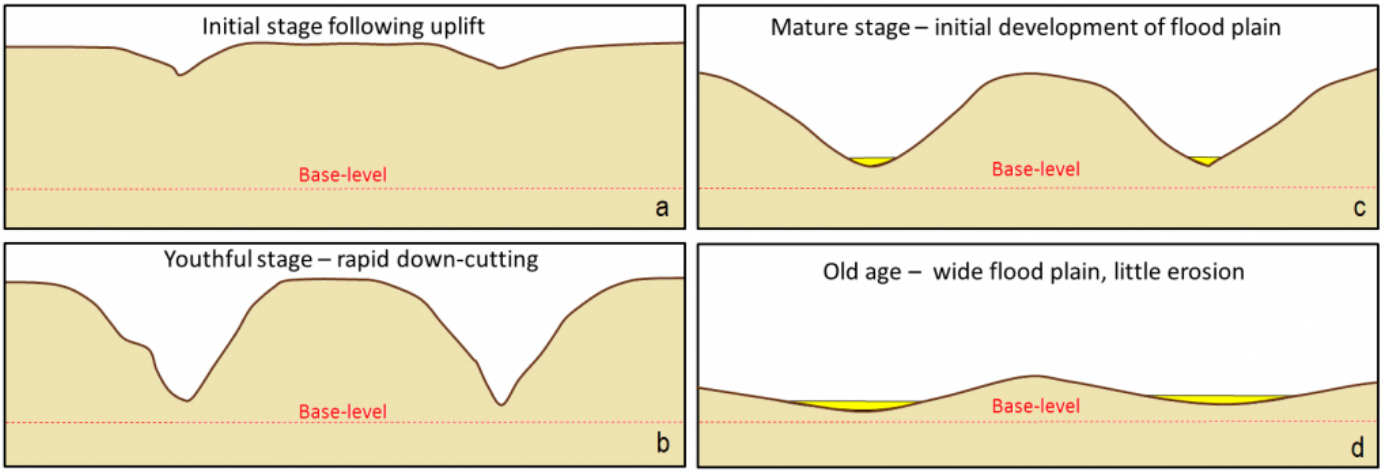


Figure 17.2.2.11 A depiction of the Davis cycle of erosion: a: initial stage, b: youthful stage, c: mature stage, and d: old age.

Image Descriptions

Figure 17.2.2.1 image description: Cawston Creek is a drainage basin made up of a number of small streams that flow into the Similkameen River. The Cawston Creek drainage basin also has a smaller tributary drainage basin that flows into it. [Return to Figure 17.2.2.1]

Figure 17.2.2.2 image description: The graph shows the topographical profile of the Cawston Creek channel. Starting at 10 kilometers from the Similkameen River, the channel is at an elevation of 1800 meters. It flows downhill towards the river at a

steady decline for seven kilometers. At three kilometers from the Similkameen River at an elevation of 275 meters, the channel levels out and stays at that elevation until it joins with the Similkameen River. This is conveyed in the following table.

Distance from the Similkameen River (kilometers)	Elevation (meters)
10	1,800
9	1,750
8	1,500
7	1,300
6	1,125
5	950
4	600
3	280
2	275
1	275
0	275

To determine the gradient of the stream, you must divide the change in elevation (rise) by the change in distance (run). For example, the elevation at 4.5 kilometers from the river is 800 meters and at 8.0 kilometers from the river, the elevation is 1500 meters. That means that in 3.5 kilometers (run), elevation increased by 700 meters (rise). To determine the gradient, you must divide rise (700 meters) by run (3.5 kilometers), which equals a vertical increase of 200 meters per kilometre.

[Return to Figure 17.2.2.2]

Figure 17.2.2.4 image description: Three types of drainage patterns:

- Dendritic: a pattern of drainage channels that resembles the branches in a tree.
- Trellis: a drainage pattern in which tributaries typically flow parallel to one other but meet at right angles.
- Rectangular: a drainage pattern in which tributaries typically flow at right angles to each other and meet at right angles.

[Return to Figure 17.2.2.4]

Figure 17.2.2.5 image description: Two types of drainage patterns:

- Deranged: a pattern of drainage channels that is chaotic.
- Radial: a pattern of streams radiating out from a central point, typically an isolated mountain.

[Return to Figure 17.2.2.5]

Figure 17.2.2.6 image description: The topographic profile of a typical graded stream. While the topographical profile of Cawston Creek fell at a steady rate and then leveled out for the last three kilometers, the topographical profile of a typical graded stream starts out steep, but the slope (or gradient) gradually decreases towards the bottom to form a soft curve. [Return to Figure 17.2.2.6]

Figure 17.2.2.10 image description: The Fraser River snaking through a channel that it has worn down over the years. The sides of the bank rise steeply, level out, and then rise again to produce terraces. [Return to Figure 17.2.2.10]

Figure 17.2.2.11 image description: William Davis's cycle of erosion.

1. Initial stage: A tectonic uplift forms two sharp valleys on each side.
2. Youthful stage: Steams flowing through those valleys erode the surrounding earth quickly and cut deep to form sharp V's.
3. Mature stage: The streams become wider and the banks more gradual. This marks the initial development of a flood plain.
4. Old age: Steams for a wide flood plain where little erosion occurs.

[Return to Figure 17.2.2.11]

Media Attributions

- Figures 13.2.1, 13.2.2, 13.2.4, 13.2.5, 13.2.6, 13.2.7, 13.2.8, 13.2.9, 13.2.11: © Steven Earle. CC BY.
- Figure 17.2.2.3: Image by Metro Vancouver. Used with permission.
- Figure 17.2.2.10: Image by Marie Betcher. Used with permission.

This page titled 17.2.2: Drainage Basins is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Steven Earle (BCCampus) via source content that was edited to the style and standards of the LibreTexts platform.

- 13.2: Drainage Basins by Steven Earle is licensed CC BY 4.0. Original source: https://opentextbc.ca/physicalgeology2ed.

17.2.3: Stream Erosion and Deposition

As we discussed in Chapter 6, flowing water is a very important mechanism for erosion, transportation and deposition of sediments. Water flow in a stream is primarily related to the stream's gradient, but it is also controlled by the geometry of the stream channel. As shown in Figure 17.2.3.1, water flow velocity is decreased by friction along the stream bed, so it is slowest at the bottom and edges and fastest near the surface and in the middle. In fact, the velocity just below the surface is typically a little higher than right at the surface because of friction between the water and the air. On a curved section of a stream, flow is fastest on the outside and slowest on the inside.

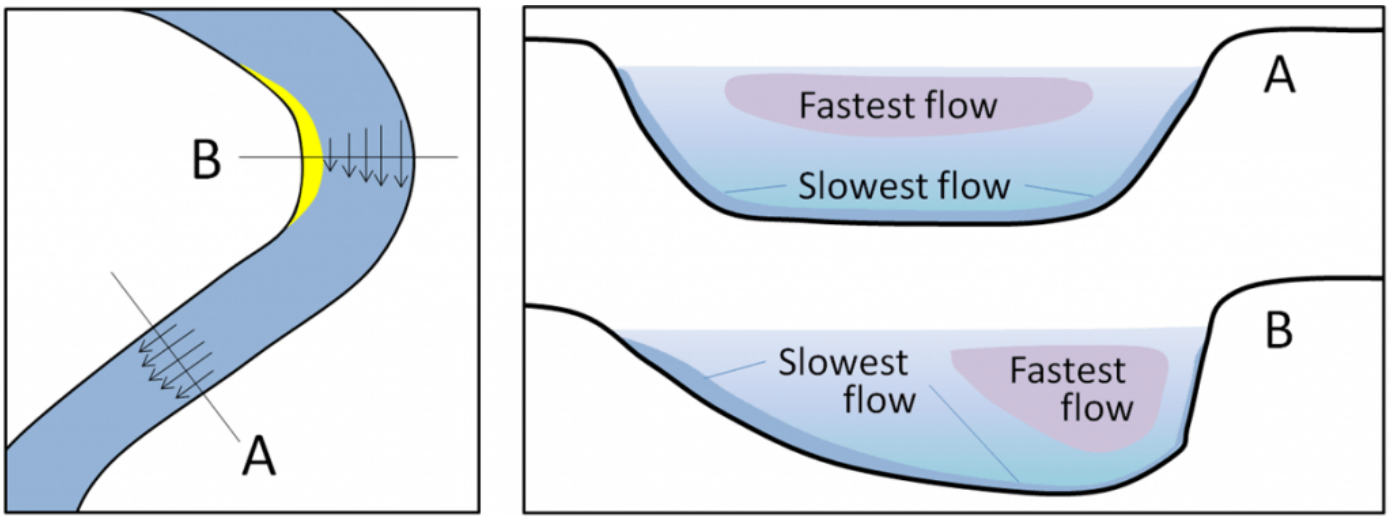


Figure 17.2.3.1 The relative velocity of stream flow depending on whether the stream channel is straight or curved (left), and with respect to the water depth (right).

Other factors that affect stream-water velocity are the size of sediments on the stream bed—because large particles tend to slow the flow more than small ones—and the **discharge**, or volume of water passing a point in a unit of time (e.g., cubic meters (m^3) per second). During a flood, the water level always rises, so there is more cross-sectional area for the water to flow in, however, as long as a river remains confined to its channel, the velocity of the water flow also increases.

Figure 17.2.3.2 shows the nature of sediment transportation in a stream. Large particles rest on the bottom—**bed load**—and may only be moved during rapid flows under flood conditions. They can be moved by **saltation** (bouncing) and by **traction** (being pushed along by the force of the flow).

Smaller particles may rest on the bottom some of the time, where they can be moved by saltation and traction, but they can also be held in suspension in the flowing water, especially at higher velocities. As you know from intuition and from experience, streams that flow fast tend to be turbulent (flow paths are chaotic and the water surface appears rough) and the water may be muddy, while those that flow more slowly tend to have laminar flow (straight-line flow and a smooth water surface) and clear water. Turbulent flow is more effective than laminar flow at keeping sediments in suspension.

Stream water also has a dissolved load, which represents (on average) about 15% of the mass of material transported, and includes ions such as calcium (Ca^{+2}) and chloride (Cl^{-}) in solution. The solubility of these ions is not affected by flow velocity.

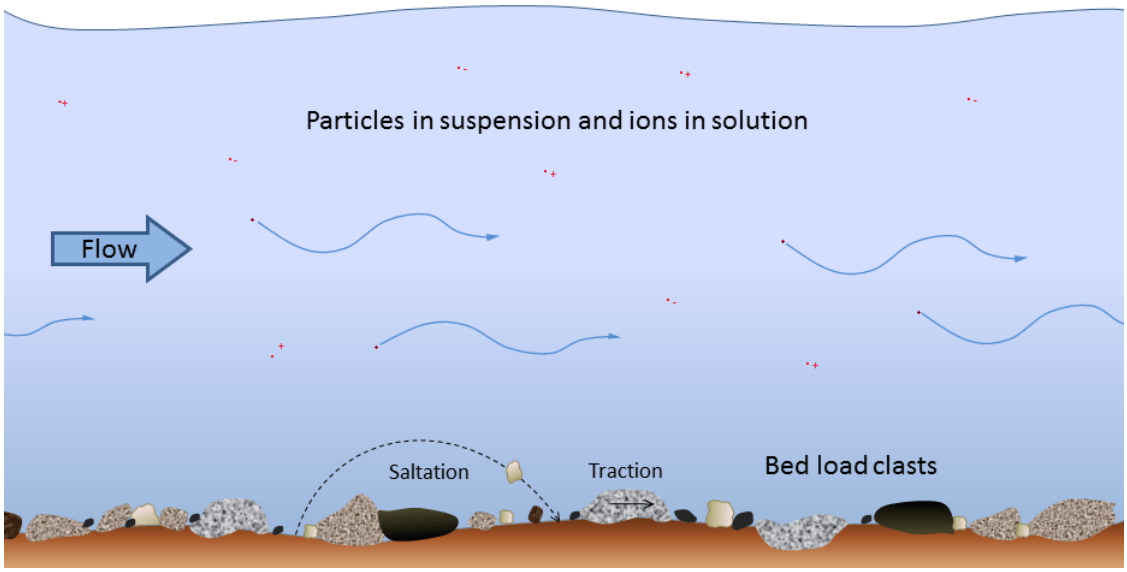


Figure 17.2.3.2 Modes of transportation of sediments and dissolved ions (represented by red dots with + and – signs) in a stream.

The faster the water is flowing, the larger the particles that can be kept in suspension and transported within the flowing water. However, as Swedish geographer Filip Hjulström discovered in the 1940s, the relationship between grain size and the likelihood of a grain being eroded, transported, or deposited is not as simple as one might imagine (Figure 17.2.3.3). Consider, for example, a 1 millimetre grain of sand. If it is resting on the bottom, it will remain there until the velocity is high enough to erode it, around 20 centimeters per second (cm/s). But once it is in suspension, that same 1 mm particle will remain in suspension as long as the velocity doesn't drop below 10 cm/s. For a 10 mm gravel grain, the velocity is 105 cm/s to be eroded from the bed but only 80 cm/s to remain in suspension.

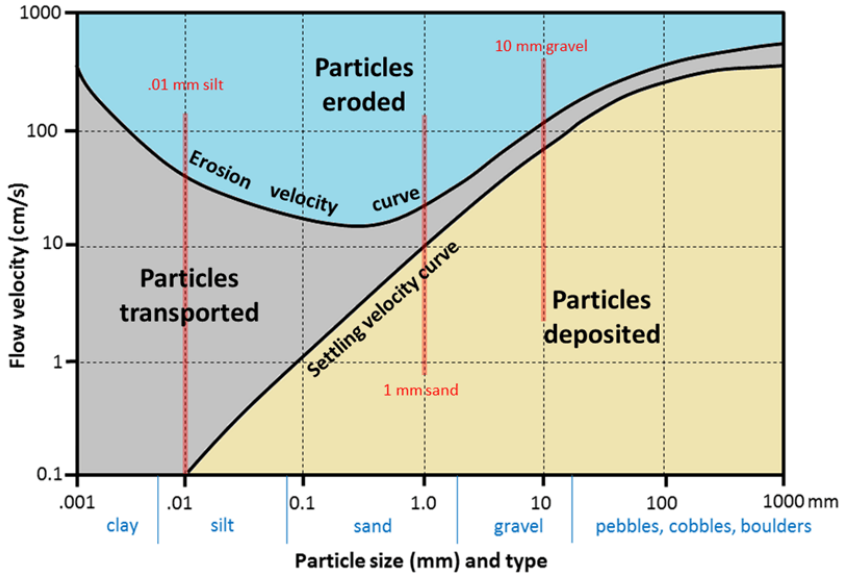


Figure 17.2.3.3 The Hjulström-Sundborg diagram showing the relationships between particle size and the tendency to be eroded, transported, or deposited at different current velocities.

On the other hand, a 0.01 mm silt particle only needs a velocity of 0.1 centimeters per second (cm/s) to remain in suspension, but requires 60 cm/s to be eroded. In other words, a tiny silt grain requires a greater velocity to be eroded than a grain of sand that is 100 times larger! For clay-sized particles, the discrepancy is even greater. In a stream, the most easily eroded particles are small sand grains between 0.2 mm and 0.5 mm. Anything smaller or larger requires a higher water velocity to be eroded and entrained in the flow. The main reason for this is that small particles, and especially the tiny grains of clay, have a strong tendency to stick together, and so are difficult to erode from the stream bed.

It is important to be aware that a stream can both erode and deposit sediments at the same time. At 100 cm/s, for example,

silt, sand, and medium gravel will be eroded from the stream bed and transported in suspension, coarse gravel will be held in suspension, pebbles will be both transported and deposited, and cobbles and boulders will remain stationary on the stream bed.

Exercise 13.3 Understanding the Hjulström-Sundborg diagram

Refer to the Hjulström-Sundborg diagram (Figure 17.2.3.3) to answer these questions.

1. A fine sand grain (0.1 millimeters) is resting on the bottom of a stream bed.
    1. What stream velocity will it take to get that sand grain into suspension?
    2. Once the particle is in suspension, the velocity starts to drop. At what velocity will it finally come back to rest on the stream bed?
2. A stream is flowing at 10 centimeters per second (which means it takes 10 seconds to go 1 meters, and that's pretty slow).
    1. What size of particles can be eroded at 10 centimeters per second?
    2. What is the largest particle that, once already in suspension, will remain in suspension at 10 centimeters per second?

See Appendix 3 for Exercise 13.3 answers.

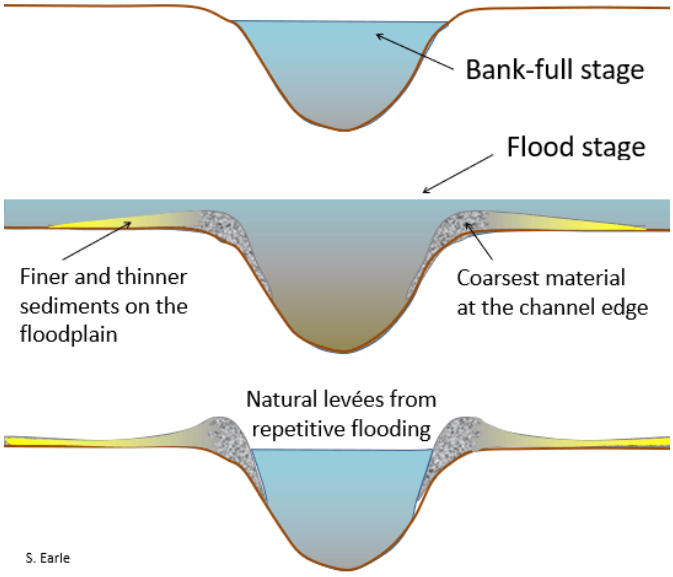


Figure 17.2.3.4 The development of natural levées during flooding of a stream. The sediments of the levée become increasingly fine away from the stream channel, and even finer sediments—clay, silt, and fine sand—are deposited across most of the flood plain.

A stream typically reaches its greatest velocity when it is close to flooding over its banks. This is known as the bank-full stage, as shown in Figure 17.2.3.4. As soon as the flooding stream overtops its banks and occupies the wide area of its flood plain, the water has a much larger area to flow through and the velocity drops significantly. At this point, sediment that was being carried by the high-velocity water is deposited near the edge of the channel, forming a natural bank or **levée**.

Image Descriptions

Figure 17.2.3.1 image description: When a stream curves, the flow of water is fastest on the outside of the curve and slowest on the inside of the curve. When the stream is straight and a uniform depth, the stream flows fastest in the middle near the top and slowest along the edges. When the depth is not uniform, the stream flows fastest in the deeper section. [Return to Figure 17.2.3.1]

Figure 17.2.3.3 image description:

- Erosion velocity curve: A 0.001 millimetre particle would erode at a flow velocity of 500 centimeters per second or greater. As the particle size gets larger, the minimum flow velocity needed to erode the particle decreases, with the lowest flow velocity being 30 centimeters per second to erode a 0.5 millimetre particle. To erode particles larger than 0.5 millimeters,

the minimum flow velocity rises again.
- Settling velocity curve: A 0.01 millimetre particle would be deposited with a flow velocity of 0.1 centimetre per second or less. As the flow velocity increases, only larger and larger particles will be deposited.
- Particles between these two curves (either moving too slow or being too small to be eroded or deposited) will be transported in the stream.

[Return to Figure 17.2.3.3]

Media Attributions

- Figure 17.2.3.1, 13.3.2, 13.3.4: © Steven Earle. CC BY.

This page titled 17.2.3: Stream Erosion and Deposition is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Steven Earle (BCCampus) via source content that was edited to the style and standards of the LibreTexts platform.

- 13.3: Stream Erosion and Deposition by Steven Earle is licensed CC BY 4.0. Original source: https://opentextbc.ca/physicalgeology2ed.

17.2.4: Stream Types

Stream channels can be straight or curved, deep and slow, or rapid and choked with coarse sediments. The cycle of erosion has some influence on the nature of a stream, but there are several other factors that are important.



Figure 17.2.4.1 The Cascade Falls area of the Kettle River, near Christina Lake, B.C. This stream has a step-pool morphology and a deep bedrock channel.

Youthful streams that are actively down-cutting their channels tend to be relatively straight and are typically ungraded (meaning that rapids and falls are common). As shown in Figures 13.0.1 and 13.4.1, youthful streams commonly have a **step-pool** morphology, meaning that the stream consists of a series of pools connected by rapids and waterfalls. They also have steep gradients and steep and narrow V-shaped valleys—in some cases steep enough to be called canyons.

In mountainous terrain, such as that in western Alberta and B.C., steep youthful streams typically flow into wide and relatively low-gradient U-shaped glaciated valleys. The youthful streams have high sediment loads, and when they flow into the lower-gradient glacial valleys where the velocity isn't high enough to carry all of the sediment **braided** patterns develop, characterized by a series of narrow channels separated by gravel bars (Figure 17.2.4.2).

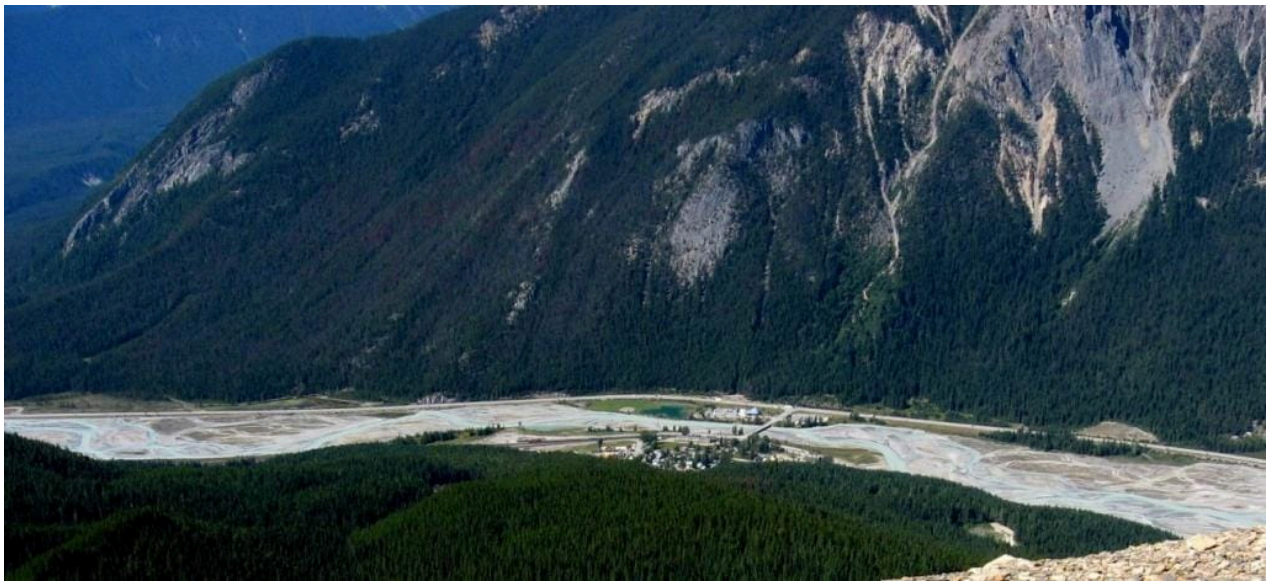


Figure 17.2.4.2 The braided channel of the Kicking Horse River at Field, B.C.



Figure 17.2.4.3 The braided channel of Meager Creek in the Mt. Meager area

Braided streams can develop anywhere there is more sediment than a stream is able to transport. One such environment is in volcanic regions, where explosive eruptions produce large amounts of unconsolidated material that gets washed into streams. Streams in the volcanic Mt. Meager area of southwestern British Columbia are good examples of this (Figure 17.2.4.3).

A stream that occupies a wide, flat flood plain with a low gradient typically carries only sand-sized and finer sediments and develops a sinuous flow pattern. As you saw in Figure 17.2.4.1, when a stream flows around a corner, the water on the outside has farther to go and tends to flow faster. This leads to erosion of the banks on the outside of the curve, deposition on the inside, and formation of a point bar (Figure 17.2.4.4). Over time, the sinuosity of the stream becomes increasingly exaggerated, and the channel migrates around within its flood plain, forming a **meandering** pattern.



Figure 17.2.4.4 The meandering channel of the Bonnell Creek, Nanoose, B.C. The stream is flowing toward the viewer. The sand and gravel point bar must have formed when the creek was higher and the flow faster than it was when the photo was taken.

A well-developed meandering river is shown in Figure 17.2.4.5. The meander in the middle of the photo has reached the point where the thin neck of land between two parts of the channel is about to be eroded through. When this happens, another **oxbow** lake will form like the others in the photo.



Figure 17.2.4.5 The meandering channel of the Nowitna River, Alaska. Numerous oxbow lakes are present and another meander cutoff will soon take place.

Exercise 13.4 Determining stream gradients

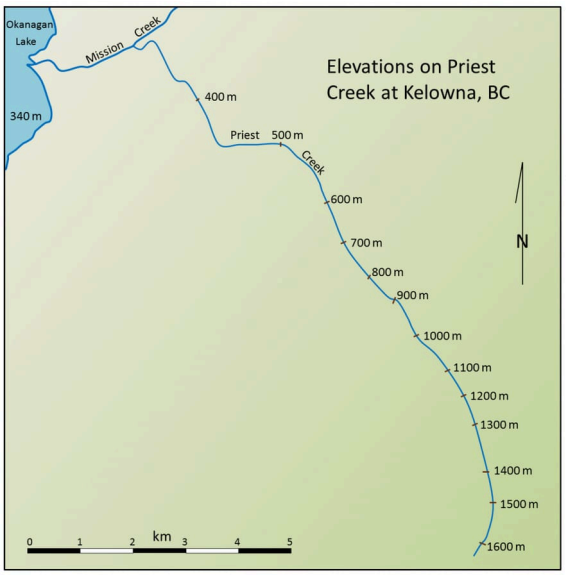


Figure 17.2.4.6 Elevations on Priest Creek at Kelowna, BC.

Gradient is the key factor controlling stream velocity, and of course, velocity controls sediment erosion and deposition. This map shows the elevations of Priest Creek in the Kelowna area. The length of the creek between 1,600 meters and 1,300 meters elevation is 2.4 kilometers, so the gradient is 300/2.4 = 125 meters per kilometre.

1. Use the scale bar to estimate the distance between 1,300 meters and 600 meters and then calculate that gradient.
2. Estimate the gradient between 600 and 400 meters.
3. Estimate the gradient between 400 meters on Priest Creek and the point where Mission Creek enters Okanagan Lake.

See Appendix 3 for Exercise 13.4 answers.

At the point where a stream enters a still body of water—a lake or the ocean—sediment is deposited and a delta forms. The Fraser River has created a large delta, which extends out into the Strait of Georgia (Figure 17.2.4.7). Much of the Fraser delta is very young in geological terms. Shortly after the end of the last glaciation (10,000 years ago), the delta did not extend past New Westminster. Since that time, all of the land that makes up Richmond, Delta, and parts of New Westminster and south Surrey has formed from sediment from the Fraser River. (You can see this in more detail at Geoscape Vancouver.)



Figure 17.2.4.7 The delta of the Fraser River and the plume of sediment that extends across the Strait of Georgia. The land outlined in red has formed over the past 10,000 years.

Image Descriptions

Figure 17.2.4.5 image description: A part of the Nowitna River has curved around so sharply that it almost forms a circle before curving the other way again. Eventually, as the barrier between these two parts of the channel erodes, they will be joined and form an oxbow lake. [Return to Figure 17.2.4.5]

Media Attributions

- Figures 13.4.1, 13.4.2, 13.4.4, 13.4.6: © Steven Earle. CC BY.
- Figure 17.2.4.3: "Meager Creek" © Isaac Earle. CC BY.
- Figure 17.2.4.5: "Nowitna river" by Oliver Kumis. CC BY-SA.
- Figure 17.2.4.7: Delta of the Fraser River by NASA. Taken September 2011. Adapted by Steven Earle. Public domain.

This page titled 17.2.4: Stream Types is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Steven Earle (BCCampus) via source content that was edited to the style and standards of the LibreTexts platform.

- 13.4: Stream Types by Steven Earle is licensed CC BY 4.0. Original source: https://opentextbc.ca/physicalgeology2ed.

17.2.5: Flooding

The discharge levels of streams are highly variable depending on the time of year and on specific variations in the weather from one year to the next. In Canada, most streams show discharge variability similar to that of the Stikine River in northwestern B.C., as illustrated in Figure 17.2.5.1. The Stikine River has its lowest discharge levels in the depths of winter when freezing conditions persist throughout most of its drainage basin. Discharge starts to rise slowly in May, and then rises dramatically through the late spring and early summer as a winter's worth of snow melts. For the year shown, the minimum discharge on the Stikine River was 56 cubic meters per second (m^3/s) in March, and the maximum was 37 times higher, 2,470 m^3/s, in May.

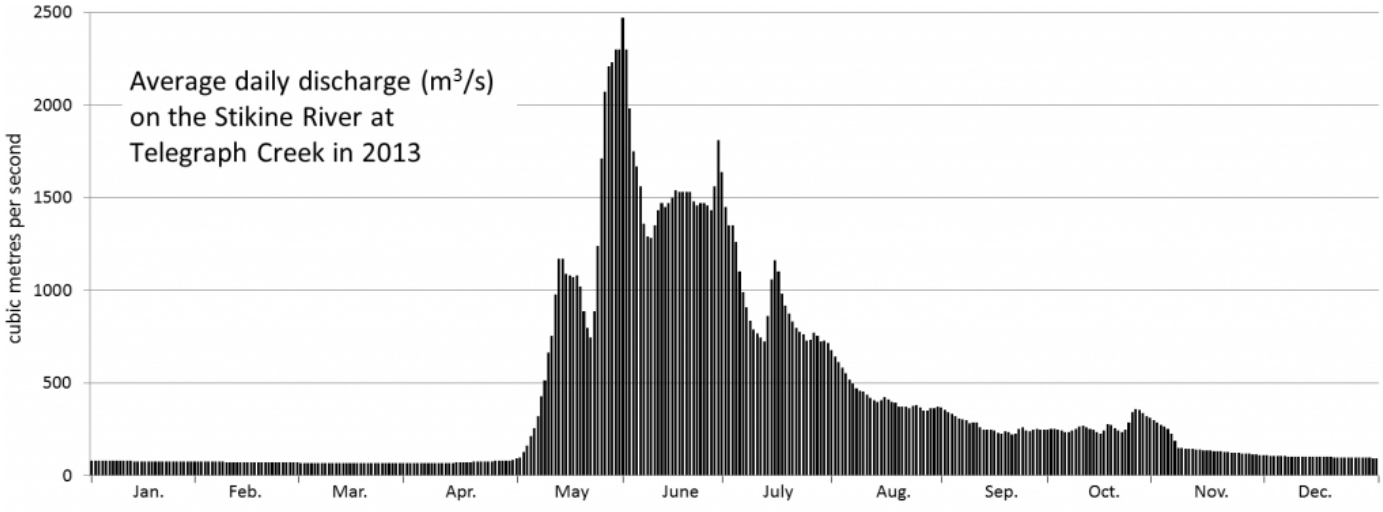


Figure 17.2.5.1 Variations in discharge of the Stikine River during 2013.

Streams in coastal areas of southern British Columbia show a very different pattern from those in most of the rest of the country because their drainage basins do not remain entirely frozen and because they receive a lot of rain (rather than snow) during the winter. The Qualicum River on Vancouver Island typically has its highest discharge levels in January or February and its lowest levels in late summer (Figure 17.2.5.2). In 2013, the minimum discharge was 1.6 m^3/s, in August, and the maximum was 34 times higher, 53 m^3/s, in March.

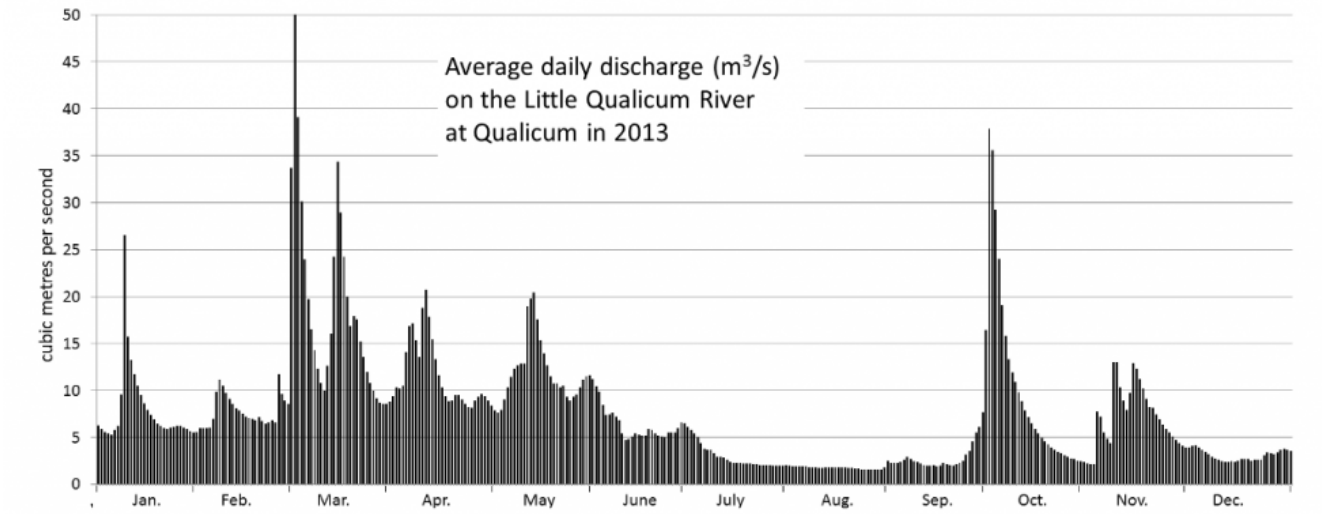


Figure 17.2.5.2 Variations in discharge of the Qualicum River during 2013.

When a stream's discharge increases, both the water level (**stage**) and the velocity increase as well. Rapidly flowing streams become muddy and large volumes of sediment are transported both in suspension and along the stream bed. In extreme

situations, the water level reaches the top of the stream's banks (the **bank-full stage**, see Figure 17.2.5.4), and if it rises any more, it floods the surrounding terrain. In the case of mature or old-age streams, this could include a vast area of relatively flat ground known as a flood plain, which is the area that is typically covered with water during a major flood. Because fine river sediments are deposited on flood plains, they are ideally suited for agriculture, and thus are typically occupied by farms and residences, and in many cases, by towns or cities. Such infrastructure is highly vulnerable to damage from flooding, and the people that live and work there are at risk.

Most streams in Canada have the greatest risk of flooding in the late spring and early summer when stream discharges rise in response to melting snow. In some cases, this is exacerbated by spring storms. In years when melting is especially fast and/or spring storms are particularly intense, flooding can be very severe.

One of the worst floods in Canadian history took place in the Fraser Valley in late May and early June of 1948. The early spring of that year had been cold, and a large snow pack in the interior was slow to melt. In mid-May, temperatures rose quickly and melting was accelerated by rainfall. Fraser River discharge levels rose rapidly over several days during late May, and the dykes built to protect the valley were breached in a dozen places. Approximately one-third of the flood plain was inundated and many homes and other buildings were destroyed, but there were no deaths. The Fraser River flood of 1948, which was the highest in the past century, was followed by very high river levels in 1950 and 1972 and by relatively high levels several times since then, the most recent being 2007 (Table 13.1). In the years following 1948, millions of dollars were spent repairing and raising the existing dykes and building new ones; since then damage from flooding in the Fraser Valley has been relatively limited.

Table 13.1 Ranking of the maximum stage and discharge values for the Fraser River at Hope between 1948 and 2008. Typical discharge levels are around 1,000 m^3/s.[1]

Rank	Year	Month	Date	Stage (meters)	Discharge (cubic meters per second)
1	1948	May	31	11.0	15,200
2	1972	June	16	10.1	12,900
3	1950	June	20	9.9	12,500
4	1964	June	21	9.6	11,600
5	1997	June	5	9.5	11,300
6	1955	June	29	9.4	11,300
7	1999	June	22	9.4	11,000
8	2007	June	10	9.3	10,850
9	1974	June	22	9.3	10,800
10	2002	June	21	9.2	10,600

Serious flooding happened in July in 1996 in the Saguenay-Lac St. Jean region of Quebec. In this case, the floods were caused by two weeks of heavy rainfall followed by one day of exceptional rainfall. July 19 saw 270 millimeters of rain, equivalent to the region's normal rainfall for the entire month of July. Ten deaths were attributed to the Saguenay floods, and the economic toll was estimated at $1.5 billion.

Just a year after the Saguenay floods, the Red River in Minnesota, North Dakota, and Manitoba reached its highest level since 1826. As is typical for the Red River, the 1997 flooding was due to rapid snowmelt. Because of the south to north flow of the river, the flooding starts in Minnesota and North Dakota, where melting starts earlier, and builds toward the north. The residents of Manitoba had plenty of warning that the 1997 flood was coming because there was severe flooding at several locations on the U.S. side of the border.

After the 1950 Red River flood, the Manitoba government built a channel around the city of Winnipeg to reduce the potential of flooding in the city (Figure 17.2.5.3). Known as the Red River Floodway, the channel was completed in 1964 at a cost of $63 million. Since then it has been used many times to alleviate flooding in Winnipeg, and is estimated to have saved many billions of dollars in flood damage. The massive 1997 flood was almost too much for the floodway; in fact the amount of water diverted was greater than the designed capacity. The floodway has recently been expanded so that it can be used to divert more of the Red River's flow away from Winnipeg.

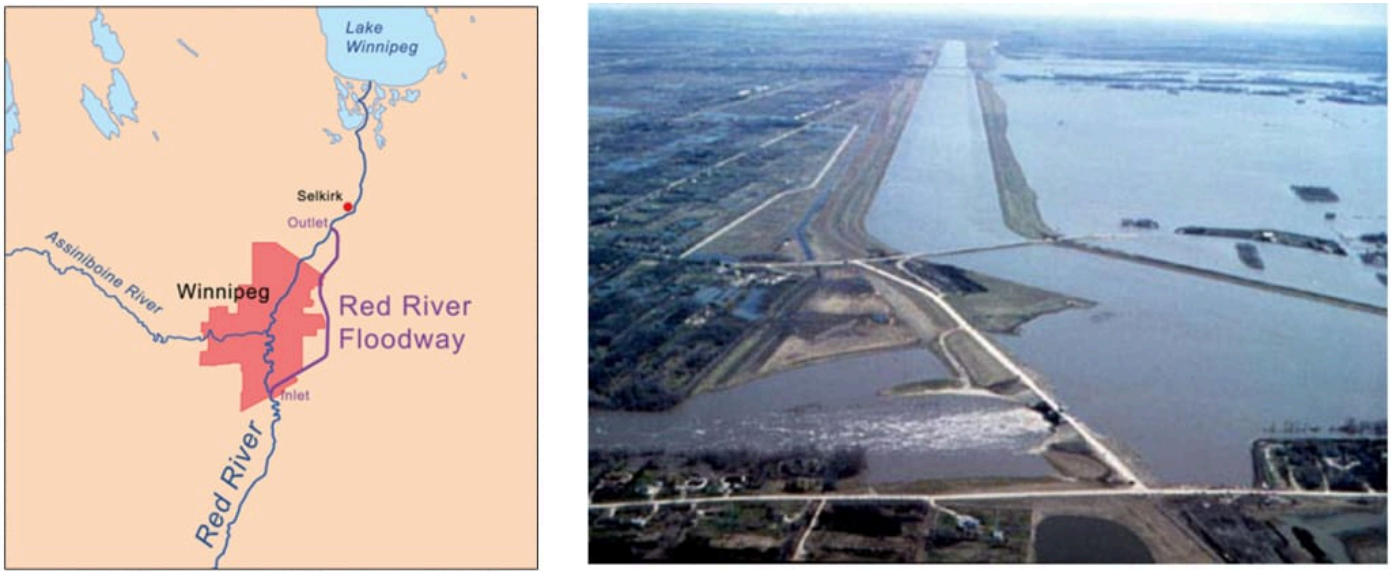


Figure 17.2.5.3 Map of the Red River floodway around Winnipeg, MB (left), and aerial view of the southern (inlet) end of the floodway (right).

Canada's most costly flood ever was the June 2013 flood in southern Alberta. The flooding was initiated by snowmelt and worsened by heavy rains in the Rockies due to an anomalous flow of moist air from the Pacific and the Caribbean. At Canmore, rainfall amounts exceeded 200 millimeters in 36 hours, and at High River, 325 millimeters of rain fell in 48 hours.

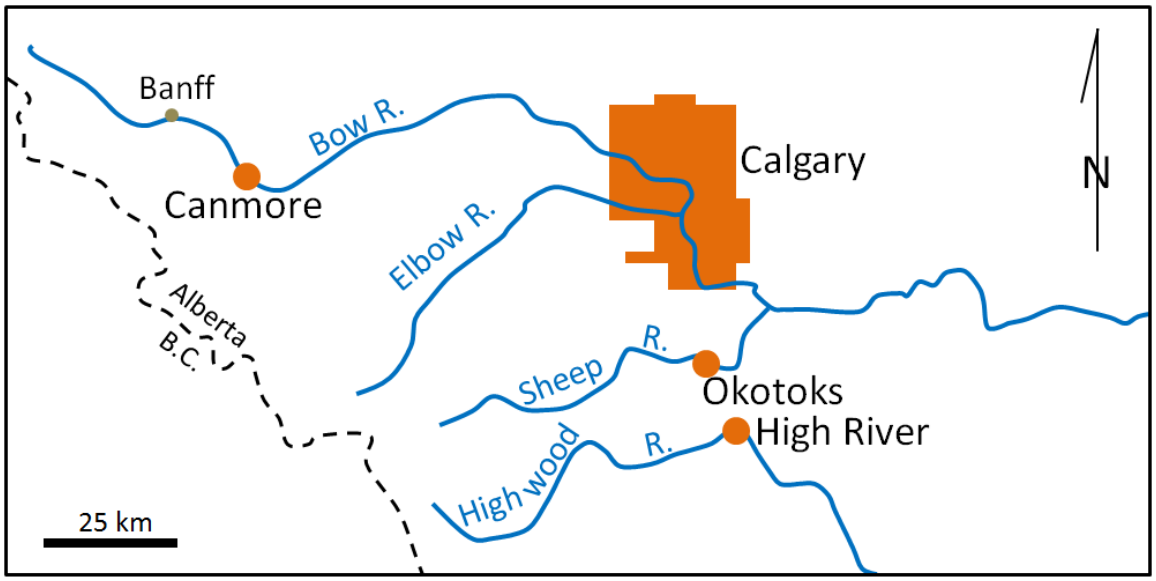


Figure 17.2.5.4 Map of the communities most affected by the 2013 Alberta floods (in orange).

In late June and early July, the discharges of several rivers in the area, including the Bow River in Banff, Canmore, and Exshaw, the Bow and Elbow Rivers in Calgary, the Sheep River in Okotoks, and the Highwood River in High River, reached levels that were 5 to 10 times higher than normal for the time of year (see Exercise 13.5). Large areas of Calgary, Okotoks, and High River were flooded and five people died (see Figures 13.5.2 and 13.5.3). The cost of the 2013 flood is estimated to be approximately $5 billion. Learn more about Alberta's flood of the century.

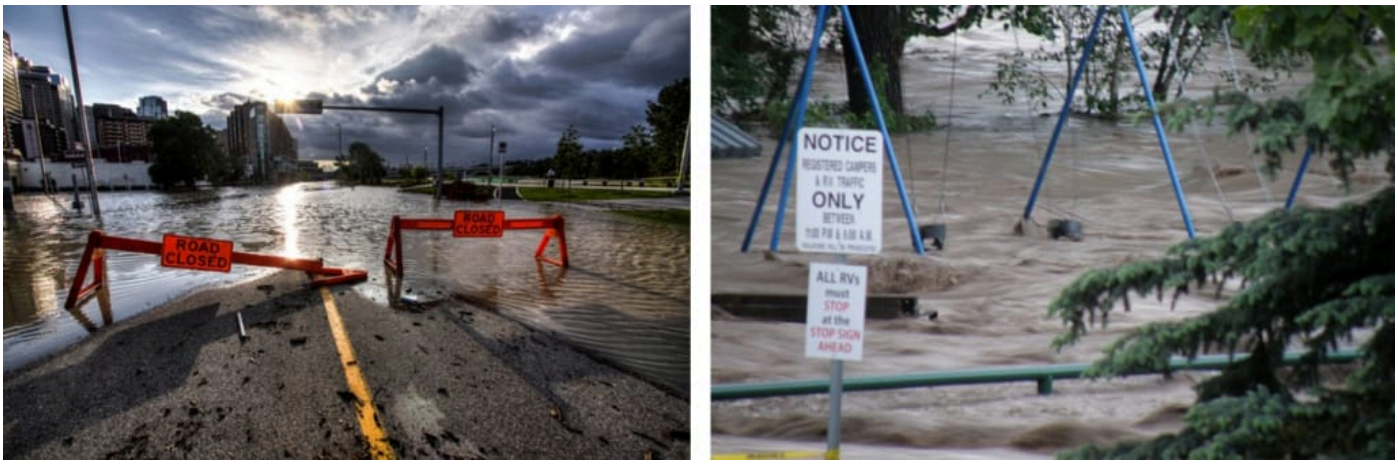


Figure 17.2.5.5 Flooding in Calgary (June 21, left) and Okotoks (June 20, right) during the 2013 southern Alberta flood.

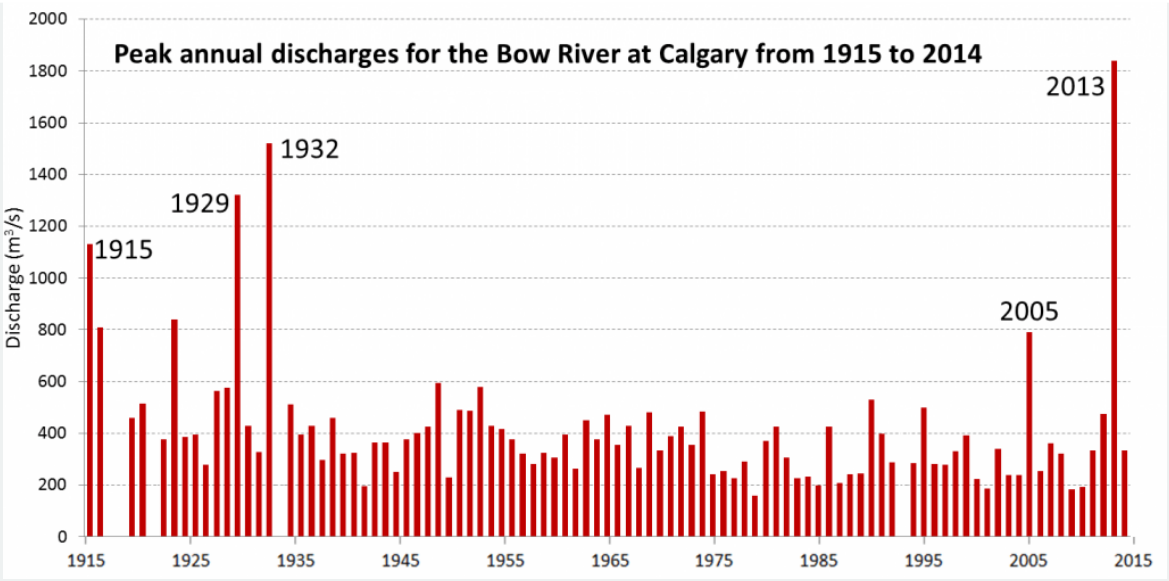


Figure 17.2.5.6 Graph of the highest discharge from the Bow River each year from 1915 to 2014.

Figure 17.2.5.6 shows the highest discharges per year between 1915 and 2014 on the Bow River at Calgary. Using this data set, we can calculate the recurrence interval (Ri) for any particular flood magnitude using the equation: Ri = (n+1)/r (where n is the number of floods in the record being considered, and r is the rank of the particular flood). There are a few years missing in this record, and the actual number of data points is 95.

The largest flood recorded on the Bow River over that period was the one in 2013, 1,840 cubic meters per second (m^3/s) on June 21. Ri for that flood is (95+1)/1 = 96 years. The probability of such a flood in any future year is 1/Ri, which is 1%. The fifth largest flood was just a few years earlier in 2005, at 791 m^3/s. Ri for that flood is (95+1)/5 = 19.2 years. The recurrence probability is 5%.

1. Calculate the recurrence interval for the second largest flood (1932, 1,520 m^3/s).
2. What is the probability that a flood of 1,520 m^3/s will happen next year?
3. Examine the 100-year trend for floods on the Bow River. If you ignore the major floods (the labelled ones), what is the general trend of peak discharges over that time?

See Appendix 3 for Exercise 13.5 answers.

One of the things that the 2013 flood on the Bow River teaches us is that we can't predict when a flood will occur or how big it will be, so in order to minimize damage and casualties we need to be prepared. Some of the ways of doing that are as follows:

- Mapping flood plains and not building within them
- Building dykes or dams where necessary
- Monitoring the winter snowpack, the weather, and stream discharges
- Creating emergency plans
- Educating the public

Image Descriptions

Figure 17.2.5.1 image description: The highest and lowest daily average discharge from the Stikine River by month in 2013 in cubic meters per second (m^3/s).

Month	Lowest daily average discharge (m^3/s)	Highest daily average discharge (m^3/s)
January	100	100
February	100	100
March	100	100
April	100	100
May	150	2,450
June	1,250	1,800
July	700	1,450
August	300	700
September	250	300
October	250	750
November	150	600
December	100	100

[Return to Figure 17.2.5.1]

Figure 17.2.5.2 image description: The highest and lowest daily average discharge from the Qualicum River by month in 2013 in cubic meters per second (m^3/s).

Month	Lowest daily average discharge (m^3/s)	Highest daily average discharge (m^3/s)
January	5	24
February	6	14
March	8	50
April	7	21
May	7	20
June	5	12
July	3	7
August	2	3
September	3	8
October	3	37
November	3	13
December	3	4

[Return to Figure 17.2.5.2]

Figure 17.2.5.6 image description: Peak annual discharge from the Bow River from 1915 to 2014 in cubic meters per second (m3/s)

Year	Peak Annual Discharge (m^3/s)
1915	1,125 (major)
1916	800
1919	475
1920	525
1922	375
1923	850
1924	400
1925	400
1926	275
1927	575
1928	580
1929	1,325 (major)
1930	425
1931	325
1932	1,520 (major)
1934	525
1935	400
1936	425
1937	300
1938	475
1939	325
1940	325
1941	200
1942	375
1943	375
1944	250
1945	375
1946	400
1947	425
1948	600
1949	225
1950	500
1951	500
1952	575
1953	425

Year	Peak Annual Discharge (m^3/s)
1954	400
1955	375
1956	325
1957	275
1958	325
1959	300
1960	400
1961	250
1962	450
1963	375
1964	475
1965	350
1966	425
1967	275
1968	475
1969	325
1970	400
1971	425
1972	350
1973	350
1974	250
1975	250
1976	225
1977	300
1978	150
1979	375
1980	425
1981	300
1982	225
1983	225
1984	200
1985	425
1986	200
1987	250
1988	250
1989	550
1990	400

Year	Peak Annual Discharge (m^3/s)
1991	300
1994	300
1995	500
1996	300
1997	300
1998	350
1999	400
2000	225
2001	175
2002	350
2003	250
2004	250
2005	791 (major)
2006	250
2007	375
2008	325
2009	175
2010	200
2011	350
2012	475
2013	1,840 (major)
2014	350

[Return to Figure 17.2.5.6]

Media Attributions

- Figure 17.2.5.1: © Steven Earle. CC BY. Based on from data at Water Survey of Canada, Environment Canada.
- Figure 17.2.5.2: © Steven Earle. CC BY. Based on from data at Water Survey of Canada, Environment Canada.
- Figure 17.2.5.3: Left image: “Rednorthfloodwaymap” © Kmusser. CC BY-SA. Right image: From Natural Resources Canada 2012, courtesy of the Geological Survey of Canada (Photo 2000-118 by G.R. Brooks).
- Figure 17.2.5.4: © Steven Earle. CC BY.
- Figure 17.2.5.5: “Riverfront Ave Calgary Flood” © Ryan L. C. Quan. CC BY-SA. “Okotoks – June 20, 2013 – Flood waters in local campground playground-03” © Stephanie N. Jones. CC BY-SA.
- Figure 17.2.5.6: © Steven Earle. CC BY. Based on data from Water Surveys of Canada, Environment Canada.

---

1. Mannerström, M, 2008, Comprehensive Review of Fraser River at Hope Flood Hydrology and Flows Scoping Study, Report prepared for the B.C. Ministry of the Environment [PDF]. Available at: http://www.env.gov.bc.ca/wsd/public_safety/flood/pdfs_word/review_fraser_flood_flows_hope.pdf↵

This page titled 17.2.5: Flooding is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Steven Earle (BCCampus) via source content that was edited to the style and standards of the LibreTexts platform.

- 13.5: Flooding by Steven Earle is licensed CC BY 4.0. Original source: https://opentextbc.ca/physicalgeology2ed.

17.3: Groundwater

This section explores water stored and moving beneath Earth's surface. It introduces groundwater, aquifers, the water table, and the properties that control how easily water can move through rock and sediment. It also examines groundwater extraction and water quality, connecting underground water resources to their importance for human use.

This page titled 17.3: Groundwater is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Robert Wagner.

17.3.1: Groundwater and Aquifers

Groundwater is stored in the open spaces within rocks and within unconsolidated sediments. Rocks and sediments near the surface are under less pressure than those at significant depth and therefore tend to have more open space. For this reason, and because it's expensive to drill deep wells, most of the groundwater that is accessed by individual users is within the first 100 meters of the surface. Some municipal, agricultural, and industrial groundwater users get their water from greater depth, but deeper groundwater tends to be of lower quality than shallow groundwater, so there is a limit as to how deep we can go.

Porosity is the percentage of open space within an unconsolidated sediment or a rock. Primary porosity is represented by the spaces between grains in a sediment or sedimentary rock. Secondary porosity is porosity that has developed after the rock has formed. It can include fracture porosity—space within fractures in any kind of rock. Some volcanic rock has a special type of porosity related to vesicles, and some limestone has extra porosity related to cavities within fossils.

Porosity is expressed as a percentage calculated from the volume of open space in a rock compared with the total volume of rock. The typical ranges in porosity of a number of different geological materials are shown in Figure 17.3.1.1. Unconsolidated sediments tend to have higher porosity than consolidated ones because they have no cement, and most have not been strongly compressed. Finer-grained materials (e.g., silt and clay) tend to have greater porosity—some as high as 70%—than coarser materials (e.g., gravel). Primary porosity tends to be higher in well-sorted sediments compared to poorly sorted sediments, where there is a range of smaller particles to fill the spaces made by the larger particles. Glacial till, which has a wide range of grain sizes and is typically formed under compression beneath glacial ice, has relatively low porosity.

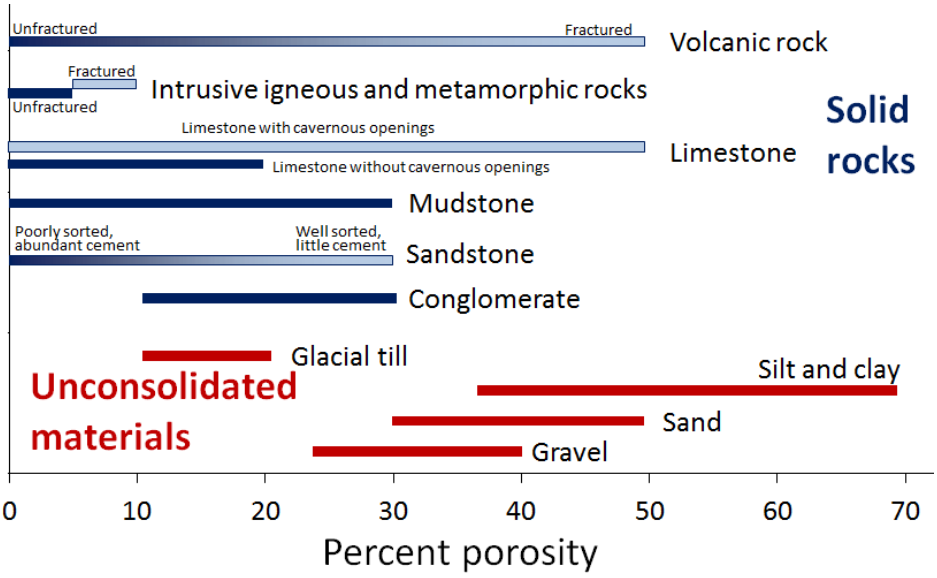


Figure 17.3.1.1 Variations in porosity of unconsolidated materials (in red) and rocks (in blue).

Consolidation and cementation during the process of lithification of unconsolidated sediments into sedimentary rocks reduces primary porosity. Sedimentary rocks generally have porosities in the range of 10% to 30%, some of which may be secondary (fracture) porosity. The grain size, sorting, compaction, and degree of cementation of the rocks all influence primary porosity. For example, poorly sorted and well-cemented sandstone and well-compressed mudstone can have very low porosity. Igneous or metamorphic rocks have the lowest primary porosity because they commonly form at depth and have interlocking crystals. Most of their porosity comes in the form of secondary porosity in fractures. Of the consolidated rocks, well-fractured volcanic rocks and limestone that has cavernous openings produced by dissolution have the highest potential porosity, while intrusive igneous and metamorphic rocks, which formed under great pressure, have the lowest.

Porosity is a measure of how much water can be stored in geological materials. Almost all rocks contain some porosity and therefore contain groundwater. Groundwater is found under your feet and everywhere on the planet. Considering that sedimentary rocks and unconsolidated sediments cover about 75% of the continental crust with an average thickness of a few hundred meters, and that they are likely to have around 20% porosity on average, it is easy to see that a huge volume of water can be stored in the ground.

Porosity is a description of how much space there could be to hold water under the ground, while **permeability** describes how those pores are shaped and interconnected. This determines how easy it is for water to flow from one pore to the next. Larger pores mean there is less friction between flowing water and the sides of the pores. Smaller pores mean more friction along pore walls, and also more twists and turns for the water to have to flow-through. A permeable material has a greater number of larger, well-connected pores spaces, whereas an impermeable material has fewer, smaller pores that are poorly connected. Permeability is the most important variable in groundwater. Permeability describes how easily water can flow through the rock or unconsolidated sediment and how easy it will be to extract the water for our purposes. The characteristic of permeability of a geological material is quantified by geoscientists and engineers using a number of different units, but the most common is the **hydraulic conductivity**. The symbol used for hydraulic conductivity is *K*. Although hydraulic conductivity can be expressed in a range of different units, in this book we will always use meters per second.

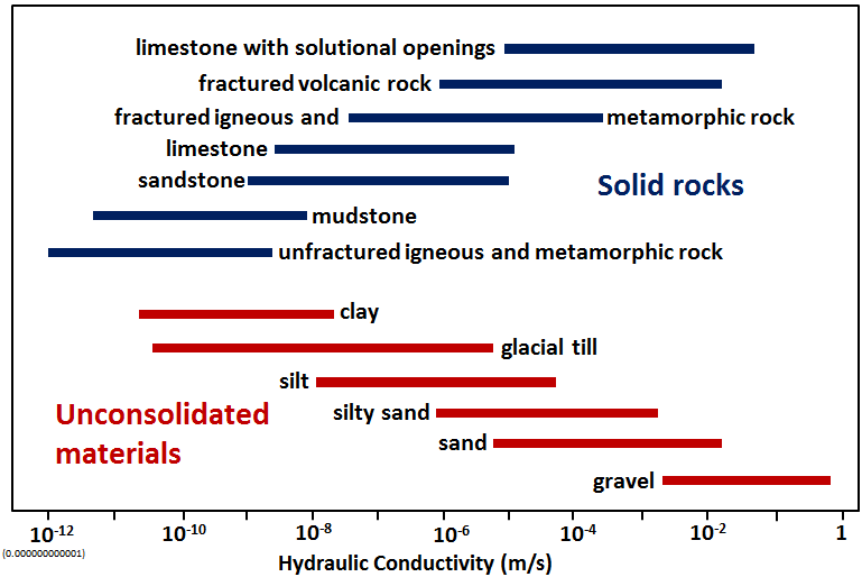
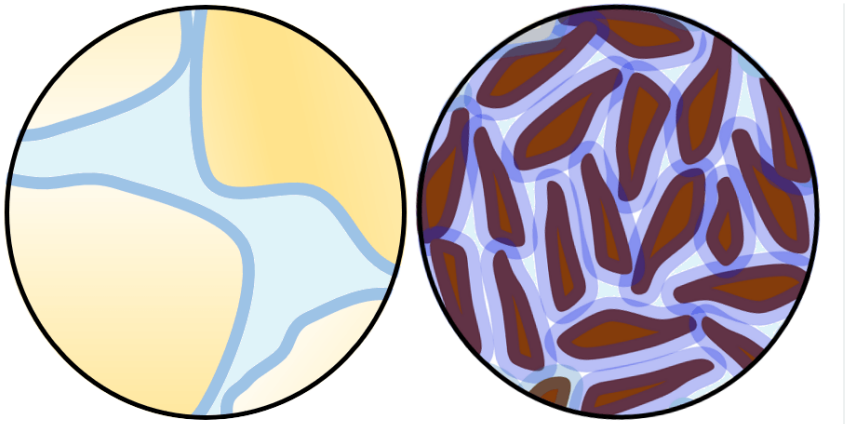


Figure 17.3.1.2 Variations in hydraulic conductivity (in meters per second) of unconsolidated materials (in red) and of rocks (in blue).

As shown on Figure 17.3.1.2 there is a wide range of permeability in geological materials from 10^{-12} meters per second (0.000000000001 meters per second) to around 1 metre per second. Unconsolidated materials are generally more permeable than the corresponding rocks (compare sand with sandstone, for example), and the coarser materials are much more permeable than the finer ones. The least permeable rocks are unfractured intrusive igneous and metamorphic rocks, followed by unfractured mudstone, sandstone, and limestone. The permeability of sandstone can vary widely depending on the degree of sorting and the amount of cement that is present. Fractured igneous and metamorphic rocks, and especially fractured volcanic rocks, can be highly permeable, as can limestone that has been dissolved along fractures and bedding planes to create solutional openings.



Sand deposit. The light blue represents water that is free to move, while the medium blue is water held to the grains by surface tension.
Most of the water is free to move.

Clay deposit. The light blue represents water that is free to move, while the blue is water held to the grains by surface tension.
Most of the water is unable to move.

Figure 17.3.1.3 Porosity versus permeability in sand and clay

Both sand and clay deposits (and sandstone and mudstone) are quite porous (30% to 50% for sand and 40% to 70% for silt and clay), but while sand can be quite permeable, clay and mudstone are not.

The surface of most silicate mineral grains has a slight negative charge due to imperfections in the mineral structure. Water (H_2O) is a polar molecule. This means that while it has no overall electrical charge, one side of the molecule has a slight positive charge (the side with the two hydrogens), compared to a slight negative charge on the other side. Water is strongly attracted to all mineral grains and water within that bound water layer (a few microns around each grain) is not able to move and flow along with the rest of the groundwater.

In the two diagrams of Figure 17.3.1.3 the bound water is represented by medium blue lines around each grain, and the water that is not bound, and can flow, is shown in light blue. In the sand, there is still a lot of water that is able to move through the sediment, but in the clay deposit almost all of the water is held tightly to the grains and this reduces the permeability.

We have now seen that there is a wide range of porosity in geological materials and an even wider range of permeability. Groundwater exists everywhere there is porosity. However, whether that groundwater is able to flow in significant quantities depends on the permeability. An **aquifer** is defined as a body of rock or unconsolidated sediment that has sufficient permeability to allow water to flow through it. Unconsolidated materials like gravel, sand, and even silt make relatively good aquifers, as do rocks like sandstone. Other rocks can be good aquifers if they are well fractured. An **aquitard** is a body that does not allow transmission of a significant amount of water, such as a clay, a till, or a poorly fractured igneous or metamorphic rock. These are relative terms, not absolute, and are usually defined based on someone's desire to pump groundwater; what is an aquifer to someone who does not need a lot of water, may be an aquitard to someone else who does. An aquifer that is exposed at the ground surface is called an **unconfined aquifer**. An aquifer where there is a lower permeability material between the aquifer and the ground surface is known as a **confined aquifer**, and the aquitard separating ground surface and the aquifer is known as the **confining layer**.

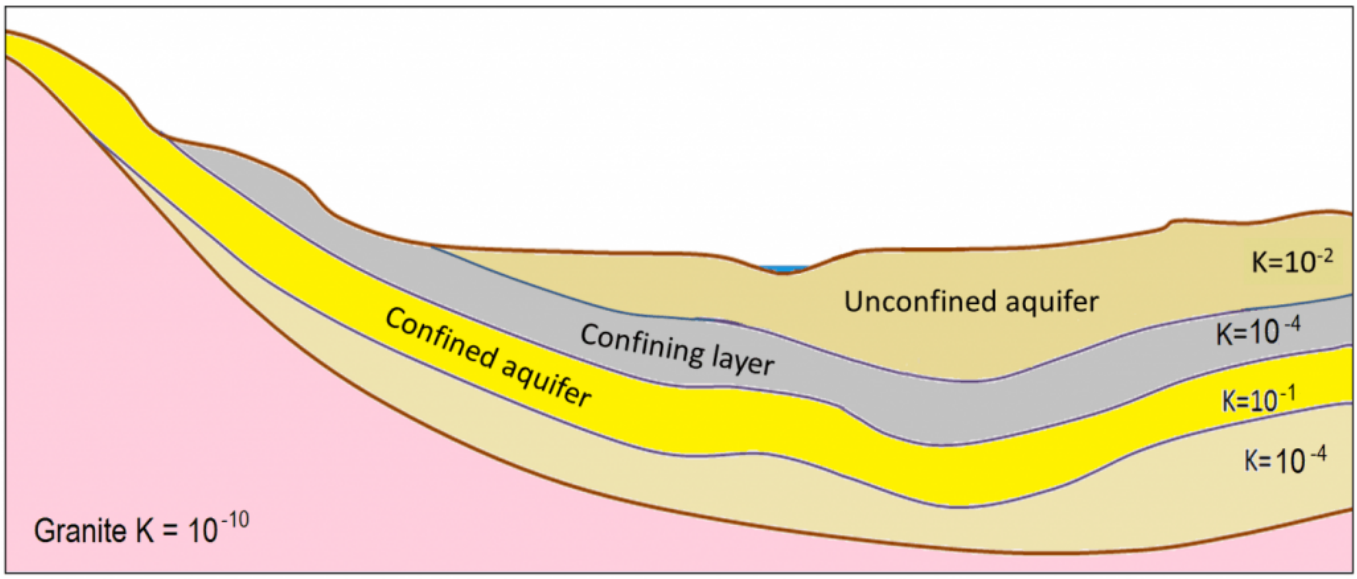


Figure 17.3.1.4 A cross-section showing materials that might serve as aquifers and confining layers. The relative permeabilities are denoted by hydraulic conductivity (K in meters per second). The pink rock is granite; the other layers are various sedimentary layers.

Figure 17.3.1.4 shows a cross-section of a series of rocks and unconsolidated materials, some of which might serve as aquifers and others as aquitards or confining layers. The granite is much less permeable than the other materials, and so is an aquitard in this context. The yellow layer is very permeable and would make an ideal aquifer. The overlying grey layer is a confining layer.

The upper buff-colored layer (K = 10^{-2} meters per second) does not have a confining layer and is an unconfined aquifer. The yellow layer (K = 10^{-1} meters per second) is "confined" by the confining layer (K = 10^{-4} meters per second), and is a confined aquifer. The confined aquifer gets most of its water from the upper part of the hill where it is exposed at the surface, and relatively little by seepage through the fine silt layer.

Image Descriptions

Figure 17.3.1.1 image description part 1: Percent porosity for solid rocks

Solid rocks	Percent porosity (Unfractured)	Percent porosity (Fractured)
Volcanic Rock	0% to roughly 20%	Roughly 20% to 50%
Intrusive igneous and metamorphic rocks	0% to 5%	5% to 10%
Limestone with cavernous openings	N/A	0% to 50%
Limestone without cavernous openings	0% to 20%	N/A
Mudstone	0% to 30%	N/A
Sandstone	0% to roughly 10%	Roughly 10% to 30%
Conglomerate	10% to 30%	N/A

Figure 17.3.1.1 image description part 2: Percent porosity of unconsolidated materials

Unconsolidated Materials	Percent porosity
Glacial Till	11% to 20%
Silt and clay	36% to 70%
Sand	30% to 50%
Gravel	24% to 40%

[Return to Figure 17.3.1.1]

Figure 17.3.1.2 image description part 1: Variations in hydraulic conductivity in meters per second in solid rocks.

Material type	Hydraulic conductivity (meters per second)
Limestone with solutional openings	10 to the power of negative 5 to 10 to the power of negative 1.2 (10^{-5} to $10^{-1.2}$)
Fractured volcanic rock	10 to the power of negative 6 to 10 to the power of negative 1.7 (10^{-6} to $10^{-1.7}$)
Fractured igneous and metamorphic rock	10 to the power of negative 7.5 to 10 to the power of negative 3.5 ($10^{-7.5}$ to $10^{-3.5}$)
Limestone	10 to the power of negative 8.4 to 10 to the power of negative 4.9 ($10^{-8.4}$ to $10^{-4.9}$)
Sandstone	10 to the power of negative 8.9 to 10 to the power of negative 5 ($10^{-8.9}$ to 10^{-5})
Mudstone	10 to the power of negative 11.3 to 10 to the power of negative 8 ($10^{-11.3}$ to 10^{-8})
Unfractured igneous and metamorphic rock	10 to the power of negative 12 to 10 to the power of negative 8.6 (10^{-12} to $10^{-8.6}$)

Figure 17.3.1.2 image description part 2: Variations in hydraulic conductivity in meters per second in unconsolidated materials.

Material Type	Hydraulic conductivity (meters per second)
Clay	10 to the power of negative 10.6 to 10 to the power of negative 7.7 ($10^{-10.6}$ to $10^{-7.7}$)
Glacial till	10 to the power of negative 10.4 to 10 to the power of negative 5.2 ($10^{-10.4}$ to $10^{-5.2}$)
Silt	10 to the power of negative 8 to 10 to the power of negative 4.3 (10^{-8} to $10^{-4.3}$)
Silty sand	10 to the power of negative 6 to 10 to the power of negative 2.7 (10^{-6} to $10^{-2.7}$)
Sand	10 to the power of negative 5.2 to 10 to the power of negative 1.7 ($10^{-5.2}$ to $10^{-1.7}$)
Gravel	10 to the power of negative 2.7 to 10 to the power of negative 0.1 ($10^{-2.7}$ to $10^{-0.1}$)

[Return to Figure 17.3.1.2]

Figure 17.3.1.4 image description: The relative permeability of various sedimentary layers. Permeability is denoted by hydraulic conductivity (K = meters per second). The buff-colored, top layer is an unconfined aquifer with K = 10^{-2}. The grey-colored second layer is a confining layer with K = 10^{-4}. The yellow-colored third layer is a confined aquifer with K = 10^{-1}. The fourth layer is unlabelled, with K = 10^{-4}. The bottom layer is granite, K = 10^{-10}. [Return to Figure 17.3.1.4]

Media Attributions

- Figure 17.3.1.1, 14.1.2, 14.1.3, 14.1.4: © Steven Earle. CC BY.

This page titled 17.3.1: Groundwater and Aquifers is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Steven Earle (BCCampus) via source content that was edited to the style and standards of the LibreTexts platform.

- 14.1: Groundwater and Aquifers by Steven Earle is licensed CC BY 4.0. Original source: https://opentextbc.ca/physicalgeology2ed.

17.3.2: Groundwater Flow

If you go out into your garden or into a forest or a park and start digging, you will find that the soil is moist (unless you're in a desert), but it's not saturated with water. This means that some of the pore space in the soil is occupied by water, and some of the pore space is occupied by air (unless you're in a swamp). This is known as the **unsaturated zone**. If you could dig down far enough, you would get to the point where all of the pore spaces are 100% filled with water (saturated) and the bottom of your hole would fill up with water. The level of water in the hole represents the **water table**, which is the surface of the **saturated zone**. In most parts of British Columbia, the water table is several meters below the surface.

Water falling on the ground surface as precipitation (rain, snow, hail, fog, etc.) may flow off a hill slope directly to a stream in the form of **runoff**, or it may **infiltrate** the ground, where it is stored in the unsaturated zone. The water in the unsaturated zone may be used by plants (transpiration), evaporate from the soil (evaporation), or continue past the root zone and flow downward to the water table, where it **recharges** the groundwater.

A cross-section of a typical hillside with an unconfined aquifer is illustrated in Figure 17.3.2.1. In areas with topographic relief, the water table generally follows the land surface, but tends to come closer to surface in valleys, and intersects the surface where there are streams or lakes. The water table can be determined from the depth of water in a well that isn't being pumped, although, as described below, that only applies if the well is within an unconfined aquifer. In this case, most of the hillside forms the **recharge area**, where water from precipitation flows downward through the unsaturated zone to reach the water table. The area at the stream or lake to which the groundwater is flowing is a **discharge area**.

What makes water flow from the recharge areas to the discharge areas? Recall that water is flowing in pores where there is friction, which means it takes work to move the water. There is also some friction between water molecules themselves, which is determined by the viscosity. Water has a low viscosity, but friction is still a factor. All flowing fluids are always losing energy to friction with their surroundings. Water will flow from areas with high energy to those with low energy. Recharge areas are at higher elevations, where the water has high gravitational energy. It was energy from the sun that evaporated the water into the atmosphere and lifted it up to the recharge area. The water loses this gravitational energy as it flows from the recharge area to the discharge area.

In Figure 17.3.2.1, the water table is sloping; that slope represents the change in gravitational potential energy of the water at the water table. The water table is higher under the recharge area (90 meters) and lower at the discharge area (82 meters). Imagine how much work it would be to lift water 8 meters high in the air. That is the energy that was lost to friction as the groundwater flowed from the top of the hill to the stream.

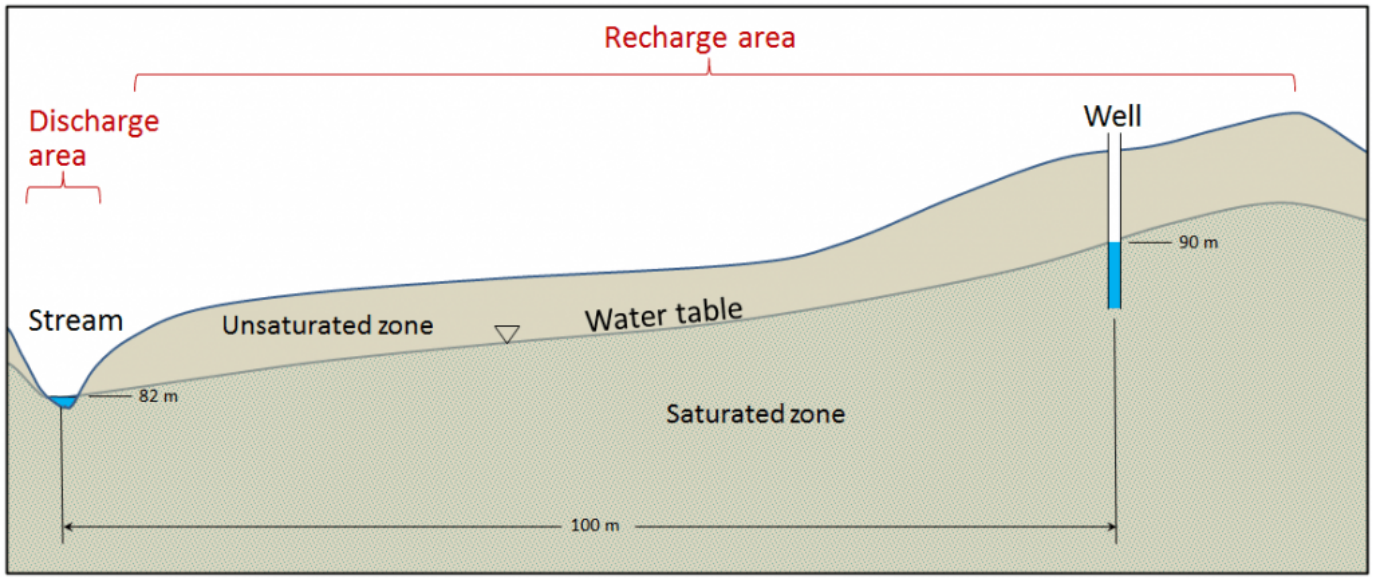


Figure 17.3.2.1 A depiction of the water table in cross-section, with the saturated zone below and the unsaturated zone above. The water table is denoted with a small upside-down triangle.

The situation gets a lot more complicated in the case of confined aquifers, but they are important sources of water so we need to understand how they work. As shown in Figure 17.3.2.2, there is always a water table, and that applies even if the geological materials at the surface have very low permeability. Where there is a confined aquifer — meaning one that is separated from the surface by a confining layer — this aquifer will have its own "water table," which is actually called a **potentiometric surface**, as it is a measure of the total potential energy of the water. The red dashed line in Figure 17.3.2.2 is the potentiometric surface for the confined aquifer, and it describes the total energy that water is under within the confined aquifer. If we drill a well into the unconfined aquifer, the water will rise to the level of the water table (well A in Figure 17.3.2.2). But if we drill a well through both the unconfined aquifer and the confining layer and into the confined aquifer, the water will rise above the top of the confined aquifer to the level of its potentiometric surface (well B in Figure 17.3.2.2). This is known as an **artesian well**, because the water rises above the top of the aquifer. In some situations, the potentiometric surface may be above the ground level. The water in a well drilled into the confined aquifer in this situation would rise above ground level, and flow out, if it's not capped (well C in Figure 17.3.2.2). This is known as a **flowing artesian well**.

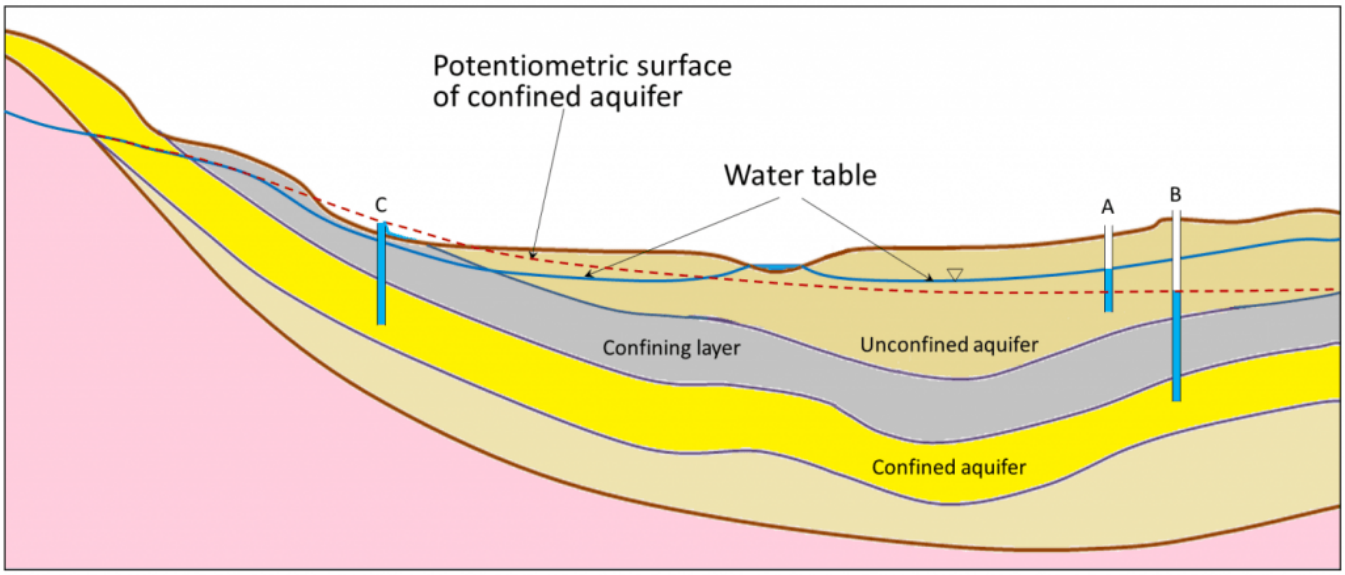


Figure 17.3.2.2 A depiction of the water table and the potentiometric surface of a confined aquifer.

In situations where there is an aquitard of limited extent, it is possible for a perched aquifer to exist as shown in Figure 17.3.2.3. Although perched aquifers may be good water sources at some times of the year, they tend to be relatively thin and small, and so can easily be depleted with over-pumping.

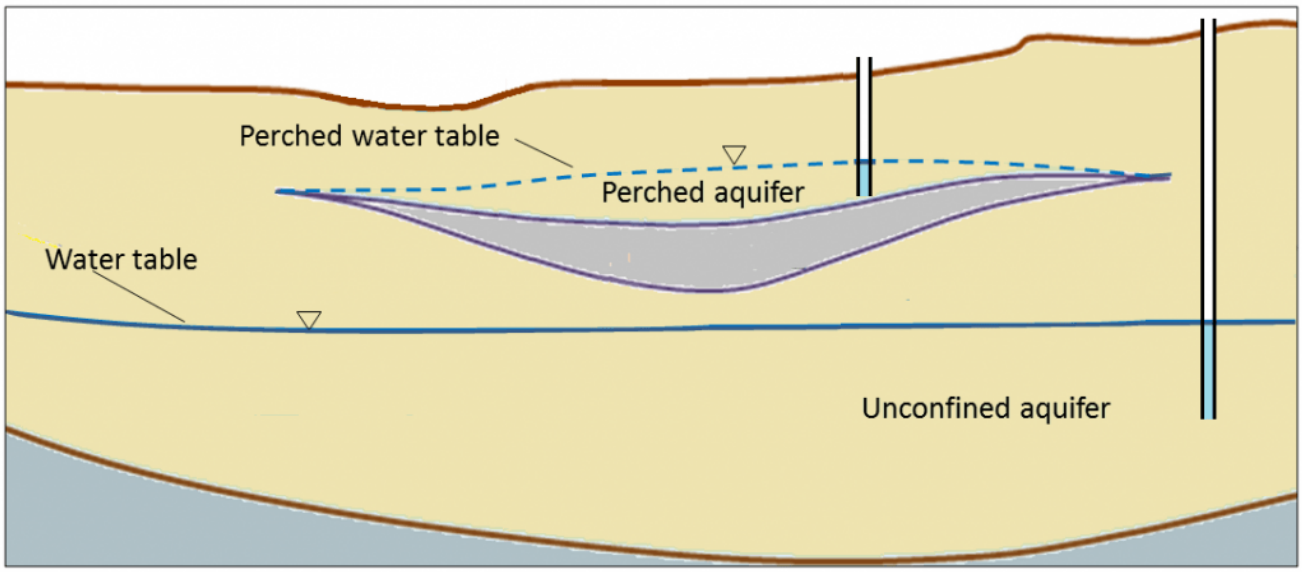


Figure 17.3.2.3 A perched aquifer above a regular unconfined aquifer.

In 1856, French engineer Henri Darcy carried out some experiments from which he derived a method for estimating the rate of groundwater flow based on the hydraulic gradient and the permeability of an aquifer, expressed using *K*, the hydraulic conductivity. Darcy's equation, which has been used widely by hydrogeologists ever since, looks like this:

$$V = K * i$$

(where *V* is the velocity of the groundwater flow, *K* is the hydraulic conductivity, and *i* is the hydraulic gradient).

We can apply this equation to the scenario in Figure 17.3.2.1. If we assume that the permeability is 0.00001 meters per second we get: *V* = 0.00001 * 0.08 = 0.0000008 meters per second. That is equivalent to 0.000048 meters per minute, 0.0029 meters/hour or 0.069 m/day. That means it would take 1,450 days (nearly four years) for water to travel the 100 meters from the vicinity of the well to the stream. Groundwater moves slowly, and that is a reasonable amount of time for water to move that distance. In fact it would likely take longer than that, because it doesn't travel in a straight line.

Exercise 14.1 How long will it take?

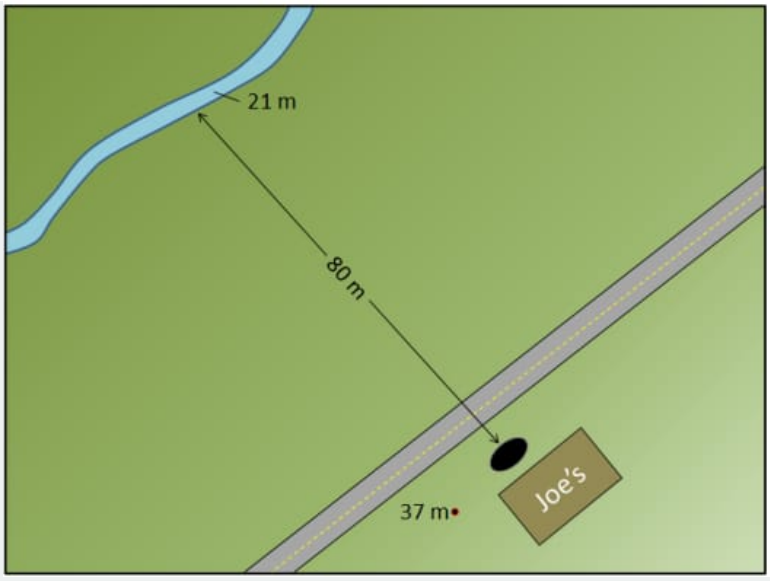


Figure 17.3.2.4

Sue, the owner of Joe's 24-Hour Gas, has discovered that her underground storage tank (UST) is leaking fuel. She calls in a hydrogeologist to find out how long it might take for the fuel contamination to reach the nearest stream. They discover that the well at Joe's has a water level that is 37 meters above sea level and the elevation of the stream is 21 meters above sea level. The sandy sediment in this area has a permeability of 0.0002 meters per second.

Using *V* = *K* * *i*, estimate the velocity of groundwater flow from Joe's to the stream, and determine how long it might take for contaminated groundwater to flow the 80 meters to the stream.

See Appendix 3 for Exercise 14.1 answers.

It's critical to understand that groundwater does not flow in underground streams, nor does it form underground lakes. With the exception of **karst** areas, with caves in limestone, groundwater flows very slowly through granular sediments, or through solid rock that has fractures in it. Flow velocities of several centimeters per day are possible in significantly permeable sediments with significant hydraulic gradients. But in many cases, permeabilities are lower than the ones we've used as examples here, and in many areas, gradients are much lower. It is not uncommon for groundwater to flow at velocities of a few millimeters to a few centimeters per year.

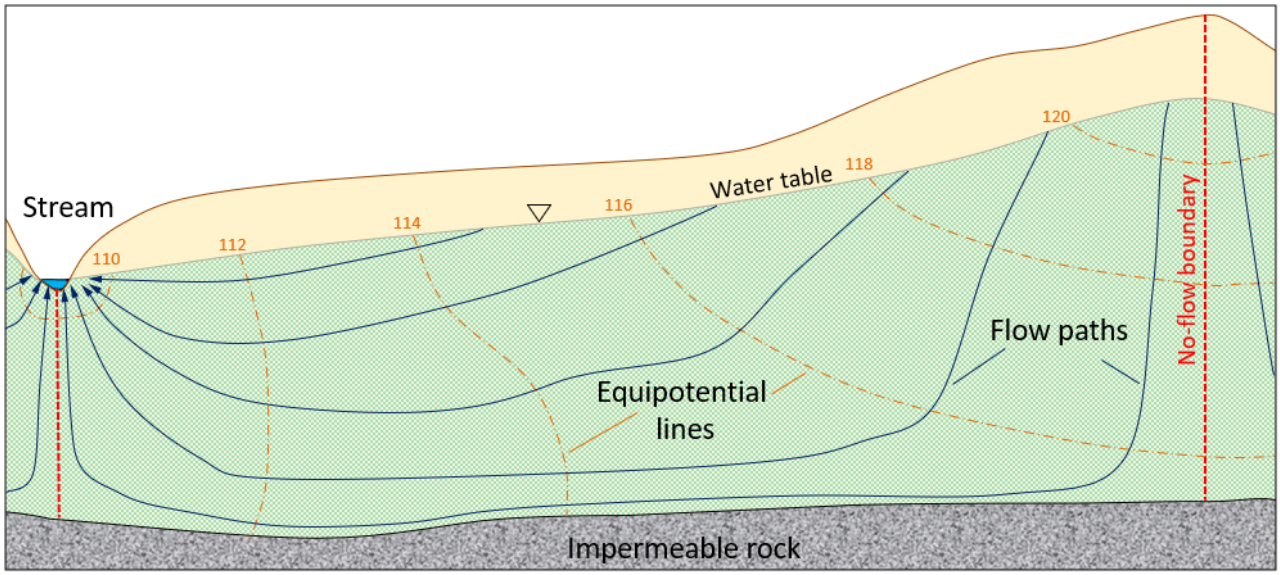


Figure 17.3.2.5 Predicted equipotential lines (orange) and groundwater flow paths (blue) in an unconfined aquifer. The orange numbers are the elevations of the water table at the locations shown, and therefore they represent the pressure along the equipotential lines.

As already noted, groundwater does not flow in straight lines. It flows from areas of higher hydraulic head to areas of lower hydraulic head, and this means that it can flow "uphill" in many situations. This is illustrated in Figure 17.3.2.5. The dashed orange lines are **equipotential lines,** meaning lines of equal pressure. The blue lines are the predicted groundwater **flow paths**. Note that in some areas the groundwater will flow steeply down, in other areas it will flow down only at a shallow angle, and in some areas—especially near to discharge zones—it will flow upward. The dashed red lines are no-flow boundaries, meaning that water cannot flow across these lines. That's not because there is something there to stop it, but because there's no pressure gradient that will cause water to flow in that direction.

Groundwater flows at right angles to the equipotential lines in the same way that water flowing down a slope would flow at right angles to the contour lines. The stream in this scenario is the location with the lowest hydraulic potential, so the groundwater that flows to the lower parts of the aquifer has to flow upward to reach this location. It is forced upward by the pressure differences, for example, the difference between the 112 and 110 equipotential lines.

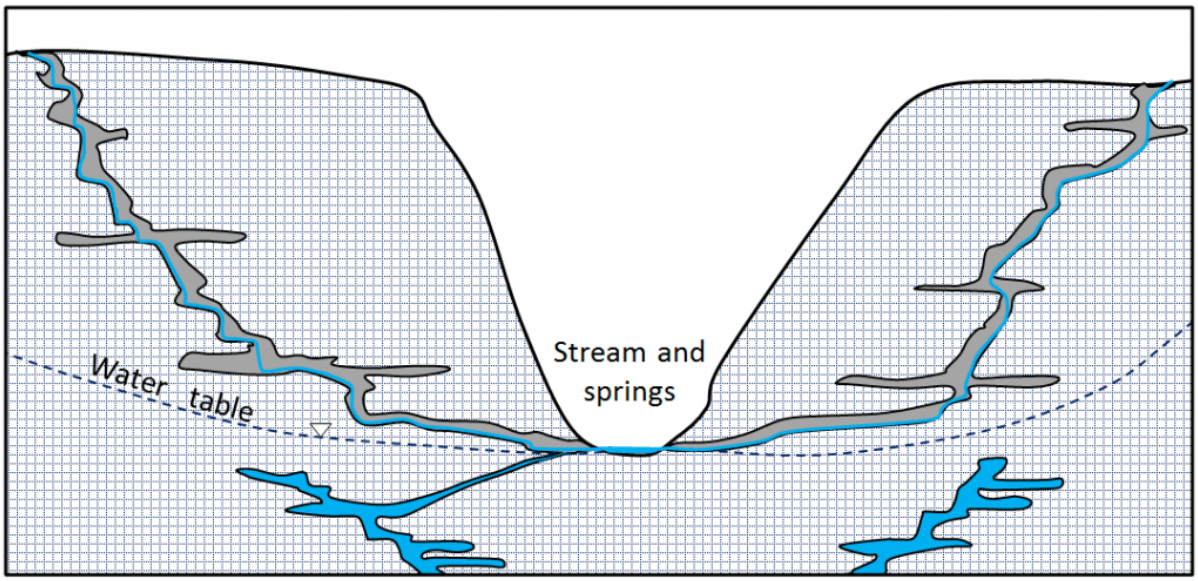


Figure 17.3.2.6 Groundwater in a limestone karst region. The water in the caves above the water table does not behave like true groundwater because its flow is not controlled by water pressure, only by gravity. The water below the water table does behave like true groundwater.

Groundwater that flows through caves, including those in **karst** areas—where caves have been formed in limestone because of dissolution—behaves differently from groundwater in other situations. Caves above the water table are air-filled conduits, and the

water that flows within these conduits is not under pressure; it responds only to gravity. In other words, it flows downhill along the gradient of the cave floor (Figure 17.3.2.6). Many limestone caves also extend below the water table and into the saturated zone. Here water behaves in a similar way to any other groundwater, and it flows according to the hydraulic gradient and Darcy's law.

Image Descriptions

Figure 17.3.2.2 image description: A cross-section showing three layers. The top layer is an unconfined aquifer. A well has been drilled into the unconfined aquifer and filled up to the height of the water table. Below the unconfined aquifer is a confining layer, and below that is a confined aquifer. A well has been drilled through the confining layer and into the confined aquifer. Water fills that well up to the potentiometric surface of the confined aquifer, which, in this case, is above the confining layer. [Return to Figure 17.3.2.2]

Media Attributions

- Figure 17.3.2.1, 14.2.3, 14.2.4, 14.2.5, 14.2.6: © Steven Earle. CC BY.

This page titled 17.3.2: Groundwater Flow is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Steven Earle (BCCampus) via source content that was edited to the style and standards of the LibreTexts platform.

- 14.2: Groundwater Flow by Steven Earle is licensed CC BY 4.0. Original source: https://opentextbc.ca/physicalgeology2ed.

17.3.3: Groundwater Extraction

Except in areas where groundwater comes naturally to the surface at a **spring** (a place where the water table intersects the ground surface), we have to construct wells in order to extract it. If the water table is relatively close to the surface, a well can be dug by hand or with an excavator, but in most cases we need to use a drill to go down deep enough. There are many types of drills that can be used; an example is shown in Figure 17.3.3.1. A well has to be drilled at least as deep as the water table, but in fact must go much deeper; first, because the water table may change from season to season and from year to year, and second, because when water is being pumped, the water level will drop, at least temporarily.



Figure 17.3.3.1 A water-well drilling rig in operation near Nanaimo, B.C. In the photo on the right the well is being test-pumped with air pressure. The casing (yellow arrow) is about 40 centimeters in diameter.

Where a well is drilled in unconsolidated sediments or relatively weak rock, it has to be lined with casing (steel pipe in most cases) in order to ensure that it doesn't cave in. A specially designed well screen is installed at the bottom of the casing. The size of the holes in the screen is carefully chosen to make sure that it allows the water to move into the well freely, but prevents aquifer particles from entering the well. A submersible pump is typically used to lift water from within the well up to the surface where it is needed. The well shown in Figure 17.3.3.1 has casing that is about 40 centimeters in diameter, which might be typical for a municipal water supply well, or a very large well for irrigation. Most domestic wells have 15 centimetre casing.

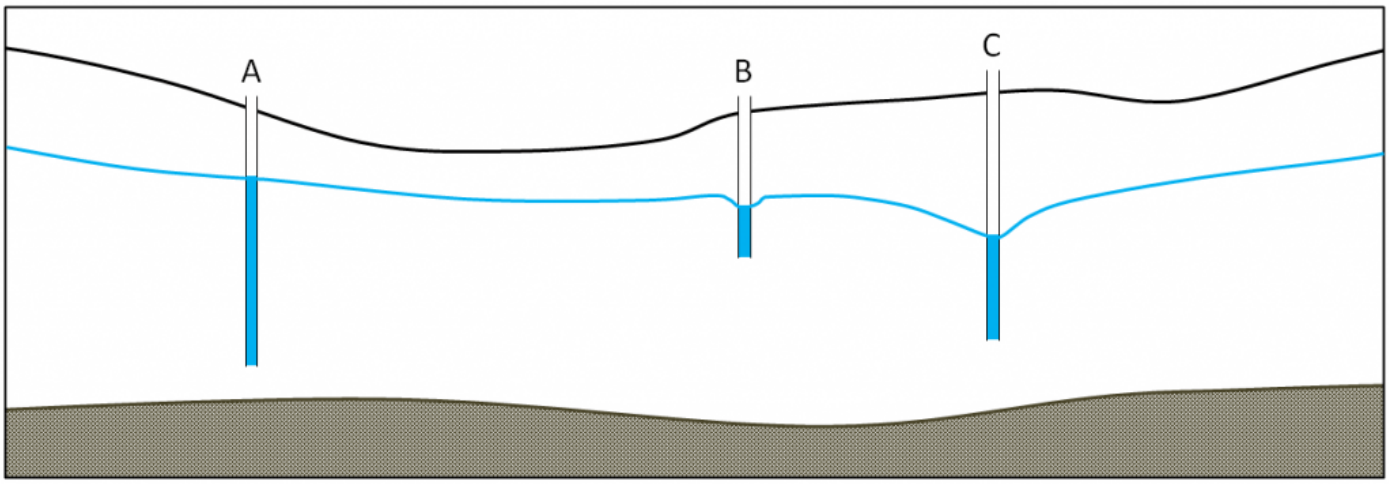


Figure 17.3.3.2 Three wells in an unconfined aquifer. Well A is not being pumped. Well B is being pumped at a slow rate and well C, which has a larger cone of depression, is being pumped at a faster rate.

Pumping water from the well removes water from inside the well at first. That lowers the water level inside the well. This means that water will flow from the surrounding aquifer toward the pumping well where the water level is now lower.

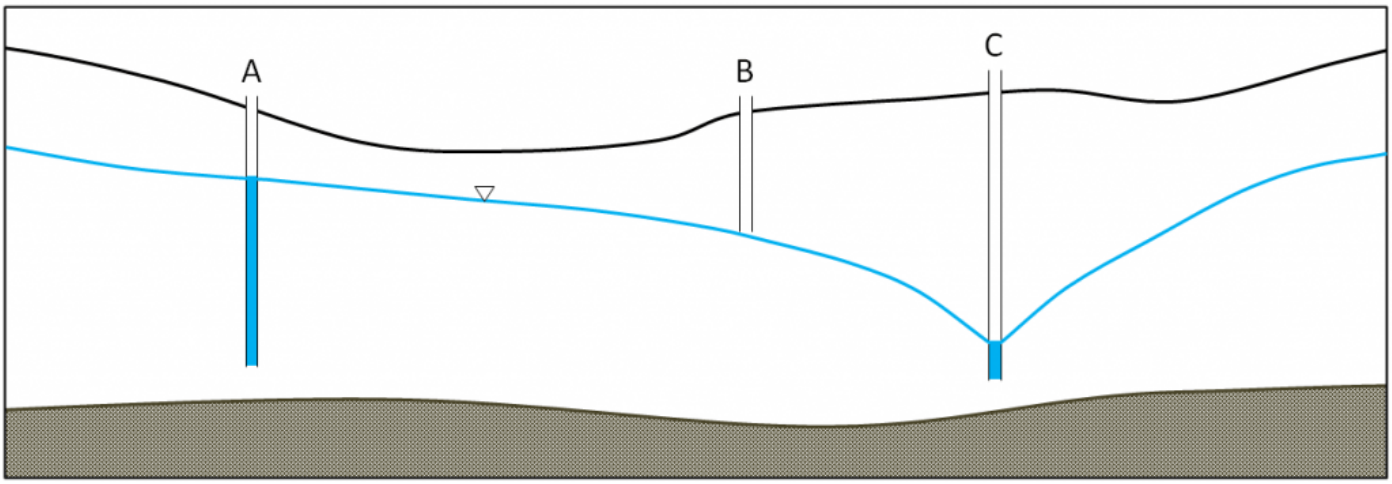


Figure 17.3.3.3 A similar scenario to that in Figure 17.3.3.2, but in this case, well C has been pumped unsustainably for a long time. The cone of depression from well C has reached well B and has contributed to it going dry because the water level is now below the bottom of well B.

That is how a well gets water from the ground. The water table, or potentiometric surface, will slope in toward the well where the water is being withdrawn. That indicates the energy gradient that is allowing water to flow toward the well. This creates a shape known as a **cone of depression** surrounding the well, as illustrated in Figure 17.3.3.2. If pumping from a well continues for hours to days, the cone of depression may result in a loss of water in nearby wells. As shown in Figure 17.3.3.3, pumping of well C has contributed to well B going dry because the water level is now below the bottom of well B. If pumping continues in well C, it too may go dry.

Exercise 14.2 Cone of depression

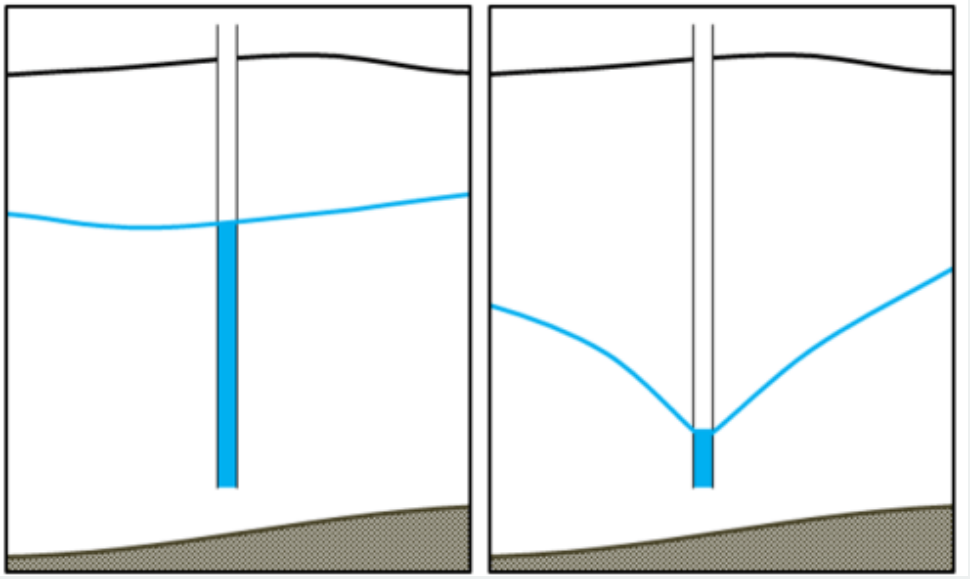


Figure 17.3.3.4

The two diagrams here show the same well before (left) and after (right) long-term pumping. A cone of depression has developed. This provides the energy gradient for water to flow toward the well so that it can be pumped out.

How will this likely affect the rate of flow into the well?

See Appendix 3 for Exercise 14.2 answers.

Like other provinces in Canada and states in the US, British Columbia has a network of observation wells administered by the Ministry of the Environment. These are wells that are installed to measure water levels; they are not pumped. There are 145 active observation wells in B.C. (in 2015), most equipped with automatic recorders that monitor water levels continuously. The main purpose of the observation wells is to monitor water table levels so that we can see if there are long-term natural fluctuations in groundwater quantity, and shorter-term fluctuations related to overuse of the resource. Most are also sampled regularly to monitor groundwater chemistry and quality.

An example of an observation well is illustrated in Figure 17.3.3.5. This one is situated at Cassidy on Vancouver Island and is used to monitor an unconsolidated aquifer that is widely used by residents with private wells.



Figure 17.3.3.5 B.C. observation well 228 near Cassidy, Vancouver Island. The installation also has a solar panel, which is not visible in this view.

The water-level data from B.C.'s observation wells are available to the public, and an example data set is illustrated in Figure 17.3.3.6. The water level in Ministry of Environment observation well 232 (OW-232), situated in Lantzville on Vancouver Island, dropped significantly from 1979 (average depth around 1.5 meters), to 2003 (average depth around 5.5 meters), but has recovered a little since then.

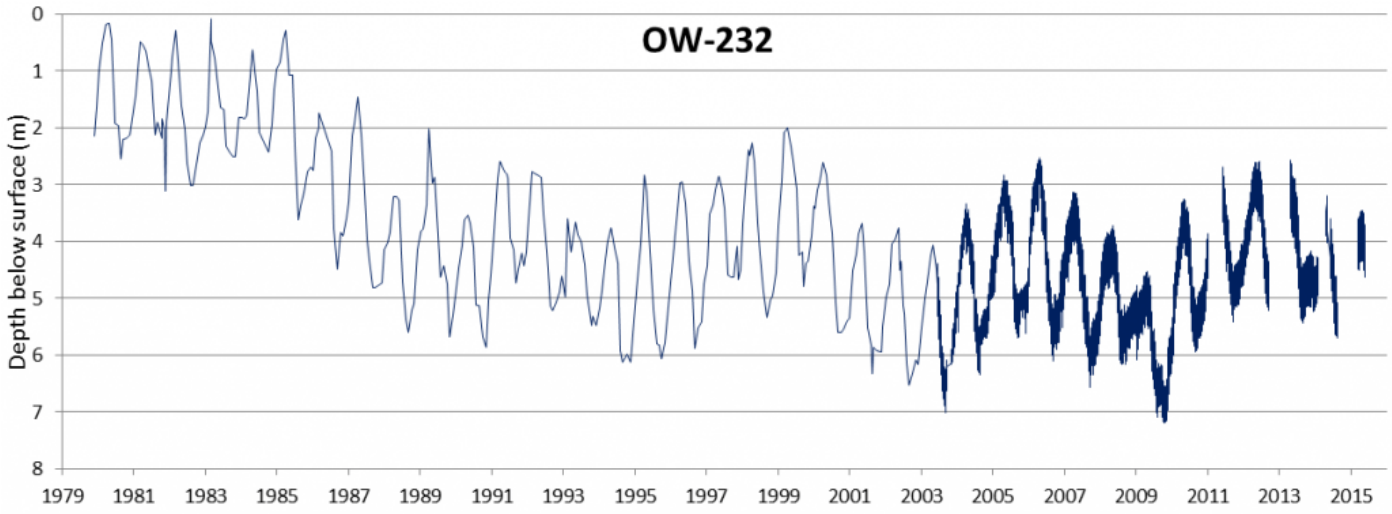


Figure 17.3.3.6 Water level data for B.C. observation well 232 on Harby Rd., Lantzville, Vancouver Island. From 1979 to 2003, depths were recorded monthly. Automated equipment was installed in 2003, and the depths were recorded hourly since that time.

The short-term variations in the level of well 232 are at a period of one year and are related to annual cycles of recharge and discharge governed by the wet winter climate and drier summers. The data for part of the period are shown in more detail in

Figure 17.3.3.7. In most of British Columbia most wells drop to their lowest levels in September or October after the relatively dry summer period. Levels increase rapidly from October through February as high winter precipitation adds recharge to the aquifer, and water is stored. The water table reaches a peak in March or April. Most wells then drop over the summer as groundwater continues to flow, but no new recharge is added. The water is drained from storage into streams or lakes and eventually into the ocean, and as a result, the water table decreases, reaching its lowest level again in September or October. Similar fluctuations are observed at most observation wells around the province, although the timing is slightly different from region to region.

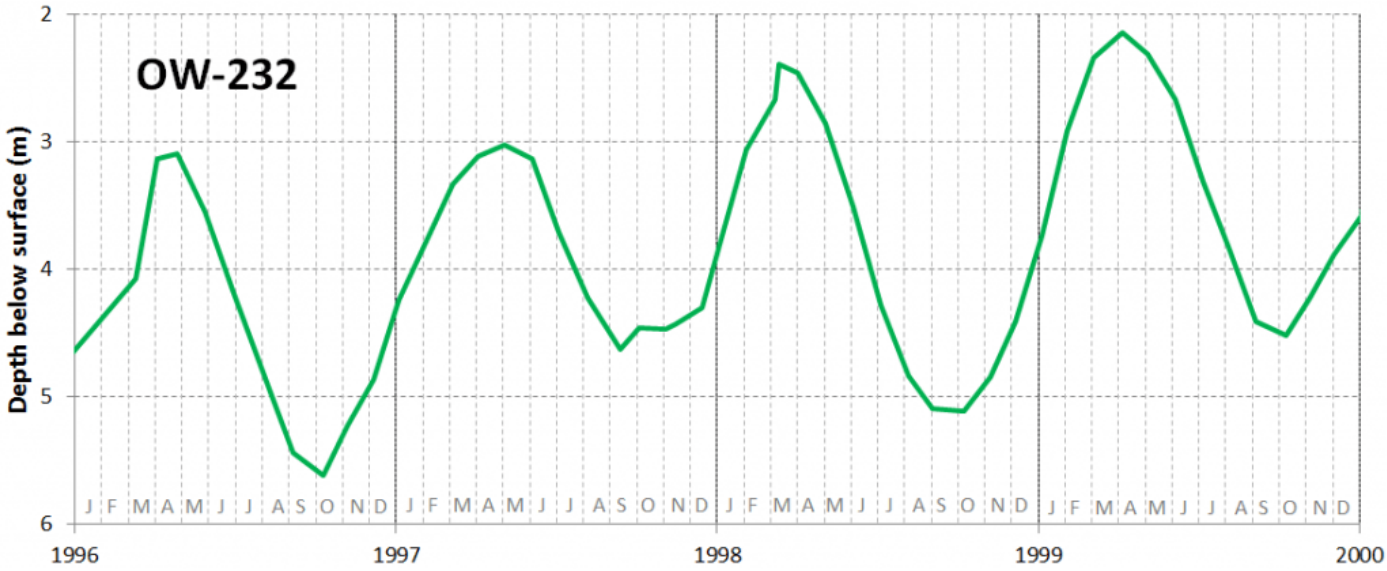


Figure 17.3.3.7 Water level data for B.C. observation well 232 for the period 1996 to 2000, showing seasonal variations.

The long-term fluctuations in levels in observation wells in different places are also quite variable. Long-term changes in climate can lead to gradual natural changes in water levels. These long-term cycles lasting years or decades are mixed with the effects of well pumping. Some observation wells show consistent decreases in water level that may indicate long-term over-extraction. Many others show generally consistent levels over several decades, and some show increases in water level. One of the important jobs performed by hydrogeologists working for government agencies is to examine these long-term records of water levels for indications of how sustainable the groundwater use might be.

Exercise 14.3 What's your water table doing?

Visit the B.C. Ministry of the Environment observation well website and use the map to find an observation well near you. When you click on a point, a window pops up with a link that says, "Click for details about this well." Click on that link and then choose one of the options available. The "Graphs" tab will show you a graph of the water levels, and you should be able to tell if the level is generally increasing or decreasing. If there isn't much data to see, choose a different well.

(Similar data are available for Alberta and Saskatchewan and other jurisdictions as well. Search using terms like "Alberta observation wells.")

See Appendix 3 for Exercise 14.3 answers.

In 2016 the new Water Sustainability Act came into effect in British Columbia, requiring licensing groundwater extraction for the first time. The new Act also includes provisions for determining "environmental flow needs"—the amount of water that must be in surface water streams at different times of the year to meet the needs of the ecosystem that depends on the streamflow. For example, many streams in B.C. support populations of salmon that live in the stream for part of their life cycle or return to their home stream for spawning. Groundwater forms a part of the baseflow in a watershed, and is therefore an important part of the environmental flow needs. Careful work is needed in the coming years to ensure that the amount of water licensed to be extracted from surface water and groundwater for human use does not interfere with the amount of water needed for the natural water-dependent ecosystems to function.

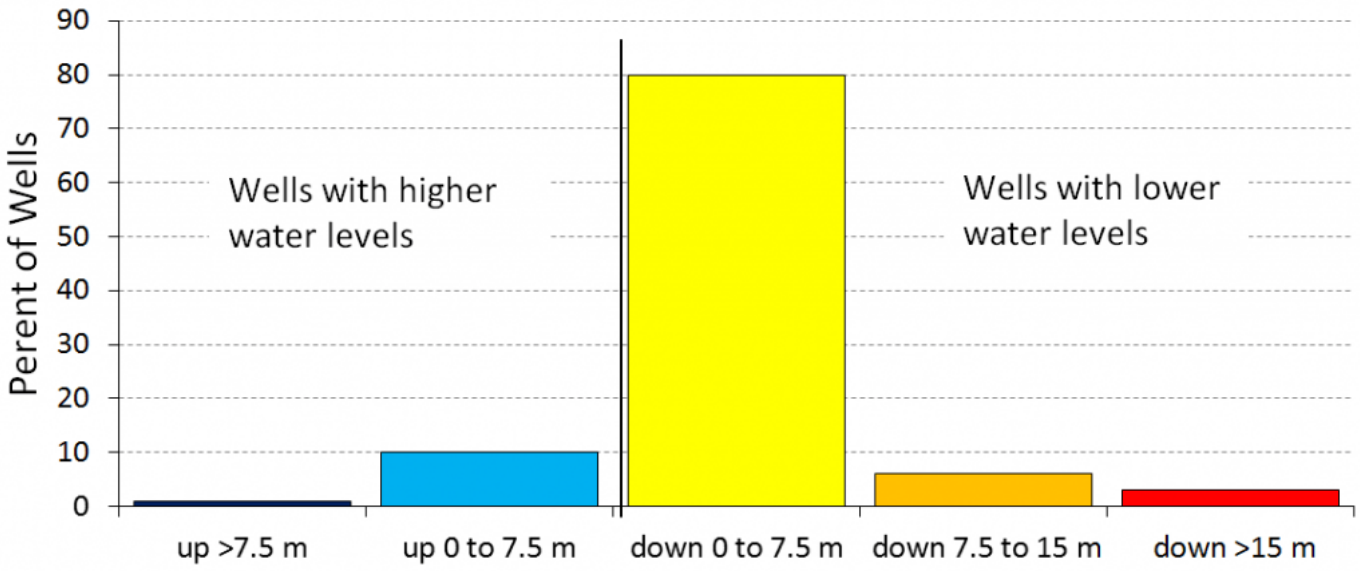


Figure 17.3.3.8 Changes in water levels in wells in California from 2011 to 2013.

The situation in California, where groundwater extraction over large areas is leading to declining water levels, is quite different from that in B.C. According to the state Department of Water Resources, 80% of groundwater wells showed drops in water level of 0 meters to 7.5 meters between 2011 and 2013, another 6% dropped by 7.5 meters to 15 meters, and 3% dropped by more than 15 meters (Figure 17.3.3.8). Over the same time period, only 10% of well levels increased by 0 meters to 7.5 meters, and 1% increased by more than 7.5 meters. The drought that gripped California in 2013 had worsened significantly by 2015, and California farmers — and the people across North America that eat the food they produce — continue to have a prodigious appetite for irrigation water. California, like B.C., is introducing new groundwater regulations to try to control water usage and halt water table declines.

Impermeable Surfaces

Even if groundwater supplies are not being depleted by overuse, or by a changing climate, we are continuing to put stress on aquifers by covering vast areas with impermeable surfaces that don't allow rain and snow-melt to infiltrate and become groundwater. Instead, water that falls on these surfaces is channeled into drainage systems, then into storm sewers, and then directly into rivers and the ocean. In cities and their suburbs, the biggest culprits are parking lots, roads, and highways. While it would great if we didn't dedicate such huge areas to cars, that's not about to change quickly, so we need to think about ways that we can improve surface water infiltration in cities. One way is to use road and parking surfaces that will allow water to seep through, although this is not practical in many cases. Another way is to ensure that runoff from pavement is channeled into existing or constructed wetlands that serve to decontaminate the water, and then allow it to infiltrate into the ground.

Image Descriptions

Figure 17.3.3.5 image description: Inside a box attached to the well casing, there is a battery, a depth sensor wheel, a spare float, and an instrument that provides a depth readout. This one says 4.047 meters. There is also a communication antenna. [Return to Figure 17.3.3.5]

Media Attributions

- Figure 17.3.3.1, 14.3.2, 14.3.3, 14.3.4, 14.3.5: © Steven Earle. CC BY.
- Figure 17.3.3.6: © Steven Earle. CC BY. Based on data from "Observation Well #232.
- Figure 17.3.3.7: © Steven Earle. CC BY. Based on data from "Observation Well #232."
- Figure 17.3.3.8:© Steven Earle. CC BY. Based on State of California Department of Water Resources data.

This page titled 17.3.3: Groundwater Extraction is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Steven Earle (BCCampus) via source content that was edited to the style and standards of the LibreTexts platform.

- 14.3: Groundwater Extraction by Steven Earle is licensed CC BY 4.0. Original source: https://opentextbc.ca/physicalgeology2ed.

17.3.4: Groundwater Quality

As was noted at the very beginning of this chapter, one of the good things about groundwater as a source of water is that it is not as easily contaminated as surface water is. But there are two caveats to that: one is that groundwater can become naturally contaminated because of its very close connection to the materials of its aquifer, and the second is that once contaminated by human activities on surface, groundwater is very difficult to clean up.

Natural Contamination of Groundwater

Groundwater moves slowly through an aquifer, and unlike the surface water of a stream, it has a lot of contact with the surrounding rock or sediment. In most aquifers, the geological materials that make up the aquifer are relatively inert, or are made up of minerals that dissolve very slowly into the groundwater. Over time, however, all groundwater gradually has more and more material dissolved within it as it remains in contact with the aquifer. In some areas, that rock or sediment includes some minerals that could potentially contaminate the water with elements that might make the water less than ideal for human consumption or agricultural use. Examples include copper, arsenic, mercury, fluorine, sodium, and boron. In some cases, contamination may occur because the aquifer material has particularly high levels of the element in question. In other cases, the aquifer material is just normal rock or sediment, but some particular feature of the water or the aquifer allows the contaminant to build up to significant levels.

An example of natural contamination takes place in the bedrock aquifers of the east coast of Vancouver Island and the adjacent Gulf Islands. The aquifer is the Cretaceous (90 Ma to 65 Ma) Nanaimo Group, which is made up of sandstone, mudstone, and conglomerate (Figure 17.3.4.1).



Figure 17.3.4.1 Cretaceous Nanaimo Group sandstone exposed in a Nanaimo parking lot.

The rocks of the Nanaimo Group are not particularly enriched in any trace elements, but the submarine-fan sandstone that makes up much of the group has relatively high levels of clay (for a sandstone). This clay is good at adsorbing[1] some elements from the water and desorbing others, and in the process, its pH goes up (it becomes alkaline). At high pH levels (some as high as 9 in groundwater in the Nanaimo Group), the element fluorine that is naturally present in the rock (as it is in almost any rock) has an increased tendency to dissolve in the water. In some areas, groundwater in the Nanaimo Group has fluorine levels that are well above recommended levels for drinking water. The World Health Organization (WHO) maximum acceptable concentration (MAC) for fluorine is 1.5 milligrams per litre (milligrams per litre). Between 5% and 10% of the domestic wells around Nanaimo and adjacent Gabriola Island have more than that, some as much as 10 milligrams per litre. A small amount of fluorine in the human diet is considered important for maintaining dental health, but high levels can lead to malformation and discoloration of teeth, and long-term exposure can lead to other more serious health effects such as skeletal problems.

Nanaimo Group groundwater can also have elevated levels of boron, again related to pH and adsorption from clay minerals. While

boron at the levels found there is not toxic to humans, there is enough boron in some wells to be toxic to plants, and the water cannot be used for irrigation.

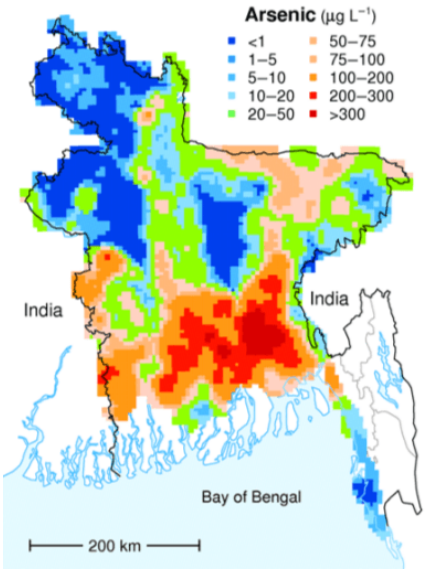


Figure 17.3.4.2 The distribution of arsenic in groundwater in Bangladesh. The WHO recommended safe level for arsenic is 10 µg/L. All of the green, orange, and red areas on the map exceed that limit.

Rural residents in the densely populated country of Bangladesh (over 1,000 residents per square kilometre, compared with 3.4 per square kilometre in Canada) used to rely mostly on surface supplies for their drinking water, and many of these were subject to bacterial contamination. Infant mortality rates were among the highest in the world and other illnesses such as diarrhea, dysentery, typhoid, cholera, and hepatitis were common. In the 1970s, international agencies, including UNICEF, started a program of drilling wells to access abundant groundwater supplies at depths of 20 meters to 100 meters. Eventually over 8 million such wells were drilled. Infant mortality and illness rates dropped dramatically, but it was later discovered that the water from a high proportion of these wells has arsenic above safe levels (Figure 17.3.4.2).

Most of the wells in the affected areas are drilled into relatively recent sediments of the vast delta of the Ganges and Brahmaputra Rivers. While these sediments are not particularly enriched in arsenic, they have enough organic matter in them to use up any oxygen present. This leads to water with a naturally low oxidation potential (anoxic conditions); arsenic is highly soluble under these conditions, and so any arsenic present in the sediments easily gets dissolved into the groundwater. Arsenic poisoning leads to headaches, confusion, and diarrhea, and eventually to vomiting, stomach pain, and convulsions. If not treated, the final outcomes are heart disease, stroke, cancer, diabetes, coma, and death. There are ways to treat arsenic-rich groundwater, but it is a challenge in Bangladesh to implement the simple and effective technology that is available.

Anthropogenic Contamination of Groundwater

Groundwater can become contaminated by pollution at the surface (or at depth), and there are many different anthropogenic (human-caused) sources of contamination.

The vulnerability of aquifers to pollution depends on several factors, including the depth to the water table, the permeability of the material between the surface and the aquifer, the permeability of the aquifer, the slope of the surface, and the amount of precipitation. Confined aquifers tend to be much less vulnerable than unconfined ones, and deeper aquifers are less vulnerable than shallow ones. Steeper slopes mean that surface water tends to run off rather than infiltrate (and this can reduce the possibility of contamination). Contamination risk is also less in dry areas than in areas with heavy rainfall.

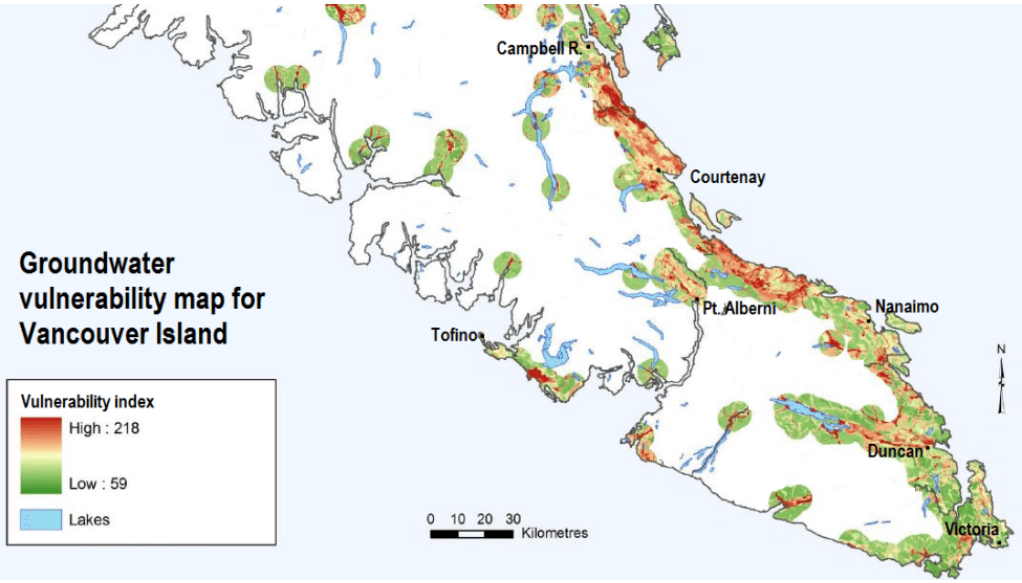


Figure 17.3.4.3 The vulnerability to anthropogenic contamination of aquifers on southern Vancouver Island. Much of the island is not mapped (shown as white) because of a lack of aquifer information in areas without wells.

Studies of groundwater vulnerability have been completed for various regions of British Columbia. A groundwater vulnerability map for southern Vancouver Island is shown in Figure 17.3.4.3. The yellow to red areas are considered to have high vulnerability to pollution from surface sources, and most of these are where the aquifers are unconfined in quite permeable unconsolidated sediments of either glacial or fluvial origin, where the water table is relatively shallow and the terrain is relatively flat.

The important sources of anthropogenic groundwater contamination include the following:

- Chemicals and animal waste related to agriculture, and chemicals applied to golf courses and domestic gardens
- Landfills
- Industrial operations
- Mines, quarries, and other rock excavations
- Leaking fuel storage tanks (especially those at gas stations)
- Septic systems
- Runoff from roads (e.g., winter salting) or chemical spills of materials being transported

Agriculture

Intensive agricultural operations and golf courses can have a significant impact on the environment, especially where chemicals and other materials are used to enhance growth or control pests. An example of agricultural contamination is in the Abbotsford area of the Fraser Valley, where nitrate levels above the 44 milligrams per litre maximum acceptable level (expressed as nitrate) in the Abbotsford-Sumas aquifer have been observed since the 1950s; however, the problem became much worse as agriculture intensity increased in the 1980s. By 2004, groundwater with nitrate levels in excess of 44 milligrams per litre was reported over an area of about 75 kilometers square around Abbotsford, and the problem extended across the border into the Sumas area of Washington State.

This region is intensively used for berry crops (especially raspberries and blueberries) and large poultry operations, as well as lesser amounts of grazing and forage crops. Chicken manure is typically stored in fields adjacent to chicken barns, and may release nitrogen to the environment from runoff water, and from releases of ammonia gas. Over decades, both chemical fertilizers and chicken manure and other manures have been applied to the berry crops to provide extra nitrogen to help maximize berry growth. If the fertilizer added is in excess of what the plants need, or is poorly timed compared to when it is needed, then the extra nitrogen may be leached into the groundwater below. Berry crops are irrigated over the summer to help the crops grow. Summer irrigation and winter rainfall may carry excess nitrate from the near surface to the aquifer below.

Since the 1990s, agricultural practices have been tightened up to reduce the rate of groundwater contamination, but it will take decades for nitrate levels to drop in the Abbotsford-Sumas aquifer. Agriculture and Agri-Food Canada and many others are conducting research on better irrigation and nitrate management techniques to reduce the amount of nitrogen that leaches to

groundwater.

Landfills

In the past, domestic and commercial refuse was commonly trucked to a “dump” (typically a hole in the ground), and when the hole was filled, it was covered with soil and forgotten. In situations like this, rain and melting snow can easily pass through the soil used to cover the refuse. This water passes into the waste itself, and the resulting landfill **leachate** that flows from the bottom of the landfill can seriously contaminate the surrounding groundwater and surface water. In the past few decades, regulations around refuse disposal have been significantly strengthened, and important steps have been taken to reduce the amount of landfill waste by diverting recyclable and compostable materials to other locations.

A modern engineered landfill has an impermeable liner (typically heavy plastic, although engineered clay liners or natural clay may be adequate in some cases), a plumbing system for draining leachate (the rainwater that flows through the refuse and becomes contaminated), and a network of monitoring wells both within and around the landfill (Figure 17.3.4.4). Once part or all of a landfill site is full, it is sealed over with a plastic cover, and a system is put in place to extract **landfill gas** (typically a mixture of carbon dioxide and methane). That gas can be sent to a nearby location where it is burned to create heat or used to generate electricity. The leachate must be treated, and that can be done in a normal sewage treatment plant.

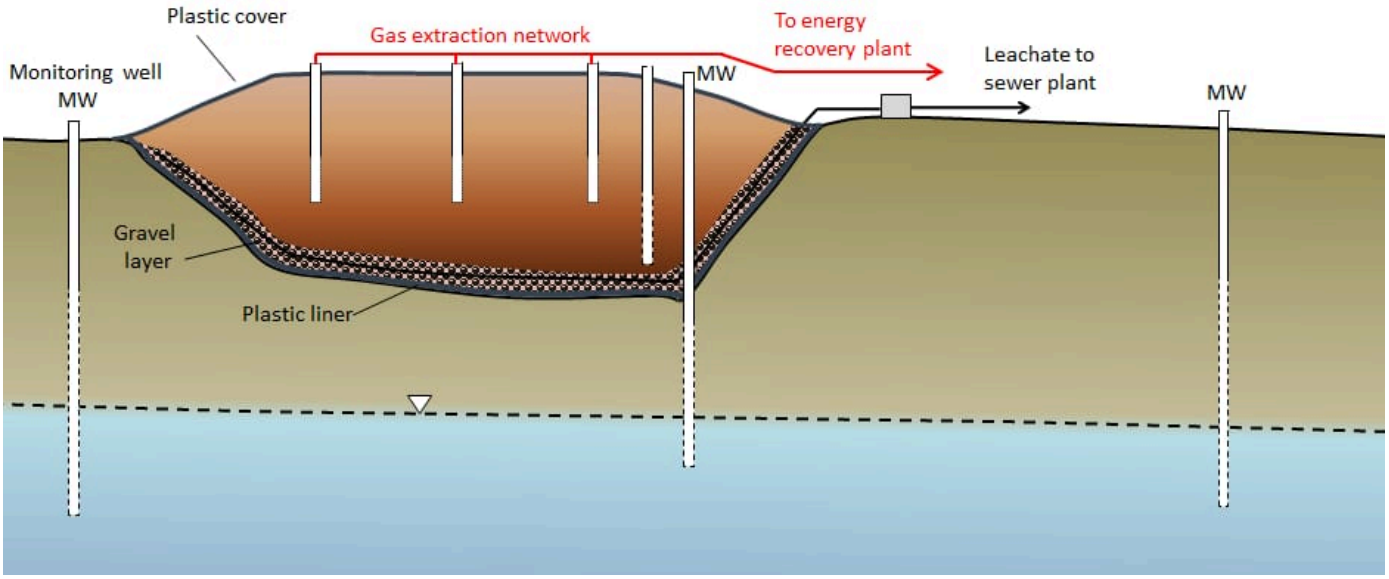


Figure 17.3.4.4 A cross-section of a typical modern landfill.

The monitoring wells are used to assess the level of the water table around the landfill and to collect groundwater samples so that any leakage can be detected. Because some leakage is almost inevitable, the ideal placement for landfills is in areas where the depth to the water table is significant (tens of meters if possible) and where the aquifer material is relatively impermeable. Landfills should also be situated far from streams, lakes, or wetlands so that contamination of aquatic habitats can be avoided.

Today there are hundreds of abandoned dumps scattered across the country; most have been left to contaminate groundwater that we might wish to use sometime in the future. In many cases, it’s unlikely that we’ll be able to do so.

Exercise 14.4 What goes on at your landfill?



Figure 17.3.4.5

Unless you live in a remote rural area, there's a good chance that the refuse you can't recycle is picked up at the curb and taken to a landfill. Most landfills are operated by cities or districts, and you should be able to find information about yours on the appropriate local government website. See if you can answer some the following questions:

1. Which government body operates your landfill?
2. Where is the landfill situated?
3. Is your waste all placed in a landfill, or are there other processes in use (e.g., incineration or composting)?
4. Are landfill gases captured, and, if so, what is done with them?
5. What could be changed to improve the waste disposal situation in your community (e.g., more recycling, compost collection, waste-to-energy technology)?

See Appendix 3 for Exercise 14.4 answers.

Industrial Operations

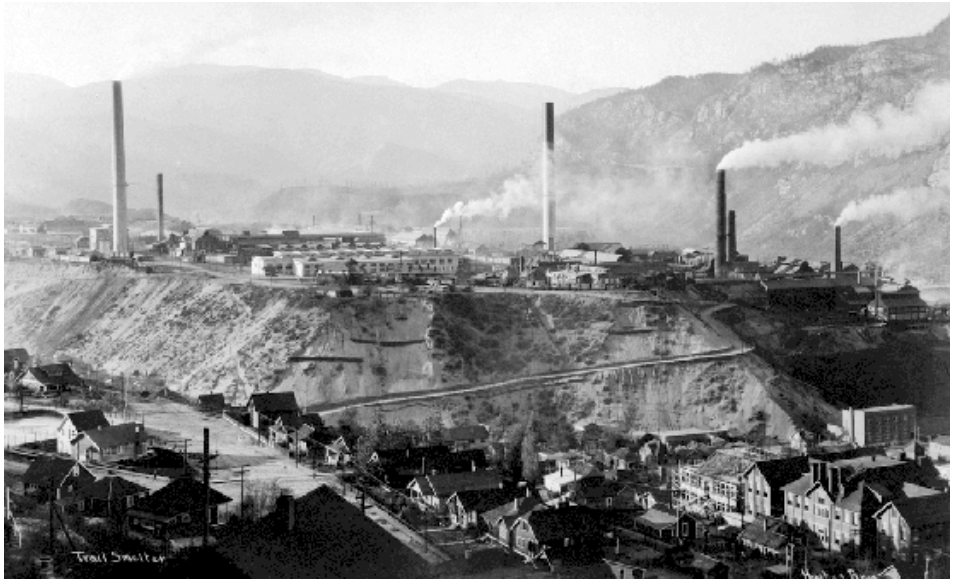


Figure 17.3.4.6 The Trail lead-zinc smelter in 1929.

Although western Canada doesn't have the same extent of industrial pollution as other parts of the country, there are still seriously contaminated sites in the west, most with the potential to contaminate groundwater. One example is the lead and zinc smelter at Trail, B.C. The largest in the world, it has been operating for over 100 years and has left a residue of metal contamination around the region (Figure 17.3.4.6). In some parts of Trail, the contamination is serious enough that existing soil has been removed

from residential properties and replaced with clean soil brought in from elsewhere. This contaminated soil has contributed to contamination of groundwater in the Trail area. Groundwater beneath the actual smelter site is contaminated, and the operator (Teck Resources) is currently working on plans to prevent that water from reaching the nearby Columbia River.

Mines, Quarries, and Rock Excavations

Mines and other operations that involve the excavation of large amounts of rock (e.g., highway construction) have the potential to create serious environmental damage. The exposure of rock that has previously not been exposed to air and water can lead to the oxidation of sulfide-bearing minerals, such a pyrite (FeS2), within the rock. The combination of pyrite, water, oxygen, and a special type of bacteria (*Acidithiobacillus ferrooxidans*) that thrives in acidic conditions leads to the generation of acidity, in some cases to pH less than 2. Water that acidic is hazardous by itself, but the low pH also has the property of increasing the solubility of certain heavy metals. The water that is generated by this process is known as acid rock drainage (ARD). ARD can occur naturally where sulfide-bearing rocks are near the surface.



Figure 17.3.4.7 Acidic runoff at the abandoned Mt. Washington Mine near Courtenay, B.CThe issue of ARD is a major environmental concern at both operating mines and abandoned mines (see Chapter 20). In streams around the Mt. Washington Mine on Vancouver Island (Figure 17.3.4.7), copper levels are high enough to be toxic to fish. Groundwater adjacent to the contaminated streams in the area is very likely contaminated as well.

Leaking Fuel Tanks



Figure 17.3.4.8 A closed and fenced gas station site in Nanaimo, B.C. The white pipes in the background are wells for monitoring groundwater contamination on the site.

Underground storage tanks (USTs) are used to store fuel at gas stations, industrial sites, airports, and anywhere that large volumes of fuel are used. They do not last forever, and eventually they start to leak their contents into the ground. This is a particular problem at older gas stations — although it may also become a future problem at newer gas stations. You may have noticed gas stations that have been closed and then surrounded by chain-link fence (Figure 17.3.4.8). In virtually all such cases the closure has been triggered by the discovery of leaking USTs and the requirement to cease operations and remediate the site.

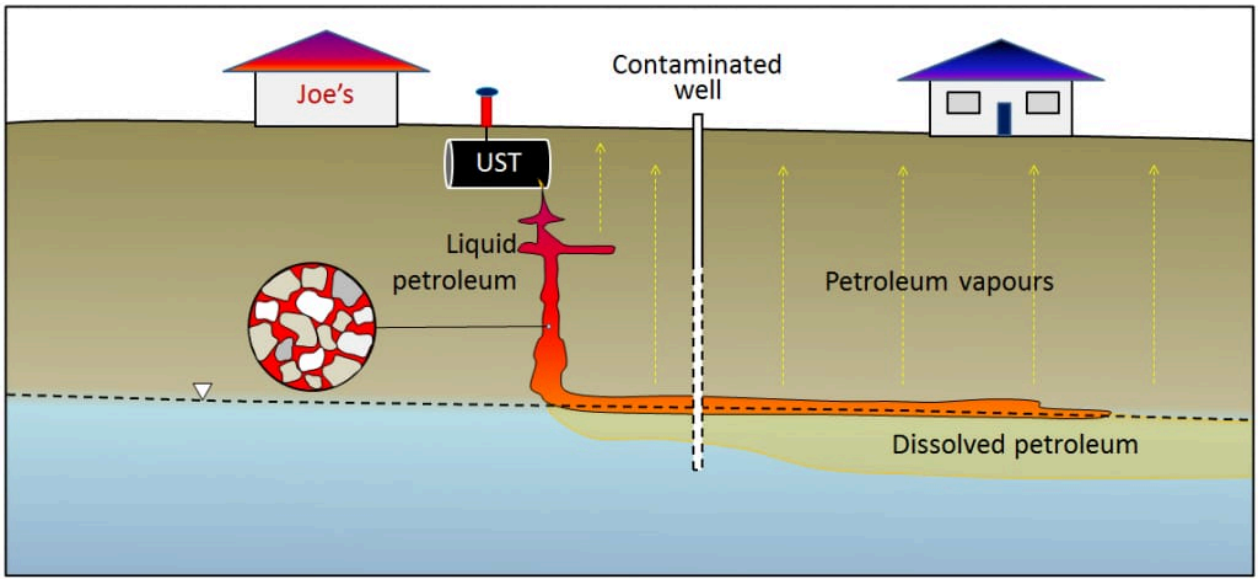


Figure 17.3.4.9 A depiction of the fate of different components of a petroleum spill from an underground storage tank.

Petroleum fuels are complex mixtures of hydrocarbon compounds and the properties of their components—such as density, viscosity, solubility in water, and volatility—tend to vary widely. As a result, a petroleum spill is like several spills for the price of one. The petroleum liquid slowly settles through the unsaturated zone and then tends to float on the surface of the groundwater (Figure 17.3.4.9). The more readily soluble components of the spill dissolve in the groundwater and are dispersed along with the normal groundwater flow, and the more volatile components of the spill rise toward the surface, potentially contaminating buildings.

Exercise 14.5 Find a leaking UST in your community

There is almost certainly a leaking UST at a former gas station near you. Look for an empty property that is surrounded by a chain-link fence with “No Trespassing” signs. You might see evidence of monitoring wells (like those shown in Figure 17.3.4.9 , and there could be some petroleum barrels around that are being used to store contaminated water. Once you’ve identified one of these, you’ll probably start seeing them everywhere!

See Appendix 3 for Exercise 14.5 answers.

Septic Systems

In areas that are not served by sewage networks leading to a central sewage treatment plant, most homeowners rely on **septic systems** for disposal of sewage. There are two primary components to a simple septic system, the septic tank and the drainage field (Figure 17.3.4.10). A typical septic tank is constructed of either concrete or plastic and has a volume of 5,000 litres to 10,000 litres (5 m^3 to 10 m^3). This forms the first treatment and is designed to be **anaerobic** (without oxygen). That promotes the activity of certain bacteria that help break down the waste. As the waste is degraded, some portions tend to sink to form sludge at the base of the tank, and others float to the surface, forming a scum layer. A septic tank may be divided into two parts to keep the sludge at the bottom and the scum on the top from draining out. The water then moves to the drainage field, which provides the right conditions for a different set of bacteria that operate in **aerobic** conditions. The drainage field includes an array of plastic pipes that are perforated to allow the effluent to drain out over a large area and seep slowly into the ground. In order to install a drainage field, it is first necessary to test the soil below, as it must be sufficiently permeable to allow the effluent to percolate away, but not so permeable that it flows too quickly and the soil is not able to filter out the pathogenic bacteria.

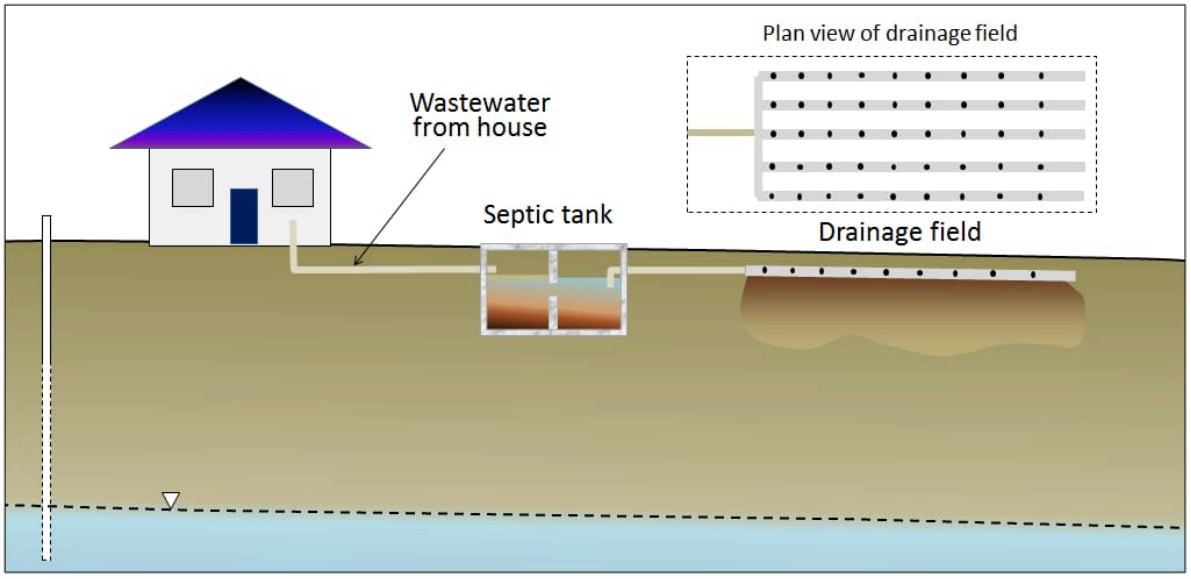


Figure 17.3.4.10 A typical septic system.

If they are properly installed and used, and if the sludge is periodically removed from the tank, a septic system should be effective in treating the sewage for decades. The anaerobic and aerobic bacteria should be able to break down the incoming waste and there should be little risk to the surface environment or groundwater. But many things can go wrong with a septic system, including the following:

- If inappropriate chemicals are added to the waste stream, they may interfere with the natural breakdown of the sewage.
- If the tank is not periodically pumped out, solids can get into the drainage field and compromise the drainage, resulting in the flow of effluent toward the surface.
- If the soil is either not sufficiently permeable or too permeable, the effluent will not drain away (and will start to pool at the surface) or it will drain too quickly.
- If the drainage field is constructed in an area where the water table is close to surface, some of the effluent is likely to flow into the groundwater without being treated.

Prevention and Mitigation of Groundwater Contamination

As illustrated in the landfill example above, there are two fairly simple ways to significantly reduce the chance and degree of groundwater contamination from surface sources. One is to prevent rainwater from infiltrating down to the water table and picking up contaminants; this can be achieved by simply capping or roofing over the landfill, mine tailings, or spill site. The second is to provide an impermeable barrier beneath the contaminant. Modern landfills and mine tailings impoundments are all built using some combination of clay and engineered plastic barriers. Both of these solutions — caps and liners — are subject to failure due to leaks.

Once contaminants are in the groundwater, the main form of remediation is to pump out the contaminated water and treat it at the surface. This can be a slow process, and preventing the contaminant from travelling significantly during this process can be accomplished by manipulating local groundwater flow through the extraction or injection of water at certain locations. Consider this in the exercise below.

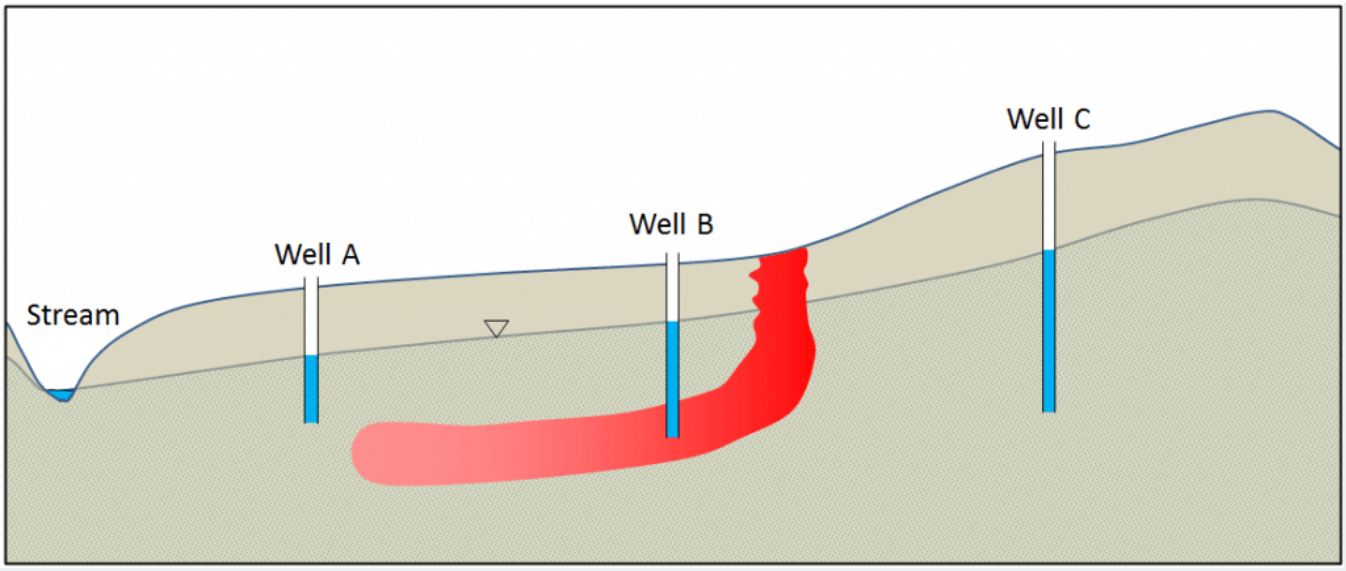


Figure 17.3.4.11 A plume of contamination (red) from a surface source

Figure 17.3.4.11 shows a groundwater contaminant plume in red. The source of the contamination has been removed but if the plume is not dealt with, it will eventually enter the stream and threaten the health of wildlife. Pumping the contaminant from well B for treatment will not be sufficient to prevent some of the contamination from making it to the stream.

What could you do at wells A and C to prevent this? Explain and use the diagram below to illustrate the expected changes to the water table and the movement of the plume.

See Appendix 3 for Exercise 14.6 answers.

Image Descriptions

Figure 17.3.4.11 image description: A cross section of a stream, three wells, and a contaminant plume. The stream marks the lowest point. As you move to the right, the ground and water table rises and you meet Well A, Well B, and Well C . The contaminant was introduced on the surface between Well B and Well C. It sunk down into the ground, entered the groundwater, and spread out to the left to surround the bottom part of Well B. It has continued to flow downhill towards Well A and the stream, but has yet to reach Well A. [Return to Figure 17.3.4.11]

Media Attributions

- Figure 17.3.4.1: © Steven Earle. CC BY.
- Figure 17.3.4.2: “Distribution of Arsenic” by the British Geological Survey (BGS) and the Department of Public Health

Engineering (DPHE). Reproduced with the permission of the British Geological Survey © UKRI. All rights reserved.
- Figure 17.3.4.3: Image from Newton, P. and Gilchrist, A. 2010. Technical summary of intrinsic vulnerability mapping methods of Vancouver Island, Vancouver Island Water Resources Vulnerability Mapping Project, Vancouver Island University, 45pp. Used with permission.
- Figures 14.4.4, 14.4.5: © Steven Earle. CC BY.
- Figure 17.3.4.6: "Trail Smelter in Year 1929" by Hughes Brothers. Public domain.
- Figures 14.4.7, 14.4.8, 14.4.9, 14.4.10, 14.4.11: © Steven Earle. CC BY.

---

1. "Adsorb" (with a "d") is not the same as "absorb" (with a "b"). Water can be absorbed by a sponge. Ions dissolved in water can be adsorbed onto—or desorbed from—the surfaces of clay minerals. ↵

This page titled 17.3.4: Groundwater Quality is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Steven Earle (BCCampus) via source content that was edited to the style and standards of the LibreTexts platform.

- 14.4: Groundwater Quality by Steven Earle is licensed CC BY 4.0. Original source: https://opentextbc.ca/physicalgeology2ed.

17.4: Mass Wasting

This section examines the downslope movement of rock, soil, and sediment under the influence of gravity. It considers the factors that determine slope stability and distinguishes among different types of mass wasting. The section also explores how these hazards can be recognized, monitored, and sometimes prevented or mitigated.

This page titled 17.4: Mass Wasting is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Robert Wagner.

17.4.1: Factors That Control Slope Stability

Mass wasting happens because tectonic processes have created uplift. Erosion, driven by gravity, is the inevitable response to that uplift, and various types of erosion, including mass wasting, have created slopes in the uplifted regions. Slope stability is ultimately determined by two factors: the angle of the slope and the strength of the materials on it.

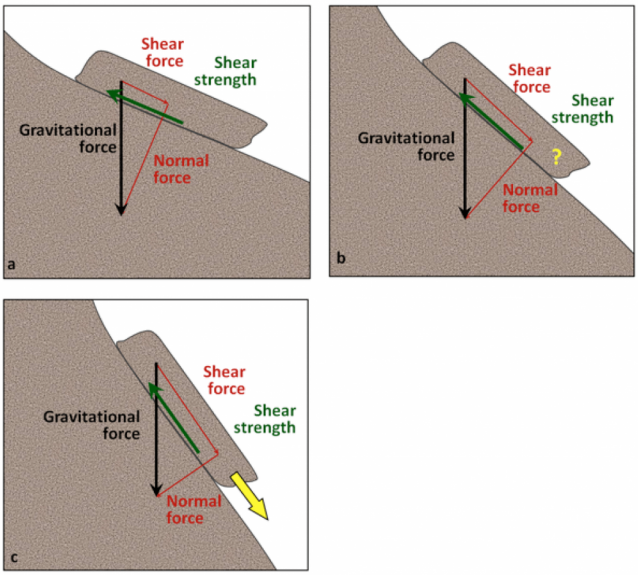


Figure 17.4.1.1 Differences in the shear and normal components of the gravitational force on slopes with differing steepness. The gravitational force is the same in all three cases. In (a) the shear force is substantially less than the shear strength, so the block should be stable. In (b) the shear force and shear strength are about equal, so the block may or may not move. In (c) the shear force is substantially greater than the shear strength, so the block is very likely to move.

Figure 15.2 shows a block of rock situated on a rock slope. The block is being pulled toward Earth's centre (vertically down) by gravity. We can split the vertical gravitational force into two components relative to the slope: one pushing the block down the slope (the **shear force**), and the other pushing into the slope (the **normal force**). The shear force, which wants to push the block down the slope, has to overcome the strength of the connection between the block and the slope, which may be quite weak if the block has split away from the main body of rock, or may be very strong if the block is still a part of the rock. This is the **shear strength**, and in Figure 17.4.1.1a, it is greater than the shear force, so the block should not move. In Figure 17.4.1.1b the slope is steeper and the shear force is approximately equal to the shear strength. The block may or may not move under these circumstances. In Figure 17.4.1.1c, the slope is steeper still, so the shear force is considerably greater than the shear strength, and the block will very likely move.

As already noted, slopes are created by uplift followed by erosion. In areas with relatively recent uplift (such as most of British Columbia and the western part of Alberta), slopes tend to be quite steep. This is especially true where glaciation has taken place because glaciers in mountainous terrain create steep-sided valleys. In areas without recent uplift (such as central Canada), slopes are less steep because hundreds of millions of years of erosion (including mass wasting) has made them that way. However, as we'll see, some mass wasting can happen even on relatively gentle slopes.

The strength of the materials on slopes can vary widely. Solid rocks tend to be strong, but there is a very wide range of rock strength. If we consider just the strength of the rocks, and ignore issues like fracturing and layering, then most crystalline rocks—like granite, basalt, or gneiss—are very strong, while some metamorphic rocks—like schist—are moderately strong. Sedimentary rocks have variable strength. Dolostone and some limestone are strong, most sandstone and conglomerate are moderately strong, and some sandstone and all mudstones are quite weak.

Fractures, metamorphic foliation, or bedding can significantly reduce the strength of a body of rock, and in the context of mass wasting, this is most critical if the planes of weakness are parallel to the slope and least critical if they are perpendicular to the slope. This is illustrated in Figure 17.4.1.2. At locations A and B the bedding is nearly perpendicular to the slope and the situation is relatively stable. At location D the bedding is nearly parallel to the slope and the situation is quite unstable. At location C the

bedding is nearly horizontal and the stability is intermediate between the other two extremes.

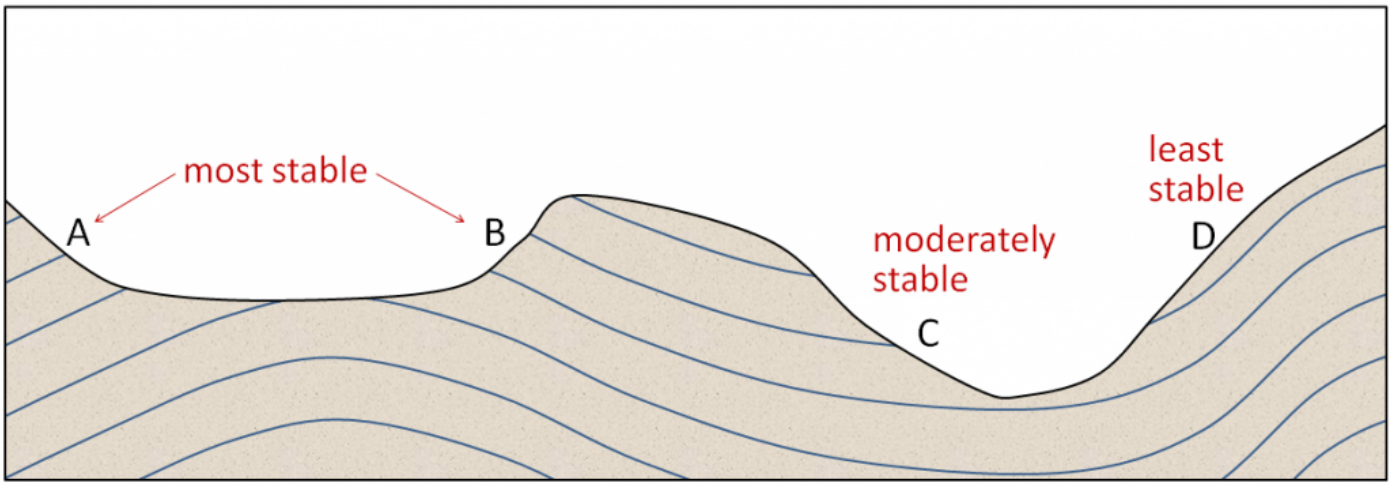


Figure 17.4.1.2 Relative stability of slopes as a function of the orientation of weaknesses (in this case bedding planes) relative to the slope orientations.

Internal variations in the composition and structure of rocks can significantly affect their strength. Schist, for example, may have layers that are rich in sheet silicates (mica or chlorite) and these will tend to be weaker than other layers. Some minerals tend to be more susceptible to weathering than others, and the weathered products are commonly quite weak (e.g., the clay formed from feldspar). The side of Johnson Peak that failed in 1965 (Hope Slide) is made up of chlorite schist (metamorphosed sea-floor basalt) that has feldspar-bearing sills within it (they are evident within the inset area of Figure 17.4.1.1). The foliation and the sills are parallel to the steep slope. The schist is relatively weak to begin with, and the feldspar in the sills, which has been altered to clay, makes it even weaker.

Unconsolidated sediments are generally weaker than sedimentary rocks because they are not cemented and, in most cases, have not been significantly compressed by overlying materials. This binding property of sediment is sometimes referred to as *cohesion*. Sand and silt tend to be particularly weak, clay is generally a little stronger, and sand mixed with clay can be stronger still. The deposits that make up the cliffs at Point Grey in Vancouver include sand, silt, and clay overlain by sand. As shown in Figure 17.4.1.3 (left) the finer deposits are relatively strong (they maintain a steep slope), while the overlying sand is relatively weak, and has a shallower slope that has recently failed. Glacial till—typically a mixture of clay, silt, sand, gravel, and larger clasts—forms and is compressed beneath tens to thousands of meters of glacial ice so it can be as strong as some sedimentary rock (Figure 17.4.1.3, right).



Figure 17.4.1.3 Left: Glacial outwash deposits at Point Grey, in Vancouver. The dark lower layer is made up of sand, silt, and clay. The light upper layer is well-sorted sand. Right: Glacial till on Quadra Island, B.C. The till is strong enough to have formed a near-vertical slope.

Apart from the type of material on a slope, the amount of water that the material contains is the most important factor controlling its strength. This is especially true for unconsolidated materials, like those shown in Figure 17.4.1.3, but it also applies to bodies of rock. Granular sediments, like the sand at Point Grey, have lots of spaces between the grains. Those spaces may be completely dry (filled only with air); or moist (often meaning that some spaces are water filled, some grains have a film of water around them, and small amounts of water are present where grains are touching each other); or completely saturated (Figure 17.4.1.4). Unconsolidated sediments tend to be strongest when they are moist because the small amounts of water at the grain boundaries hold the grains together with surface tension. Dry sediments are held together only by the friction between grains, and if they are well sorted or well rounded, or both, that cohesion is weak. Saturated sediments tend to be the weakest of all because the large amount of water actually pushes the grains apart, reducing the mount friction between grains. This is especially true if the water is under pressure.

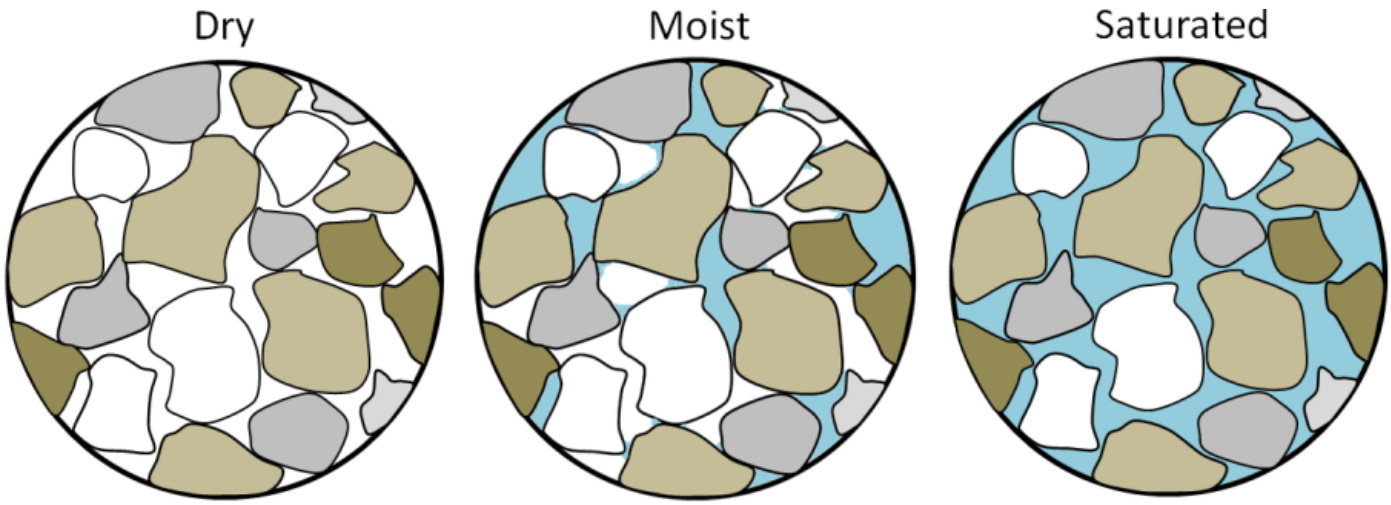


Figure 17.4.1.4 Depiction of dry, moist, and saturated sand.

Exercise 15.1 Sand and water



Figure 17.4.1.5

If you've ever been to the beach, you'll already know that sand behaves differently when it's dry than it does when it's wet, but it's worth taking a systematic look at the differences in its behaviour. Find about half a cup of clean, dry sand (or get some wet sand and dry it out), and then pour it from your hand onto a piece of paper. You should be able to make a cone-shaped pile that has a slope of around 30°. If you pour more sand on the pile, it will get bigger, but the slope should remain the same. Now add some water to the sand so that it is moist. An easy way to do this is to make it completely wet and then let the water drain away for a minute. You should be able to form this moist sand into a steep pile (with slopes of around 80°). Finally, put the same sand into

a cup and fill the cup with water so the sand is just covered. Swirl it around so that the sand remains in suspension, and then quickly tip it out onto a flat surface (best to do this outside). It should spread out over a wide area, forming a pile with a slope of only a few degrees.

Water will also reduce the strength of solid rock, especially if it has fractures, bedding planes, or clay-bearing zones. This effect is even more significant when the water is under pressure, which is why you'll often see holes drilled into rocks on road cuts to relieve this pressure. One of the hypotheses advanced to explain the 1965 Hope Slide is that the very cold conditions that winter caused small springs in the lower part of the slope to freeze over, preventing water from flowing out. It is possible that water pressure gradually built up within the slope, weakening the rock mass to the extent that the shear strength was no longer greater than the shear force.

Water also has a particular effect on clay-bearing materials. All clay minerals will absorb a little bit of water, and this reduces their strength. The **smectite** clays (such as the **bentonite** used in cat litter) can absorb a lot of water, and that water pushes the sheets apart at a molecular level and makes the mineral swell. Smectite that has expanded in this way has almost no strength; it is extremely slippery.

And finally, water can significantly increase the mass of the material on a slope, which increases the gravitational force pushing it down. A body of sediment that has 25% porosity and is saturated with water weighs approximately 13% more than it does when it is completely dry, so the gravitational shear force is also 13% higher. In the situation shown in Figure 17.4.1.1b, a 13% increase in the shear force could easily be enough to tip the balance between shear force and shear strength.

Mass-Wasting Triggers

In the previous section, we talked about the shear force and the shear strength of materials on slopes, and about factors that can reduce the shear strength. Shear force is primarily related to slope angle, and this does not change quickly. But shear strength can change quickly for a variety of reasons, and events that lead to a rapid reduction in shear strength are considered to be **triggers** for mass wasting.

An increase in water content is the most common mass-wasting trigger. This can result from rapid melting of snow or ice, heavy rain, or some type of event that changes the pattern of water flow on the surface. Rapid melting can be caused by a dramatic increase in temperature (e.g., in spring or early summer) or by a volcanic eruption. Heavy rains are typically related to major storms. Changes in water flow patterns can be caused by earthquakes, previous slope failures that dam up streams, or human structures that interfere with runoff (e.g., buildings, roads, or parking lots). An example of this is the deadly 2005 debris flow in North Vancouver (Figure 17.4.1.6). The 2005 failure took place in an area that had failed previously, and a report written in 1980 recommended that the municipal authorities and residents take steps to address surface and slope drainage issues. Little was done to improve the situation.



Figure 17.4.1.6 The debris flow in the Riverside Drive area of North Vancouver in January, 2005 happened during a rainy period, but was likely triggered by excess runoff related to the roads at the top of this slope and by landscape features, including a pool, in the area surrounding the house visible here.

In some cases, a *decrease* in water content can lead to failure. This is most common with clean sand deposits (e.g., the upper layer in Figure 17.4.1.3 (left)), which lose strength when there is no water to hold the grains together.

Freezing and thawing can also trigger some forms of mass wasting. More specifically, the thawing can release a block of rock that was attached to a slope by a film of ice.

One other process that can weaken a body of rock or sediment is shaking. The most obvious source of shaking is an earthquake, but shaking from highway traffic, construction, or mining will also do the job. Several deadly mass-wasting events (including snow avalanches) were triggered by the M7.8 earthquake in Nepal in April 2015.

Saturation with water and then seismic shaking led to the occurrence of thousands of slope failures in the Sapporo area of Hokkaido, Japan in September 2018, as shown on Figure 17.4.1.7. The area was drenched with rain from tropical storm Jebi on September 4th. On September 6th it was shaken by a M6.6 earthquake which triggered debris flows in the water-saturated volcanic materials on steep slopes. There were 41 deaths related to the slope failures.



Figure 17.4.1.7 Slope failures in the Sapporo area of Japan following a typhoon (Sept. 4th, 2018) and earthquake (Sept. 6th, 2018) (Before and after Landsat 8 images: left: July 2017, right: September 2018).

Media Attributions

- Figure 17.4.1.1, 15.1.2, 15.1.3, 15.1.4, 15.1.5: © Steven Earle. CC BY.
- Figure 17.4.1.6: © *The Province*. Used with permission.
- Figure 17.4.1.7: "Landslides in Hokkaido" by Lauren Dauphin, NASA Earth Observatory. Public domain.

This page titled 17.4.1: Factors That Control Slope Stability is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Steven Earle (BCCampus) via source content that was edited to the style and standards of the LibreTexts platform.

- 15.1: Factors That Control Slope Stability by Steven Earle is licensed CC BY 4.0. Original source: https://opentextbc.ca/physicalgeology2ed.

17.4.2: Classification of Mass Wasting

It's important to classify slope failures so that we can understand what causes them and learn how to mitigate their effects. The three criteria used to describe slope failures are:

- The type of material that failed (typically either bedrock or unconsolidated sediment)
- The mechanism of the failure (how the material moved)
- The rate at which it moved

The type of motion is the most important characteristic of a slope failure, and there are three different types of motion:

- If the material drops through the air, vertically or nearly vertically, it's known as a **fall**.
- If the material moves as a mass along a sloping surface (without internal motion within the mass), it's a **slide.**
- If the material has internal motion, like a fluid, it's a **flow**.

Unfortunately it's not normally that simple. Many slope failures involve two of these types of motion, some involve all three, and in many cases, it's not easy to tell how the material moved. The types of slope failure that we'll cover here are summarized in Table 15.1.

Table 15.1 Classification of slope failures based on type of material and type of motion

[Skip Table]

Failure Type	Type of Material	Type of Motion	Rate of Motion
Rock fall	Rock fragments	Vertical or near-vertical fall (plus bouncing in many cases)	Very fast (Greater than 10s of meters per second)
Rock slide	A large rock body	Motion as a unit along a planar surface (translational sliding)	Typically very slow (millimeters per year to centimeters per year), but some can be faster
Rock avalanche	A large rock body that slides and then breaks into small fragments	Flow (at high speeds the mass of rock fragments is suspended on a cushion of air)	Very fast (Greater than tens of meters per second)
Creep or solifluction	Soil or other overburden; in some cases, mixed with ice	Flow (although sliding motion may also occur)	Very slow (millimeters per year to centimeters per year)
Slump	Thick deposits (a metre to 10s of meters) of unconsolidated sediment	Motion as a unit along a curved surface (rotational sliding)	Slow (centimeters per year to meters per year)
Mudflow	Loose sediment with a significant component of silt and clay	Flow (a mixture of sediment and water moves down a channel)	Moderate to fast (centimeters per second to meters per second)
Debris flow	Sand, gravel, and larger fragments	Flow (similar to a mudflow, but typically faster)	Fast (meters per second)

Rock Fall

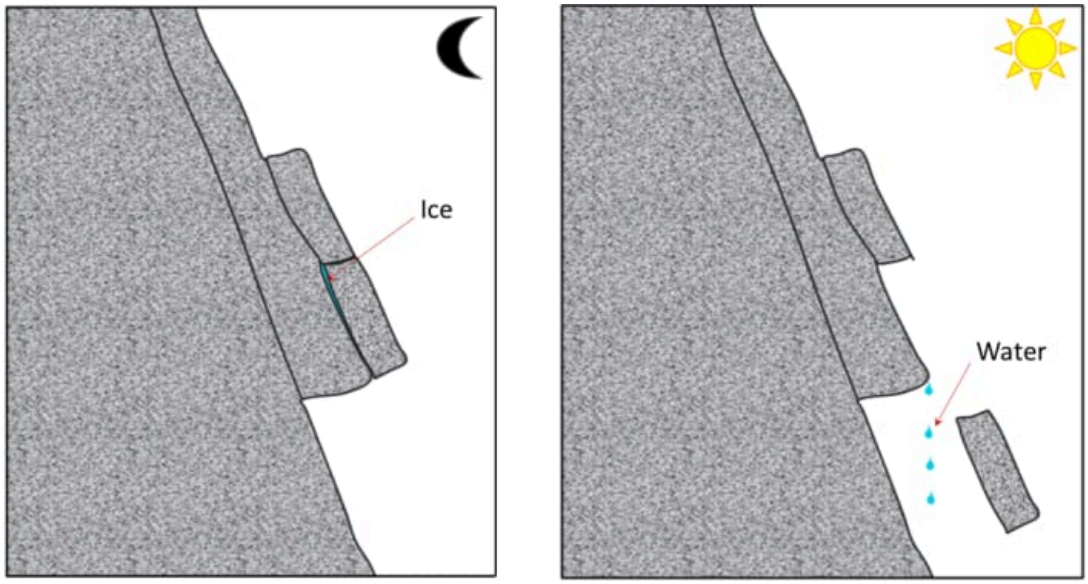


Figure 17.4.2.1 The contribution of freeze-thaw to rock fall.

Rock fragments can break off relatively easily from steep bedrock slopes, most commonly due to frost-wedging in areas where there are many freeze-thaw cycles per year. If you've ever hiked along a steep mountain trail on a cool morning, you might have heard the occasional fall of rock fragments onto a **talus slope**. This happens because the water between cracks freezes and expands overnight, and then when that same water thaws in the morning sun, the fragments that had been pushed beyond their limit by the ice fall to the slope below (Figure 17.4.2.1).

A typical talus slope, near Keremeos in southern B.C., is shown in Figure 17.4.2.2. In December 2014, a large block of rock split away from a cliff in this same area. It broke into smaller pieces that tumbled down the slope and crashed into the road, smashing the concrete barriers and gouging out large parts of the pavement. Luckily no one was hurt.



Figure 17.4.2.2 Left: A talus slope near Keremeos, B.C., formed by rock fall from the cliffs above. Right: The results of a rock fall onto a highway west of Keremeos in December 2014.

Rock Slide

A rock slide is the sliding motion of rock along a sloping surface. In most cases, the movement is parallel to a fracture, bedding, or metamorphic foliation plane, and it can range from very slow to moderately fast. The word **sackung** describes the very slow motion of a block of rock (millimeters per year to centimeters per year) on a slope. A good example is the Downie Slide north of

Revelstoke, B.C., which is shown in Figure 17.4.2.3. In this case, a massive body of rock is very slowly sliding down a steep slope along a plane of weakness that is approximately parallel to the slope. The Downie Slide, which was first recognized in the 1950s, prior to the construction of the Revelstoke Dam in the late 1970s, was moving very slowly at the time (a few centimeters per year). Geological engineers were concerned that the presence of water in the reservoir (visible in Figure 17.4.2.3) could further weaken the plane of failure, leading to an acceleration of the motion. The result would have been a catastrophic failure into the reservoir that would have sent a wall of water over the dam and into the community of Revelstoke. During the construction of the dam they tunneled into the rock at the base of the slide and drilled hundreds of drainage holes upward into the plane of failure. This allowed water to drain out so that the pressure was reduced, which reduced the rate of movement of the sliding block. BC Hydro monitors this site continuously; the slide block is currently moving more slowly than it was prior to the construction of the dam.



Figure 17.4.2.3 The Downie Slide, a sackung, on the shore of the Revelstoke Reservoir (above the Revelstoke Dam). The head scarp is visible at the top and a side-scarp along the left side.



Figure 17.4.2.4 Site of the 2008 rock slide at Porteau Cove. Notice the prominent fracture set parallel to the surface of the slope. The slope has been stabilized with rock bolts (visible near to the top of the photo) and holes have been drilled into the rock to improve drainage (one is visible in the lower right). Risk to passing vehicles from rock fall has been reduced by hanging mesh curtains (background).

In the summer of 2008, a large block of rock slid rapidly from a steep slope above Highway 99 near Porteau Cove (between Horseshoe Bay and Squamish). The block slammed into the highway and adjacent railway and broke into many pieces. The highway was closed for several days, and the slope was subsequently stabilized with rock bolts and drainage holes. As shown in Figure 17.4.2.4, the rock is fractured parallel to the slope, and this almost certainly contributed to the failure. However, it is not actually known what triggered this event as the weather was dry and warm during the preceding weeks, and there was no

significant earthquake in the region.

Rock Avalanche

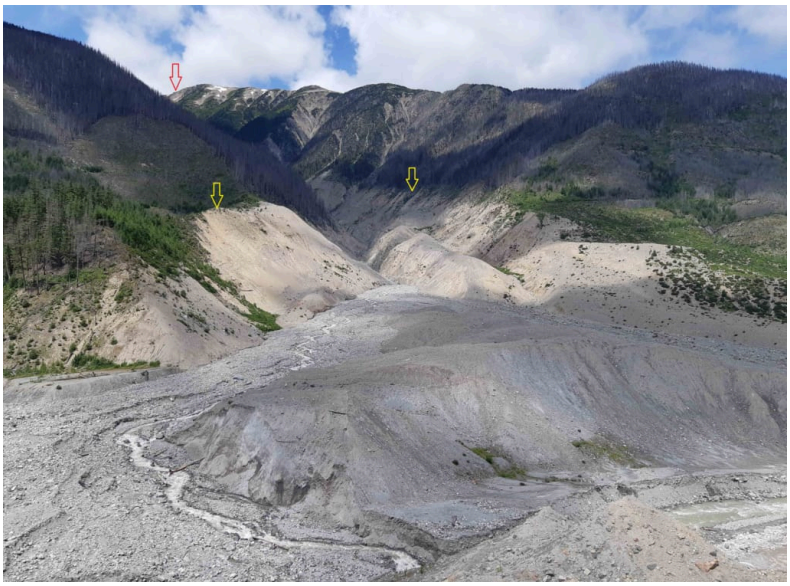


Figure 17.4.2.5 The August 2010 Mount Meager rock avalanche, showing where the slide originated (red arrow, 4 km upstream) and its path down a steep narrow valley. The yellow arrows show how far up the valley the avalanche extended.

If a rock slides and then starts moving quickly (meters per second), the rock is likely to break into many small pieces, and at that point it turns into a **rock avalanche**, in which the large and small fragments of rock move in a fluid manner supported by a cushion of air within and beneath the moving mass. The 1965 Hope Slide (Figure 17.4.2.1) was a rock avalanche, as was the famous 1903 Frank Slide in southwestern Alberta. The 2010 slide at Mount Meager (west of Lillooet) was also a rock avalanche, and rivals the Hope Slide as the largest slope failure in Canada during historical times (Figure 17.4.2.5).

Creep or Solifluction

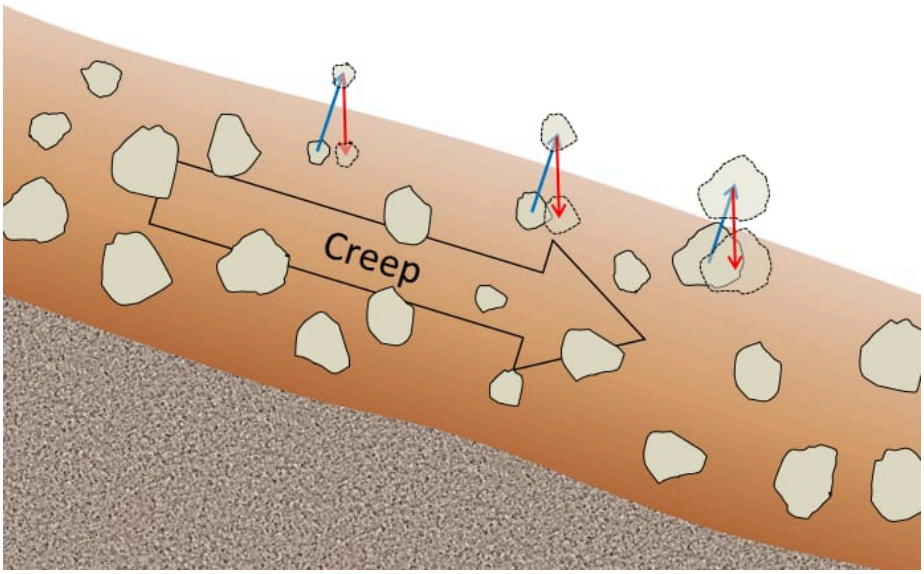


Figure 17.4.2.6 A depiction of the contribution of freeze-thaw to creep. The blue arrows represent uplift caused by freezing in the wet soil underneath, while the red arrows represent depression by gravity during thawing. The uplift is perpendicular to the slope, while the drop is vertical.

The very slow—millimeters per year to centimeters per year—movement of soil or other unconsolidated material on a slope is

known as creep. **Creep**, which normally only affects the upper several centimeters of loose material, is typically a type of very slow flow, but in some cases, sliding may take place. Creep can be facilitated by freezing and thawing because, as shown in Figure 17.4.2.6, particles are lifted perpendicular to the surface by the growth of ice crystals within the soil, and then let down vertically by gravity when the ice melts. The same effect can be produced by frequent wetting and drying of the soil. In cold environments, **solifluction** is a more intense form of freeze-thaw-triggered creep.

Creep is most noticeable on moderate-to-steep slopes where trees, fence posts, or grave markers are consistently leaning in a downhill direction. In the case of trees, they try to correct their lean by growing upright, and this leads to a curved lower trunk known as a "pistol butt." An example is shown on Figure 17.4.2.7.



Figure 17.4.2.7 Pistol-butt shaped trees on a slope that is experiencing creep

Slump

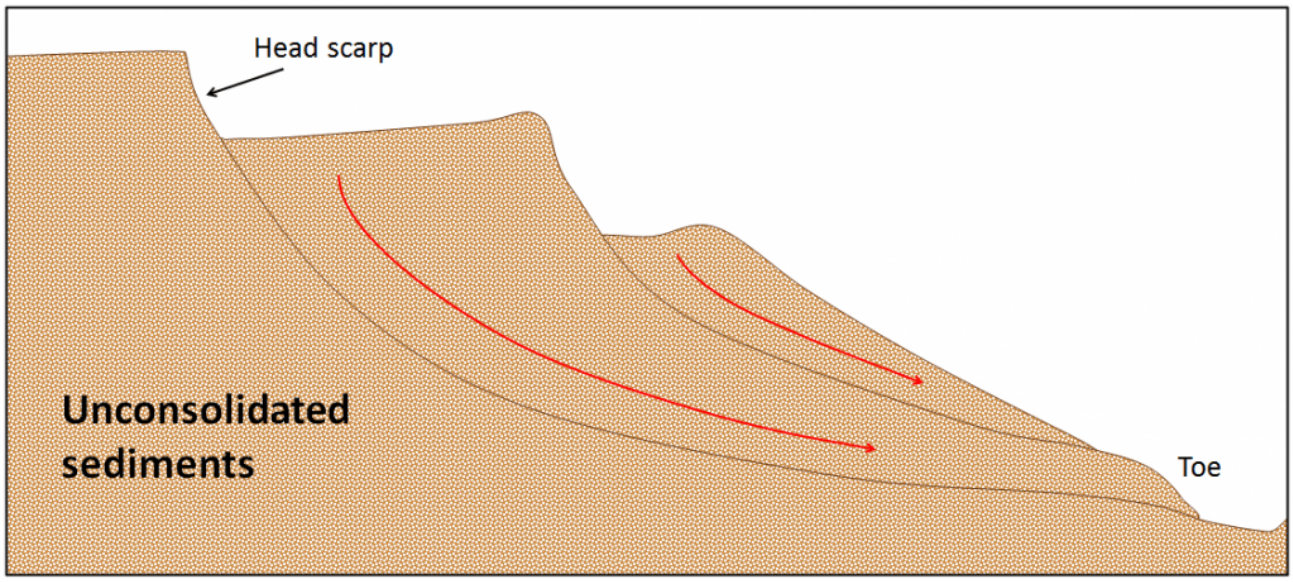


Figure 17.4.2.8 A depiction of the motion of unconsolidated sediments in an area of slumping.

Slump is a type of slide (movement as a mass) that takes place within thick unconsolidated deposits (typically thicker than 10 meters). Slumps involve movement along one or more curved failure surfaces, with downward motion near the top and outward motion toward the bottom (Figure 17.4.2.8). They are typically caused by an excess of water within these materials on a steep slope.



Figure 17.4.2.9 A slump along the banks of a small coulee near Lethbridge, Alberta. The main head-scarp is clearly visible at the top, and a second smaller one is visible about one-quarter of the way down. The toe of the slump is being eroded by the seasonal stream that created the coulee.

An example of a slump in the Lethbridge area of Alberta is shown in Figure 17.4.2.9. This feature has likely been active for many decades, and moves a little more whenever there are heavy spring rains and significant snowmelt runoff. The toe of the slump is failing because it has been eroded by the small stream at the bottom.

Mudflows and Debris Flows



Figure 17.4.2.10 A slump (left) and an associated mudflow (centre) at the same location as Figure 17.4.2.9, near Lethbridge, Alberta.

As you saw in Exercise 15.1, when a mass of sediment becomes completely saturated with water, the mass loses strength, to the extent that the grains are pushed apart, and it will flow, even on a gentle slope. This can happen during rapid spring snowmelt or heavy rains, and is also relatively common during volcanic eruptions because of the rapid melting of snow and ice. (A mudflow or debris flow on a volcano or during a volcanic eruption is a *lahar*.) If the material involved is primarily sand-sized or smaller, it is known as a mudflow, such as the one shown in Figure 17.4.2.10.

If the material involved is gravel sized or larger, it is known as a debris flow. Because it takes more gravitational energy to move larger particles, a debris flow typically forms in an area with steeper slopes and more water than does a mudflow. In many cases, a debris flow takes place within a steep stream channel, and is triggered by the collapse of bank material into the stream. This creates a temporary dam, and then a major flow of water and debris when the dam breaks. This is the situation that led to the fatal debris flow at Johnsons Landing, B.C., in 2012. A typical west-coast debris flow is shown in Figure 17.4.2.11. This event took place in November 2006 in response to very heavy rainfall. There was enough energy to move large boulders and to knock over large trees.



Figure 17.4.2.11 The lower part of debris flow within a steep stream channel near Buttle Lake, B.C., in November 2006.

Exercise 15.2 Classifying slope failures

These four photos show some of the different types of slope failures described above. Try to identify each types and provide some criteria to support your choice.



Figure 17.4.2.12a



Figure 17.4.2.12b



Figure 17.4.2.12c



Figure 17.4.2.12d

a:	b:
c:	d:

See Appendix 3 for Exercise 15.2 answers.

Media Attributions

- Figure 17.4.2.1, 15.2.2: © Steven Earle. CC BY.
- Figure 17.4.2.3: © Google Earth. Used with permission for non-commerical purposes.
- Figure 17.4.2.4: © Steven Earle. CC BY.
- Figure 17.4.2.5: “2010 Mt. Meager rock avalanche” © Isaac Earle. CC BY.
- Figure 17.4.2.6, 15.2.7, 15.2.8, 15.2.9, 15.2.10, 15.2.11, 15.2.12: © Steven Earle. CC BY.

This page titled 17.4.2: Classification of Mass Wasting is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Steven Earle (BCCampus) via source content that was edited to the style and standards of the LibreTexts platform.

- 15.2: Classification of Mass Wasting by Steven Earle is licensed CC BY 4.0. Original source: https://opentextbc.ca/physicalgeology2ed.

17.4.3: Preventing, Delaying, Monitoring, and Mitigating Mass Wasting

As already noted, we cannot prevent mass wasting in the long term as it is a natural and ongoing process; however, in many situations there are actions that we can take to reduce or mitigate its damaging effects on people and infrastructure. Where we can neither delay nor mitigate mass wasting, we should consider moving out of the way.

Preventing and Delaying Mass Wasting

It is comforting to think that we can prevent some effects of mass wasting by mechanical means, such as the rock bolts in the road cut at Porteau Cove (Figure 17.4.3.4), or the drill holes used to drain water out of a slope, as was done at the Downie Slide (Figure 17.4.3.3), or the building of physical barriers, such as retaining walls. What we have to remember is that the works of humans are mostly insignificant compared to the works of nature. The rock bolts in the road cut at Porteau Cove will slowly start to corrode after a few years, and within a few decades, many of them will begin to lose their strength. Unless they are replaced, they will no longer support that slope. Likewise, drainage holes at the Downie Slide will eventually become plugged with sediment and chemical precipitates, and unless they are periodically unplugged, their effectiveness will decrease. Eventually, unless new holes are drilled, the drainage will be so compromised that the slide will start to move again. This is why careful long-term slope monitoring by geological and geotechnical engineers is important at these sites. The point here is that our efforts to "prevent" mass wasting are only as good as our resolve to maintain those preventive measures.

Delaying mass wasting is a worthy endeavour, of course, because during the time that the measures are still effective they can save lives and reduce damage to property and infrastructure. The other side of the coin is that we must be careful to avoid activities that could make mass wasting more likely. One of the most common anthropogenic causes of mass wasting is road construction, and this applies both to remote gravel roads built for forestry and mining and large urban and regional highways. Road construction is a potential problem for two reasons. First, creating a flat road surface on a slope inevitably involves creating a cut bank that is steeper than the original slope. This might also involve creating a filled bank that is both steeper and weaker than the original slope (Figure 17.4.3.12). Second, roadways typically cut across natural drainage features, and unless great care is taken to reroute the runoff water and prevent it from forming concentrated flows, oversaturating fill of materials can result. A specific example of the contribution of construction-related impeded drainage to slope instability was shown earlier in Figure 17.4.3.6.

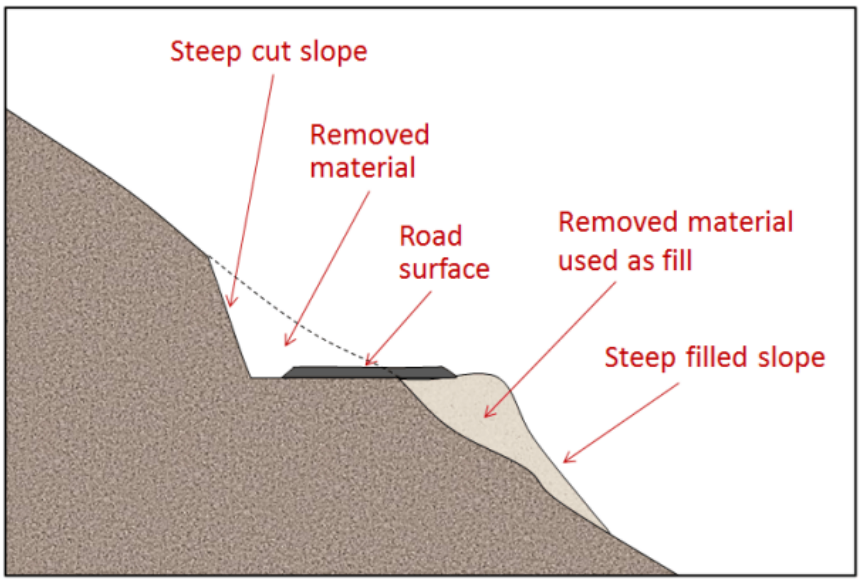


Figure 17.4.3.1 An example of a road constructed by cutting into a steep slope. The cut material was moved to outside part of the road to act as fill.

Apart from water issues, engineers building roads and other infrastructure on bedrock slopes have to be acutely aware of the

geology, and especially of any weaknesses or discontinuities in the rock related to bedding, fracturing, or foliation. If possible, situations like that at Porteau Cove (Figure 17.4.3.3) should be avoided — by building somewhere else — rather than trying to stitch the slope back together with rock bolts.

It is widely believed that construction of buildings on the tops of steep slopes can contribute to the instability of the slope. This is probably true, but in most cases that is not because of the weight of the building. As you'll see by completing Exercise 15.3, a typical house isn't usually heavier than the fill that was removed from the hole in the ground made to build it. A more likely contributor to instability of the slope around a building is the effect that it and the changes made to the surrounding area have on drainage.

Exercise 15.3 How much does a house weigh, and can it contribute to a slope failure?

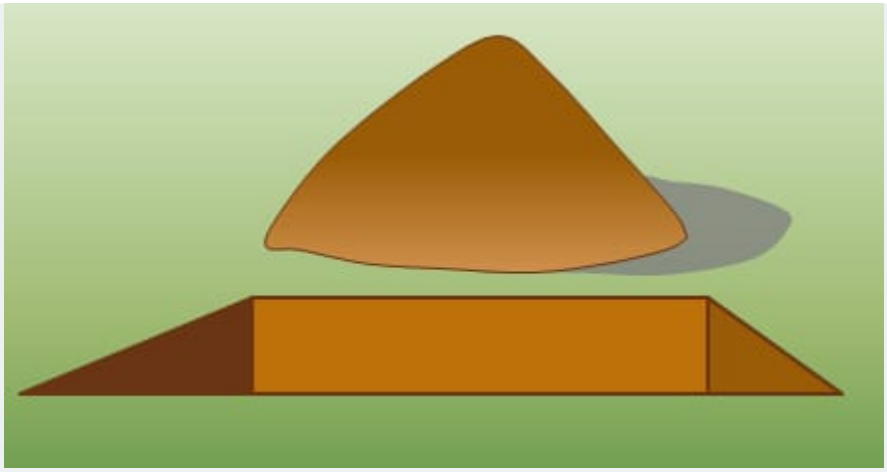


Figure 17.4.3.2

It is commonly believed that building a house (or some other building) at the top of a slope will add a lot of extra weight to the slope, which could contribute to slope failure. But what does a house actually weigh? A typical 150 square metre (approximately 1,600 square feet) wood-frame house with a basement and a concrete foundation weighs about 145 tonnes. But most houses are built on foundations that are excavated into the ground. This involves digging a hole and taking some material away, so we need to subtract what that excavated material weighs. Assuming our 150 square metre house required an excavation that was 15 meters by 11 meters by 1 metre deep, that's 165 cubic meters of "dirt," which typically has a density of about 1.6 tonnes per cubic metre.

Calculate the weight of the soil that was removed and compare that with the weight of the house and its foundation.

If you're thinking that building a bigger building is going to add more weight, consider that bigger buildings need bigger and deeper excavations, and in many cases the excavations will be into solid rock, which is much heavier than surficial materials.

Much more important than worrying about the weight of a building is to consider how a building might change the drainage on a slope. There are a number ways. Water can be collected by roofs, go into downspouts, and form concentrated flows that are directed onto or into the slope. Likewise drainage from nearby access roads, lawn irrigation, leaking pools, and septic systems can all alter the surface and groundwater flow in a slope.

See Appendix 3 for Exercise 15.3 answers.

Monitoring Mass Wasting



Figure 17.4.3.3 Part of a motion-monitoring device at the Checkerboard Slide near Revelstoke, B.C. The lower end of the cable is attached to a block of rock that is unstable. Any incremental motion of that block will move the cable, which will be detectable on this device.

In some areas, it is necessary to establish warning systems so that we know if conditions have changed at a known slide area, or if a rapid failure, such as a debris flow, is actually on its way downslope. The Downie Slide above the Revelstoke Resevoir is monitored 24/7 with a range of devices, such as inclinometers (slope-change detectors), bore-hole motion sensors, and GPS survey instruments. A simple mechanical device for monitoring the nearby Checkerboard Slide (which is also above the Revelstoke Reservoir) is shown in Figure 17.4.3.3. Both of these are very slow-moving rock slides, but it's critically important to be able to detect changes in their rates of motion because at both of these locations a rapid failure would result in large bodies of rock plunging into the reservoir and sending a wall of water over the Revelstoke Dam, potentially destroying the nearby town of Revelstoke.

Mount Rainier, a glacier-covered volcano in Washington State, has the potential to produce massive mudflows or debris flows (lahars) with or without a volcanic eruption. Over 100,000 people in the Tacoma, Puyallup, and Sumner areas are in harm's way because many of them currently reside on deposits from past lahars (Figure 17.4.3.3). In 1998, a network of acoustic monitors was established around Mount Rainier. The monitors are embedded in the ground adjacent to expected lahar paths. They are intended to provide warnings to emergency officials, and when a lahar is detected, the residents of the area will have anywhere from 40 minutes to three hours to get to safe ground.



Figure 17.4.3.4 Mount Rainier, Washington, from Tacoma.

Mitigating the Impacts of Mass Wasting

In situations where we can't predict, prevent, or delay mass-wasting, some effective measures can be taken to minimize the associated risk. For example, many highways in B.C. and western Alberta have avalanche shelters like that shown in Figure 17.4.3.5 . In some parts of the world, similar features have been built to protect infrastructure from other types of mass wasting.



Figure 17.4.3.5 A snow avalanche shelter on the Coquihalla Highway. The expected path of the avalanche is the steep un-treed slope above.

Debris flows are inevitable, unpreventable, and unpredictable in many parts of B.C., but nowhere more so than along the Sea-to-Sky Highway between Horseshoe Bay and Squamish. The results have been deadly and expensive many times in the past. It would be very expensive to develop a new route in this region, so provincial authorities have taken steps to protect residents and traffic on the highway and the railway. Debris-flow defensive structures have been constructed in several drainage basins, as shown in Figure 17.4.3.6. One strategy is to allow the debris to flow quickly through to the ocean along a smooth channel. Another is to capture the debris within a constructed basin that allows the excess water to continue through, but catches the debris materials.



Figure 17.4.3.6 Two strategies for mitigating debris flows on the Sea-to-Sky Highway. Left: A concrete –lined channel on Alberta Creek allows debris to flow quickly through to the ocean. Right: A debris-flow catchment basin on Charles Creek. In 2010, a debris flow filled the basin to the level of the dotted white line.

Finally, in situations where we can't do anything to delay, predict, contain, or mitigate slope failures, we simply have to have the sense to stay away. There is a famous example of this in B.C. at a site known as Garibaldi, 25 kilometers south of Whistler. In the early 1980s the village of Garibaldi had a population of about 100, with construction underway on some new homes, and plans for many more. In the months that followed the deadly 1980 eruption of Mount St. Helens in Washington State, the B.C. Ministry of Transportation commissioned a geological study that revealed that a steep cliff known as The Barrier (Figure 17.4.3.7) had collapsed in 1855, leading to a large rock avalanche, and that it was likely to collapse again unpredictably, putting the village of Garibaldi at extreme risk. In an ensuing court case, it was ruled that the Garibaldi site was not a safe place for people to live. Those who already had homes there were compensated, and everyone else was ordered to leave.

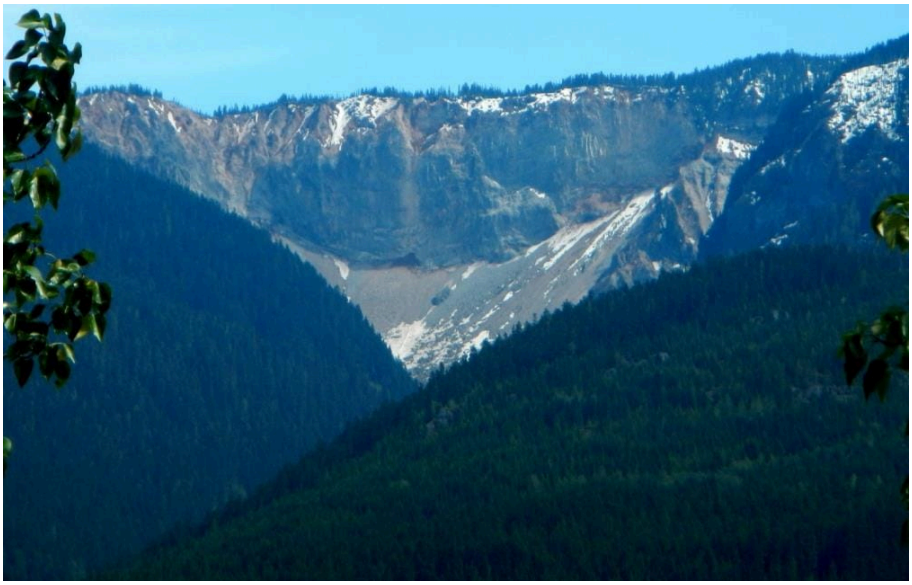


Figure 17.4.3.7 The Barrier, south of Whistler, B.C., was the site of a huge rock avalanche in 1855, which extended from the cliff visible here 4 kilometers down the valley and across the current location of the Sea-to-Sky Highway and the Cheakamus River.

Media Attributions

- Figure 17.4.3.1, 15.3.2, 15.3.3, 15.3.5, 15.3.6, 15.3.7: © Steven Earle. CC BY.
- Figure 17.4.3.4: "Mount Rainier over Tacoma" by Lyn Topinka (USGS). Public domain.

This page titled 17.4.3: Preventing, Delaying, Monitoring, and Mitigating Mass Wasting is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Steven Earle (BCCampus) via source content that was edited to the style and standards of the LibreTexts platform.

- 15.3: Preventing, Delaying, Monitoring, and Mitigating Mass Wasting by Steven Earle is licensed CC BY 4.0. Original source: https://opentextbc.ca/physicalgeology2ed.

18: GLACIERS, EARTH'S HISTORY, AND OTHER WORLDS

18.1: How Glaciers Work
18.2: Glacial Erosion
18.3: Glacial Deposition
18.4: Earth’s First 2 Billion Years
18.5: Are There Other Earths?

Chapter 18: Glaciers, Earth’s History, and Other Worlds

Earth is a dynamic planet, but some of its most important changes occur over timescales far longer than a human lifetime. Glaciers provide an especially visible example. Although ice may appear stationary, glaciers flow across the landscape, eroding rock, transporting sediment, and leaving behind distinctive landforms. By studying those features, geologists can reconstruct environments that existed thousands or even millions of years ago.

That ability to read evidence from the present also allows us to investigate much deeper into Earth’s past. The geological record becomes increasingly incomplete as we approach the planet’s earliest history, so understanding the first billions of years of Earth requires scientists to combine evidence from rocks, minerals, geochemistry, and planetary science. The result is a picture of a planet that has changed enormously since its formation.

This final chapter also brings us back to a question that connects Earth science with the astronomy we studied earlier in the course: **Is Earth unusual?** We now know much more about what makes Earth work as a planet—from its atmosphere and water to its rocks, interior, and changing surface. Those ideas give us a basis for considering other planets and asking whether worlds with Earth-like conditions might exist elsewhere.

In that sense, this chapter closes the circle of the course. We began by looking outward at the universe and our place within it. We finish by using what we have learned about our own planet to look outward again, this time with a much better understanding of what it actually means for a world to be *Earth-like*.

This page titled 18: Glaciers, Earth's History, and Other Worlds is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by Robert Wagner.

18.1: How Glaciers Work



Figure 18.1.1 Part of the continental ice sheet in Greenland, with some outflow alpine glaciers in the foreground.

There are two main types of glaciers. **Continental glaciers** cover vast areas of land in extreme polar regions, including Antarctica and Greenland (Figure 18.1.1). **Alpine glaciers** (a.k.a. valley glaciers) originate on mountains, mostly in temperate and polar regions (Figure 18.1.1), but even in tropical regions if the mountains are high enough.

Earth's two great continental glaciers, on Antarctica and Greenland, comprise about 99% of all of the world's glacial ice, and approximately 68% of all of Earth's fresh water. As is evident from Figure 18.1.2, the Antarctic Ice Sheet is vastly bigger than the Greenland Ice Sheet; it contains about 17 times as much ice. If the entire Antarctic Ice Sheet were to melt, sea level would rise by about 80 m and most of Earth's major cities would be completely submerged.

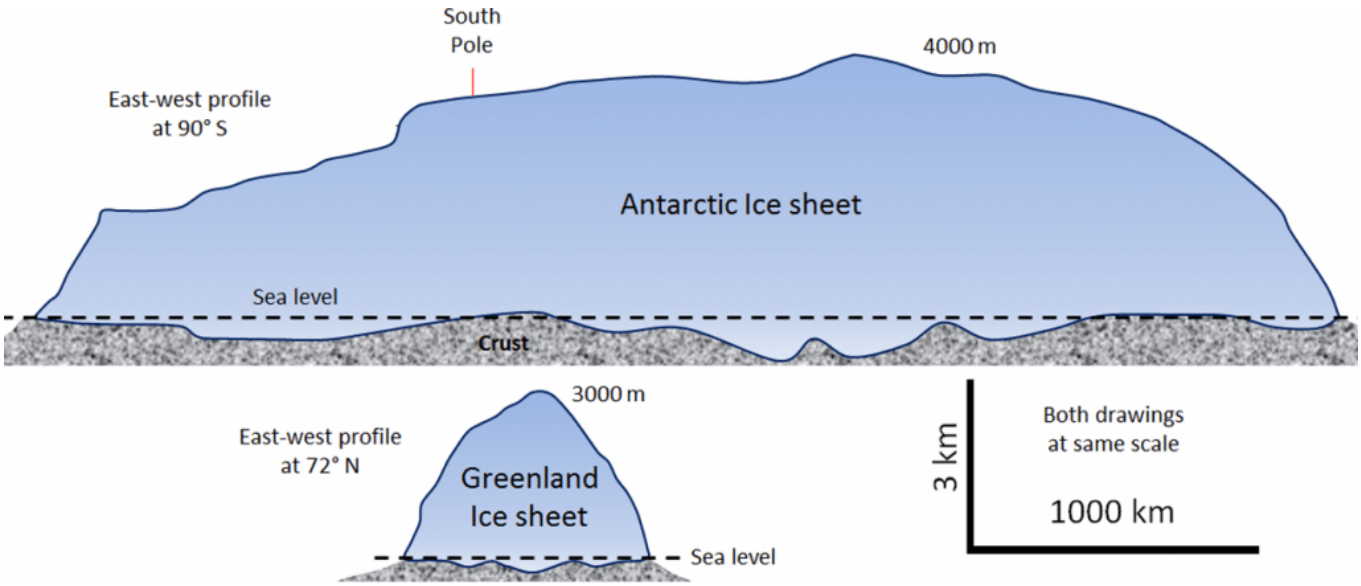


Figure 18.1.2 Simplified cross-sectional profiles the continental ice sheets in Greenland and Antarctica – both drawn to the same scale.

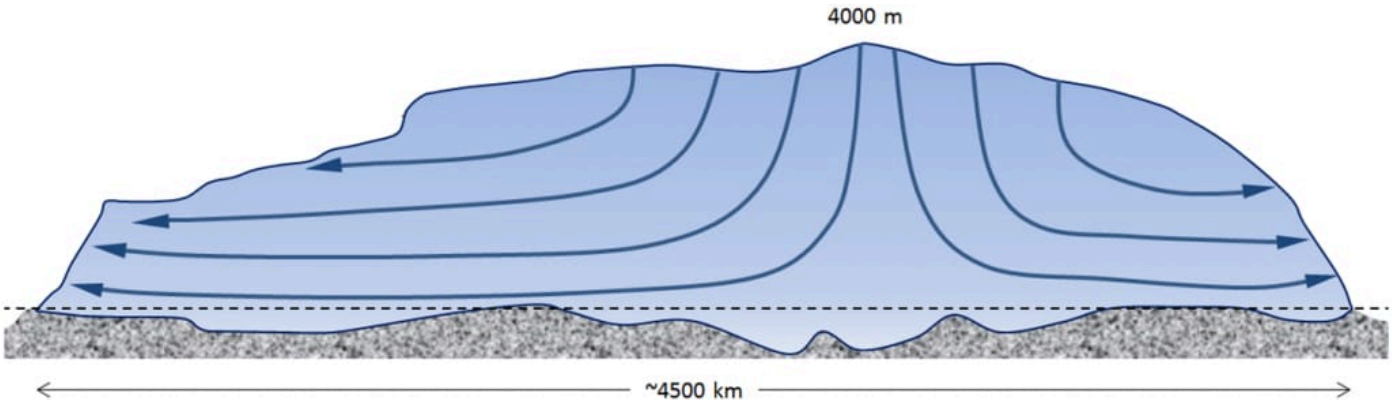


Figure 18.1.3 Schematic ice-flow diagram for the Antarctic Ice Sheet.

Continental glaciers do not flow “downhill” because the large areas that they cover are generally flat. Instead, ice flows from the region where it is thickest toward the edges where it is thinner, as shown in Figure 18.1.3. This means that in the central thickest parts, the ice flows almost vertically down toward the base, while in the peripheral parts, it flows out toward the margins. In continental glaciers like Antarctica and Greenland, the thickest parts (4,000 m and 3,000 m respectively) are the areas where the rate of snowfall and therefore the rate of ice accumulation are highest.

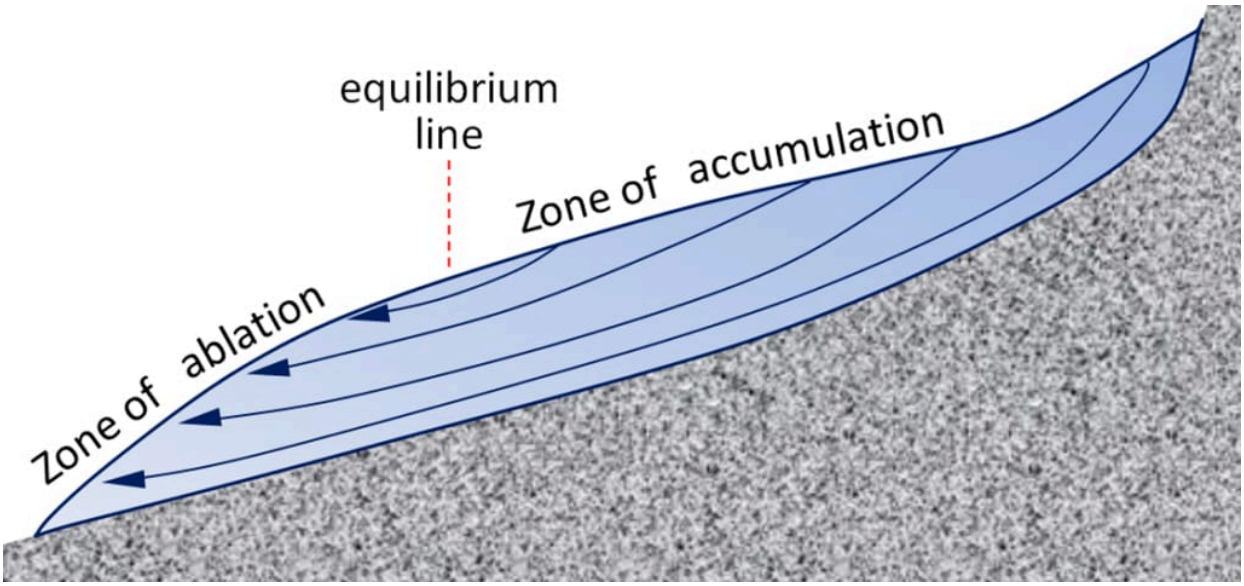


Figure 18.1.4 Schematic ice-flow diagram for an alpine glacier.

The flow of alpine glaciers is primarily controlled by the slope of the land beneath the ice (Figure 18.1.4). In the **zone of accumulation**, the rate of snowfall is greater than the rate of melting. In other words, not all of the snow that falls each winter melts during the following summer, and the ice surface is always covered with snow. In the **zone of ablation**, more ice melts than accumulates as snow. The **equilibrium line** marks the boundary between the zones of accumulation (above) and ablation (below).

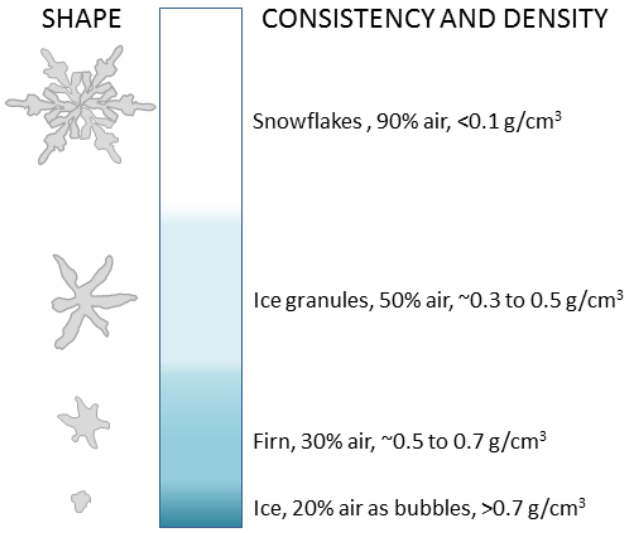


Figure 18.1.5 Steps in the process of formation of glacial ice from snow, granules, and firn.

Above the equilibrium line of a glacier, not all of the winter snow melts in the following summer, so snow gradually accumulates. The snow layer from each year is covered and compacted by subsequent snow, and it is gradually compressed and turned into **firn** within which the snowflakes lose their delicate shapes and become granules. With more compression, the granules are pushed together and air is squeezed out. Eventually the granules are "welded" together to create glacial ice (Figure 18.1.5). Downward percolation of water from melting taking place at the surface contributes to the process of ice formation.

The equilibrium line of a glacier near Whistler, B.C., is shown in Figure 18.1.6. Below that line, in the zone of ablation, bare ice is exposed because last winter's snow has all melted; above that line, the ice is still mostly covered with snow from last winter. The position of the equilibrium line changes from year to year as a function of the balance between snow accumulation in the winter and snowmelt during the summer. More winter snow and less summer melting obviously favours the advance of the equilibrium line (and of the glacier's leading edge), but of these two variables, it is the summer melt that matters most to a glacier's budget. Cool summers promote glacial advance and warm summers promote glacial retreat.



Figure 18.1.6 The approximate location of the equilibrium line (red) in September 2013 on the Overlord Glacier, near Whistler, B.C.

Glaciers move because the surface of the ice is sloped. This generates a stress on the ice, which is proportional to the slope and to the depth below the surface. As shown in Figure 18.1.6, the stresses are quite small near the ice surface but much larger at depth, and also greater in areas where the ice surface is relatively steep. Ice will deform, meaning that it will behave in a plastic manner, at stress levels of around 100 kilopascals; therefore, in the upper 50 m to 100 m of the ice (above the dashed red line),

flow is not plastic (the ice is rigid), while below that depth, ice is plastic and will flow.

When the lower ice of a glacier flows, it moves the upper ice along with it, so although it might seem from the stress patterns (red numbers and red arrows) shown in Figure 18.1.7 that the lower part moves the most, in fact while the lower part deforms (and flows) and the upper part doesn't deform at all, the upper part moves the fastest because it is pushed along by the lower ice.

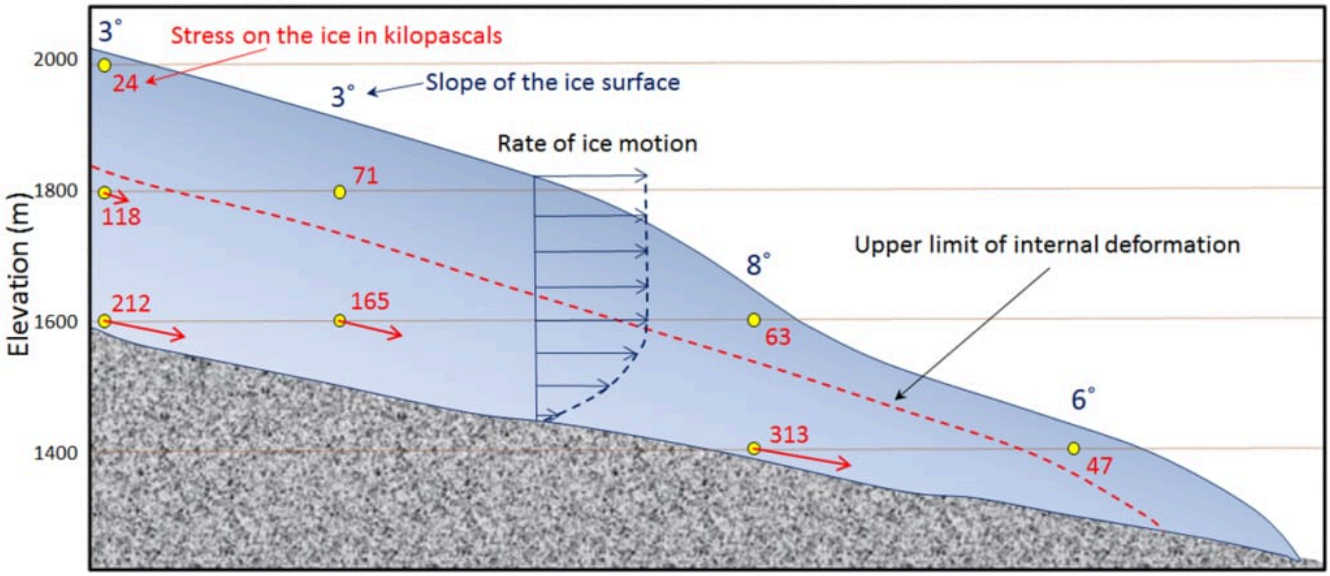


Figure 18.1.7 Stress within a valley glacier (red numbers) as determined based on the slope of the ice surface and the depth within the ice. The ice will deform and flow where the stress is greater than 100 kilopascals, and the relative extent of that deformation is depicted by the red arrows. Any deformation motion in the lower ice will be transmitted to the ice immediately above it, so although the red stress arrows get shorter toward the top, the ice velocity increases upward (blue arrows). The upper ice (above the red dashed line) does not flow, but it is pushed along with the lower ice.

The plastic lower ice of a glacier can flow like a very viscous fluid, and can therefore flow over irregularities in the base of the ice and around corners. However, the upper rigid ice cannot flow in this way, and because it is being carried along by the lower ice, it tends to crack where the lower ice has to flex. This leads to the development of **crevasses** in areas where the rate of flow of the plastic ice is changing. In the area shown in Figure 18.1.8, for example, the glacier is speeding up over the steep terrain, and the rigid surface ice has to crack to account for the change in velocity.



Figure 18.1.8 Crevasses on Overlord Glacier in the Whistler area, B.C.

The plastic lower ice of a glacier can flow like a very viscous fluid, and can therefore flow over irregularities in the base of the ice

and around corners. However, the upper rigid ice cannot flow in this way, and because it is being carried along by the lower ice, it tends to crack where the lower ice has to flex. This leads to the development of **crevasses** in areas where the rate of flow of the plastic ice is changing. In the area shown in Figure 18.1.8, for example, the glacier is speeding up over the steep terrain, and the rigid surface ice has to crack to account for the change in velocity

The base of a glacier can be cold (below the freezing point of water) or warm (above the freezing point). If it is warm, there will likely be a film of water between the ice and the material underneath, and the ice will be able to slide over that surface. This is known as **basal sliding** (Figure 18.1.9, left). If the base is cold, the ice will be frozen to the material underneath and it will be stuck—unable to slide along its base. In this case, all of the movement of the ice will be by internal flow.

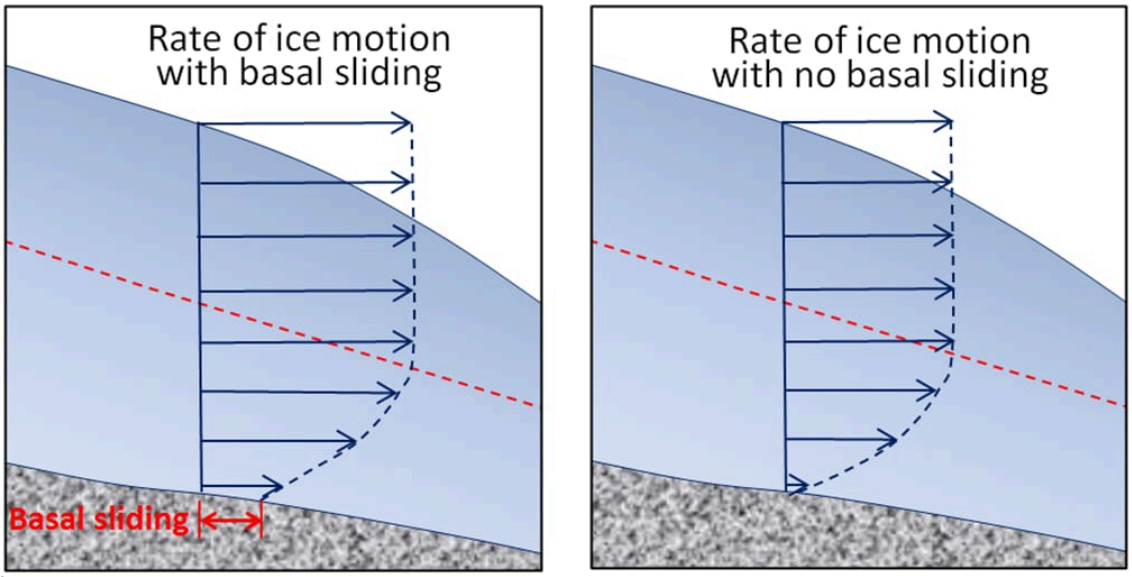


Figure 18.1.9 Differences in glacial ice motion with basal sliding (left) and without basal sliding (right). The dashed red line indicates the upper limit of plastic internal flow.

One of the factors that affects the temperature at the base of a glacier is the thickness of the ice. Ice is a good insulator. The slow transfer of heat from Earth's interior provides enough heat to warm up the base if the ice is thick, but not enough if it is thin and that heat can escape. It is typical for the leading edge of an alpine glacier to be relatively thin (see Figure 18.1.7), so it is common for that part to be frozen to its base while the rest of the glacier is still sliding. This is illustrated in Figure 18.1.10 for the Athabasca Glacier. Because the leading edge of the glacier is stuck to its frozen base, while the rest continues to slide, the ice coming from behind has pushed (or thrust) itself over top of the part that is stuck fast.

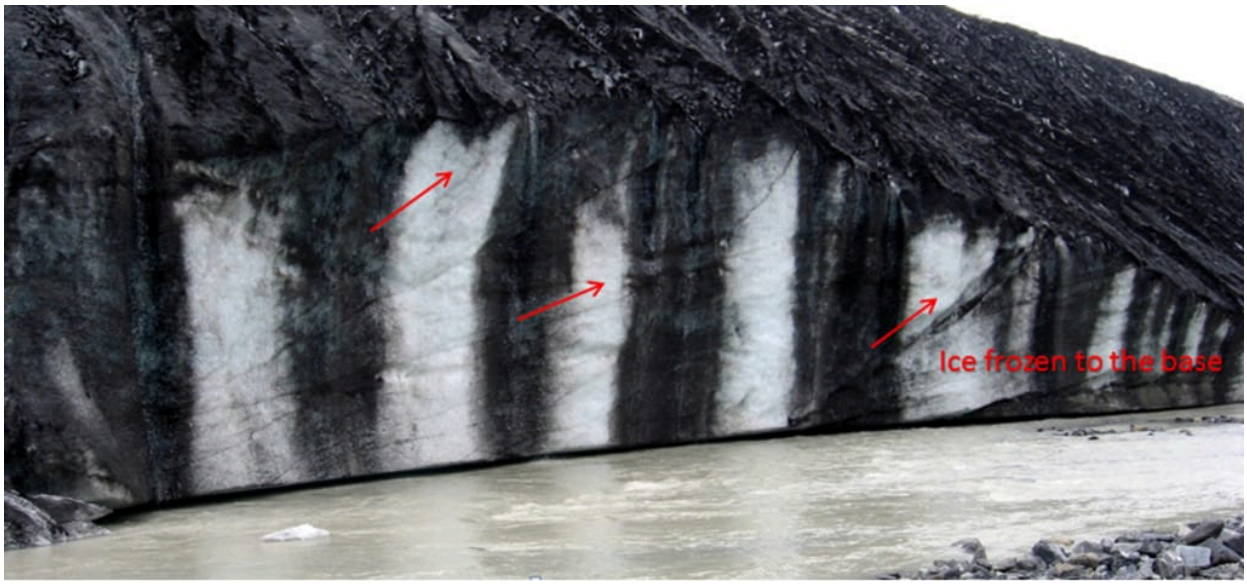


Figure 18.1.10 Thrust faults at the leading edge of the Athabasca Glacier, Alberta. The arrows show how the trailing ice has been thrust over the leading ice. (The dark vertical stripes are mud from sediments that have been washed off of the lateral moraine lying on the surface of the ice.)

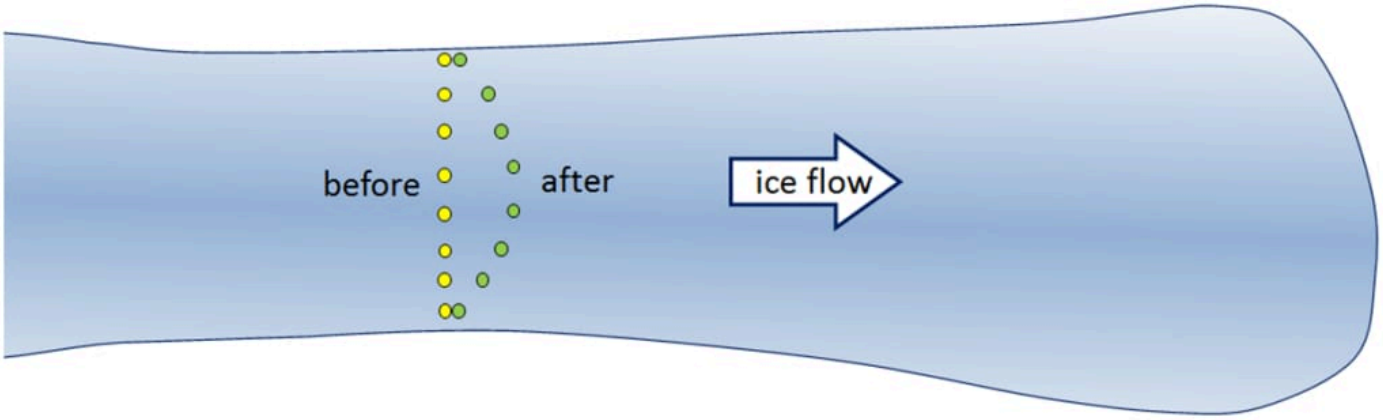


Figure 18.1.11 Markers on an alpine glacier move forward over a period of time with the ones in the middle moving faster than the ones on the edge.

Just as the base of a glacier moves more slowly than the surface, the edges, which are more affected by friction along the sides, move more slowly than the middle. If we were to place a series of markers across an alpine glacier and come back a year later, we would see that the ones in the middle had moved farther forward than the ones near the edges (Figure 18.1.11).

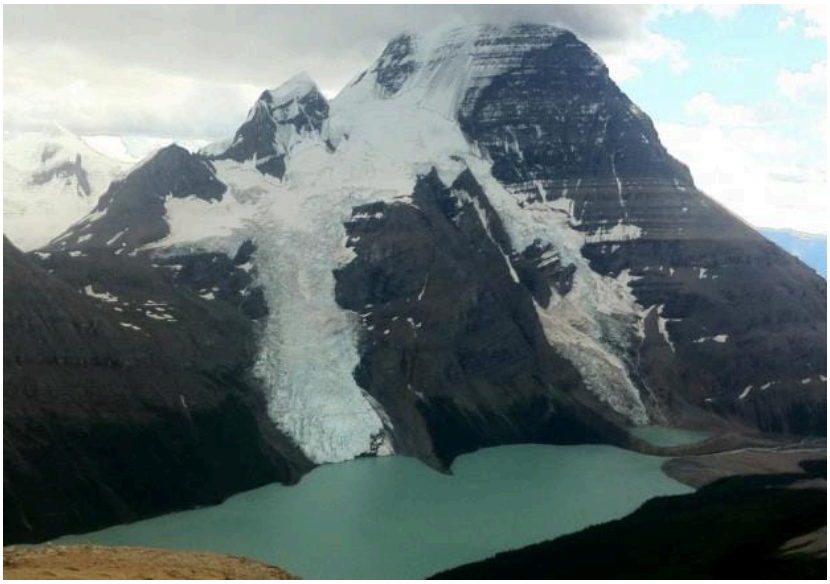


Figure 18.1.12 Mt. Robson, the tallest peak in the Canadian Rockies, Berg Glacier (centre), and Berg Lake. Although there were no icebergs visible when this photo was taken, the Berg Glacier loses mass by shedding icebergs into Berg Lake.

Glacial ice always moves downhill, in response to gravity, but the front edge of a glacier is always either melting or **calving** into water (shedding icebergs). If the rate of forward motion of the glacier is faster than the rate of **ablation** (melting), the leading edge of the glacier advances (moves forward). If the rate of forward motion is about the same as the rate of ablation, the leading edge remains stationary, and if the rate of forward motion is slower than the rate of ablation, the leading edge retreats (moves backward).

Calving of icebergs is an important process for glaciers that terminate in lakes or the ocean. An example of such a glacier is the Berg Glacier on Mt. Robson (Figure 18.1.12), which sheds small icebergs into Berg Lake. The Berg Glacier also loses mass by melting, especially at lower elevations.

Exercise 16.2 Ice advance and retreat

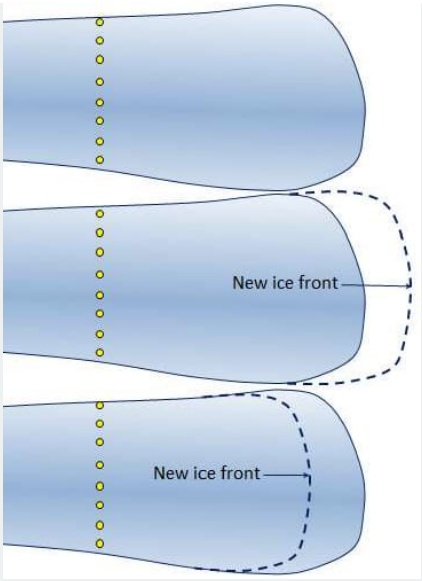


Figure 18.1.13

These diagrams represent a glacier with markers placed on its surface to determine the rate of ice motion over a one-year period. The ice is flowing from left to right.

1. In the middle diagram, the leading edge of the glacier has advanced. Draw in the current position of the markers.
2. In the lower diagram, the leading edge of the glacier has retreated. Draw in the current position of the markers.

See Appendix 3 for Exercise 16.2 answers.

Image Descriptions

Figure 18.1.2 **image description:** An east to west cross section of the ice on Antarctica is over 4000 km long and up to 4 km high in some places. An east to west cross section of the ice on Greenland is less than 1000 km long and up to 3 km high. [Return to Figure 18.1.2]

Figure 18.1.5 *image description: The comparative density of snowflakes, ice granules, firn, and ice.*

Shape	Amount of air	Density (grams per centimeters cubed)
Snowflake	90%	Less than 0.1
Ice Granules	50%	From around 0.3 to 0.5
Firn	30%	From around 0.5 to 0.7
Ice	20% (as bubbles)	Greater than 0.7

[Return to Figure 18.1.5]

Media Attributions

- Figures 16.2.1, 16.2.2, 16.2.3, 16.2.4, 16.2.5, 16.2.7: © Steven Earle. CC BY.
- Figure 18.1.6: “Overlord Glacier” by Isaac Earle. Adapted by Steven Earle. CC BY.
- Figure 18.1.8: “Crevasses at Overlord Glacier” by Isaac Earle. CC BY.
- Figures 16.2.9, 16.2.10, 16.2.11, 16.2.12, 16.2.13: © Steven Earle. CC BY.

This page titled 18.1: How Glaciers Work is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Steven Earle (BCCampus) via source content that was edited to the style and standards of the LibreTexts platform.

- 16.2: How Glaciers Work by Steven Earle is licensed CC BY 4.0. Original source: https://opentextbc.ca/physicalgeology2ed.

18.2: Glacial Erosion

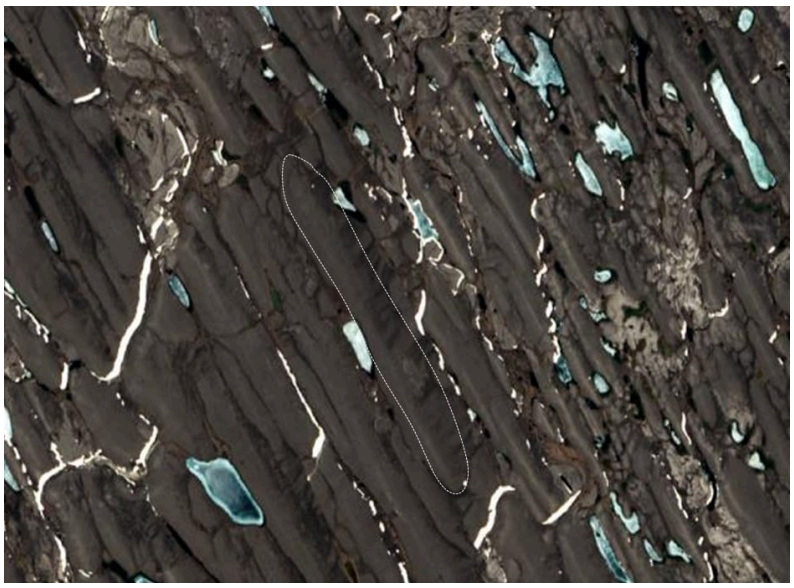


Figure 18.2.1 Drumlins — streamlined hills formed beneath a glacier, here made up of sediment — in the Amundsun Gulf region of Nunavut. The drumlins are tens of meters high, a few hundred meters across, and a few kilometers long. One of them is highlighted with a dashed white line.

Glaciers are effective agents of erosion, especially in situations where the ice is not frozen to its base and can therefore slide over the bedrock or other sediment. In fact the ice itself is not particularly effective at erosion because it is relatively soft (Mohs hardness 1.5 at 0°C); instead, it is the rock fragments embedded in the ice and pushed down onto the underlying surfaces that do most of the erosion. A useful analogy would be to compare the effect of a piece of paper being rubbed against a wooden surface, as opposed to a piece of sandpaper that has embedded angular fragments of garnet.

The results of glacial erosion are different in areas with continental glaciation versus alpine glaciation. Continental glaciation tends to produce relatively flat bedrock surfaces, especially where the rock beneath is uniform in strength. In areas where there are differences in the strength of rocks, a glacier obviously tends to erode the softer and weaker rock more effectively than the harder and stronger rock. Much of central and eastern Canada, which was completely covered by the huge Laurentide Ice Sheet at various times during the Pleistocene, has been eroded to a relatively flat surface. In many cases the existing relief is due the presence of glacial deposits—such as drumlins, eskers, and moraines (all discussed below)—rather than to differential erosion (Figure 18.2.1).

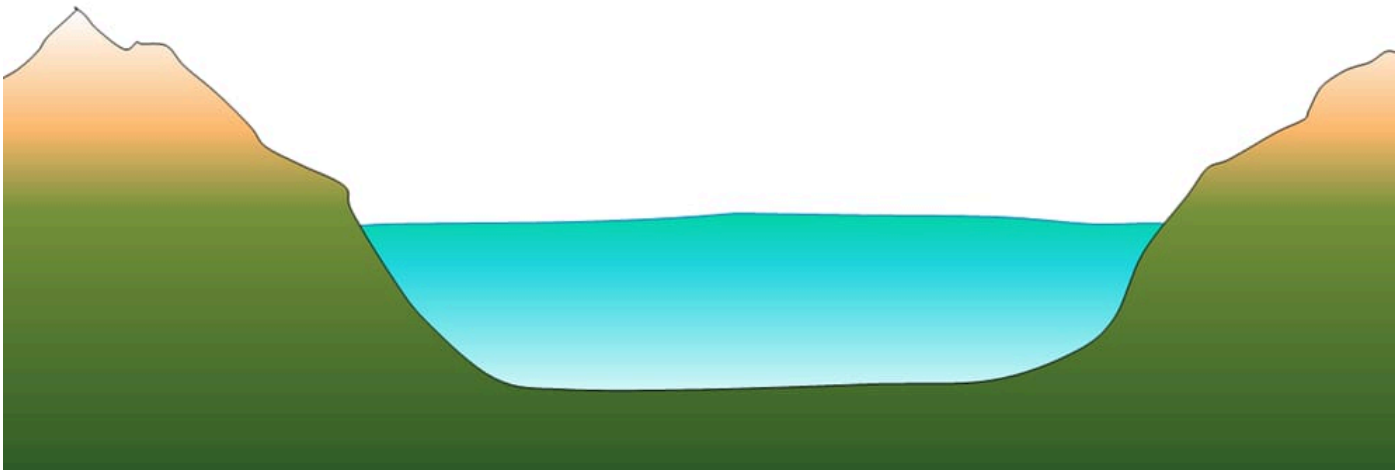


Figure 18.2.2 A depiction of a U-shaped valley occupied by a large glacier.

Alpine glaciers produce very different topography than continental glaciers, and much of the topographic variability of western Canada can be attributed to glacial erosion. In general, glaciers are much wider than rivers of similar length, and since they tend to erode more at their bases than their sides, they produce wide valleys with relatively flat bottoms and steep sides—known as **U-shaped valleys** (Figure 18.2.2). Howe Sound, north of Vancouver, was occupied by a large glacier that originated in the Squamish, Whistler, and Pemberton areas, and then joined the much larger glacier in the Strait of Georgia. Howe Sound and most of its tributary valleys have pronounced U-shaped profiles (Figure 18.2.3).

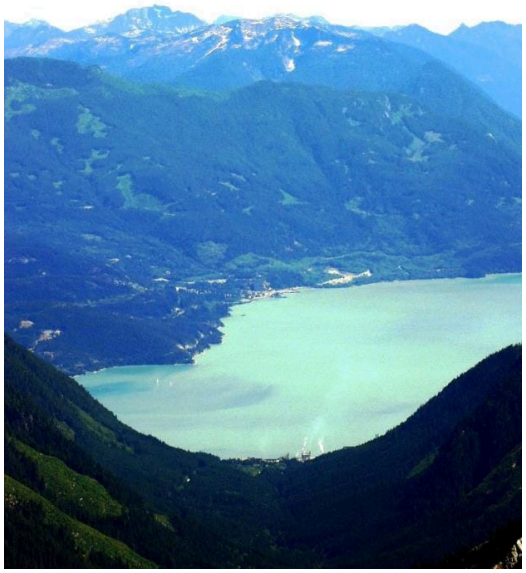


Figure 18.2.3 The view down the U-shaped valley of Mill Creek valley toward the U-shaped valley of Howe Sound (north of Vancouver BC), with the village of Britannia on the opposite side.

U-shaped valleys and their tributaries provide the basis for a wide range of alpine glacial topographic features, examples of which are visible on the International Space Station view of the Swiss Alps shown in Figure 18.2.4. This area was much more intensely glaciated during the past glacial maximum. At that time, the large U-shaped valley in the lower right was occupied by glacial ice, and all of the other glaciers shown here were longer and much thicker than they are now. But even at the peak of the Pleistocene Glaciation, some of the higher peaks and ridges would have been exposed and not directly affected by glacial erosion. A peak that extends above the surrounding glacier is called a **nunatuk**. In these areas, and in the areas above the glaciers today, most of the erosion is related to freeze-thaw effects.

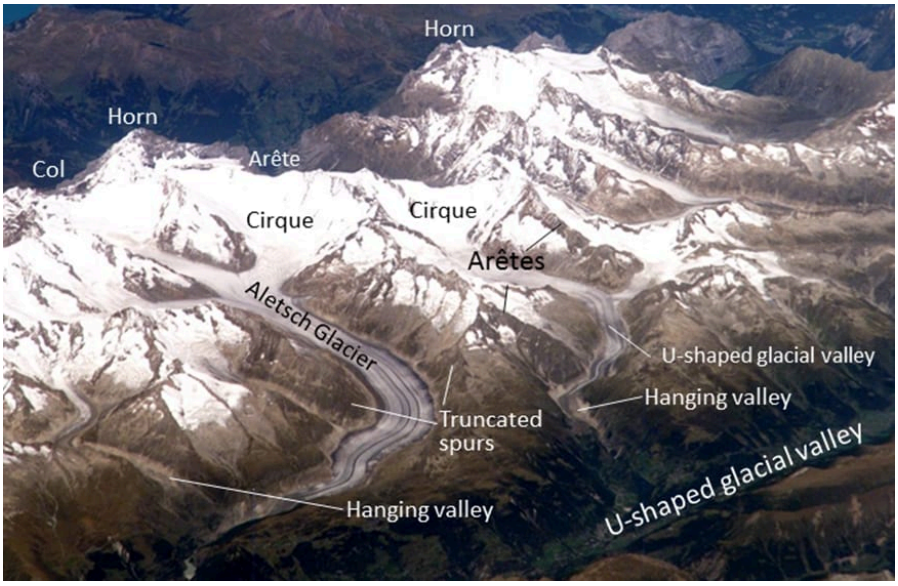


Figure 18.2.4 A view from the International Space Station of the Swiss Alps in the area of the Aletsch Glacier. The prominent peaks labelled "Horn" are the Eiger (left) and Wetterhorn (right). A variety of alpine glacial erosion features are labelled.

Figure 18.2.4 is a Space Station view of a glaciated terrain in the Swiss Alps. Some of the important features visible are **arêtes**: sharp ridges between U-shaped glacial valleys; **cols**: low points along arêtes that constitute passes between glacial valleys; **horns**: steep peaks that have been glacially and freeze-thaw eroded on three or more sides; **cirques**: bowl-shaped basins that form at the head of a glacial valley; **hanging valleys**: U-shaped valleys of tributary glaciers that hang above the main valley because the larger main-valley glacier eroded more deeply into the terrain; and **truncated spurs** (a.k.a. "spurs"): the ends of arêtes that have been eroded into steep triangle-shaped cliffs by the glacier in the corresponding main valley.

Some of these alpine-glaciation erosional features are also shown in Figure 18.2.5 in diagram form.

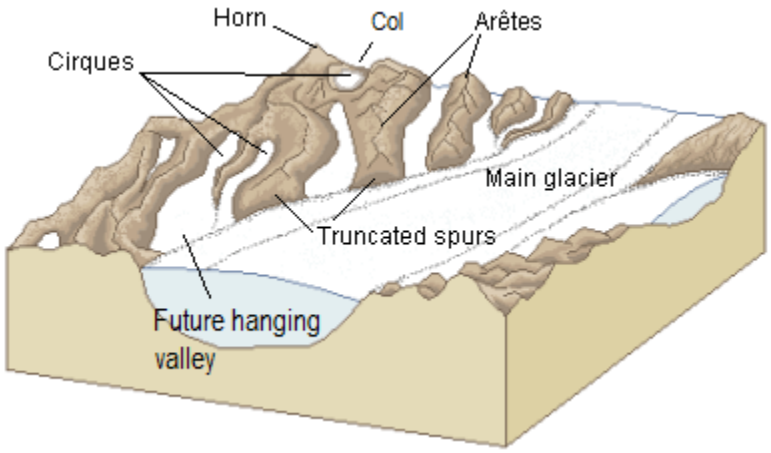


Figure 18.2.5 A diagram of some of the important alpine-glaciation erosion features.

Exercise 16.3 Identify alpine glacial erosion features

This is a photo of Mt. Assiniboine in the B.C. Rockies. What are the features at locations **a** through **e**? Look for one of each of the following: a horn, an arête, a truncated spur, a cirque, and a col. Try to identify some of the numerous other arêtes in this view, as well as another horn.

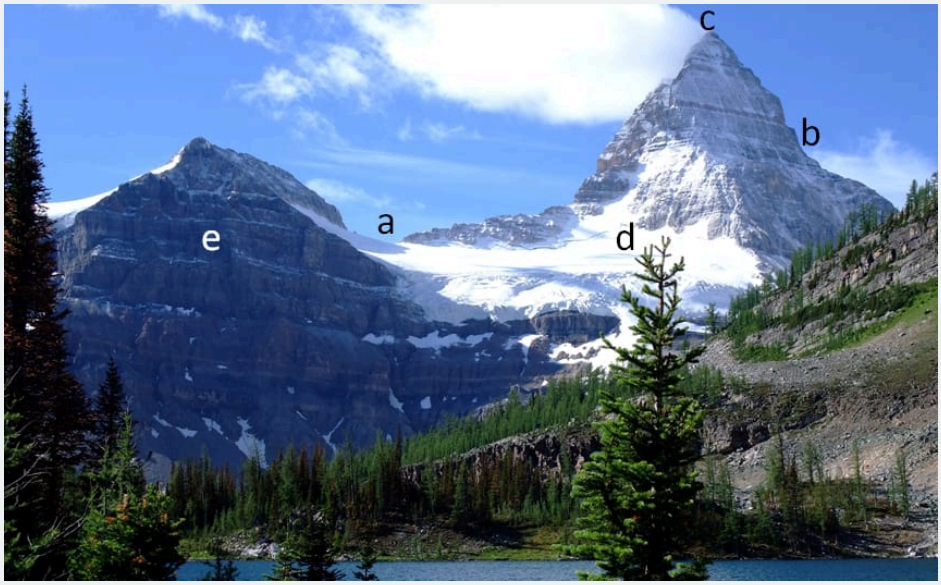


Figure 18.2.6

See Appendix 3 for Exercise 16.3 answers.

A number of other glacial erosion features exist at smaller scales. For example, a **drumlin** is an elongated feature that is streamlined at the down-ice end. The one shown in Figure 18.2.7 is larger than most, and is made up almost entirely of rock. Drumlins made up of glacial sediments are very common in some areas of continental glaciation (Figure 18.2.1).



Figure 18.2.7 Bowyer Island, a drumlin in Howe Sound, B.C. Ice flow was from right to left.

A **roche moutonée** is another type of elongated erosional feature that has a steep and sometimes jagged down-ice end (Figure 18.2.8, left). On a smaller scale still, **glacial grooves** (tens of centimeters to meters wide) and **glacial striation** (millimeters to centimeters wide) are created by fragments of rock embedded in the ice at the base of a glacier (Figure 18.2.8, left and right). Glacial striae are very common on rock surfaces eroded by both alpine and continental glaciers.



Figure 18.2.8 Left: Roches moutonées with glacial striae near Squamish, B.C. Right: Glacial striae at the same location near Squamish. Ice flow was from right to left in both cases. The brown rock in the right-hand photo is a mafic dyke about 40 cm wide intruding into granite.

Glacial lakes



Figure 18.2.9 Lower Thornton Lake, a tarn, in the Northern Cascades National Park, Washington.

Lakes are common features in glacial environments. A lake that is confined to a glacial cirque is known as a **tarn** (Figure 18.2.9). Tarns are common in areas of alpine glaciation because the ice that forms a cirque typically carves out a depression in bedrock that then fills with water. In some cases, a series of such basins will form, and the resulting lakes are called **rock basin lakes** or **paternoster lakes**.

A lake that occupies a glacial valley, but is not confined to a cirque, is known as a **finger lake**. In some cases, a finger lake is confined by a dam formed by an end moraine, in which case it may be called a **moraine lake** (Figure 18.2.10).



Figure 18.2.10 Peyto Lake in the Alberta Rockies, is both a finger lake and a moraine lake as it is dammed by an end moraine, on the right.

In areas of continental glaciation, the crust is depressed by the weight of glacial ice that is up to 4,000 m thick. Basins are formed along the edges of continental glaciers (except for those that cover entire continents like Antarctica and Greenland), and these basins fill with glacial meltwater. Many such lakes, some of them huge, existed at various times along the southern edge of the Laurentide Ice Sheet. One example is Glacial Lake Missoula, which formed within Idaho and Montana in northwestern United States. During the latter part of the last glaciation (30 ka to 15 ka), the ice holding back Lake Missoula retreated enough to allow some of the lake water to start flowing out, which escalated into a massive and rapid outflow (over days to weeks) during which much of the volume of the lake drained along the valley of the Columbia River to the Pacific Ocean. It is estimated that this type of flooding happened at least 25 times over that period, and in many cases, the rate of outflow was equivalent to the discharge

of all of Earth's current rivers combined. The record of these massive floods is preserved in the Channelled Scablands of Idaho, Washington, and Oregon (Figure 18.2.11).



Figure 18.2.11 Potholes Coulee near Wenatchee, Washington, one of many basins that received Lake Missoula floodwaters during the late Pleistocene. Here the water flowed from right to left, over the cliff and into this basin.

Another type of glacial lake is a **kettle lake**. These are discussed in section 16.4 in the context of glacial deposits.

Image Descriptions

Figure 18.2.6 image description:

1. A valley between two peaks.
2. A steep edge of a mountain.
3. A mountain peak.
4. An ice field in between two mountains.
5. A flat, rocky side of a mountain.

[Return to Figure 18.2.6]

Media Attributions

- Figure 18.2.1: © Steven Earle. CC BY.
- Figure 18.2.2: "Drumlin Field in Northern Canada" by NASA Earth Observatory. Public domain.
- Figure 18.2.3: "Woodf1a" © Keefer4. CC BY-SA.
- Figure 18.2.4: "Bernese Alps, Switzerland" by NASA Earth Observatory. Adapted by Steven Earle. Public domain.
- Figure 18.2.5: "Glacial landscape LMB" by Luis María Benítez. Public domain.
- Figure 18.2.6: "Mount Assiniboine Sunbursst Lake" © Kurt Stegmüller. Adapted by Steven Earle. CC BY.
- Figures 16.3.7, 16.3.8: © Steven Earle. CC BY.
- Figure 18.2.9: "Trappers Peak and lower Thorton Lake" © X-Weinzar. CC BY.
- Figure 18.2.10: "Peyto Lake Panorama 2006" © chensiyuan. CC BY-SA.
- Figure 18.2.11: © Steven Earle. CC BY.

This page titled 18.2: Glacial Erosion is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Steven Earle (BCCampus) via source content that was edited to the style and standards of the LibreTexts platform.

- 16.3: Glacial Erosion by Steven Earle is licensed CC BY 4.0. Original source: https://opentextbc.ca/physicalgeology2ed.

18.3: Glacial Deposition

Sediments transported and deposited during the Pleistocene glaciations are abundant throughout Canada and much of the northern USA. They are important sources of construction materials and are valuable as reservoirs for groundwater. Because they are almost all unconsolidated, they have significant implications for mass wasting.



Figure 18.3.1 Part of the Bering Glacier in southeast Alaska, the largest glacier in North America. It is about 14 km across in the centre of this view.

Figure 18.3.1 illustrates some of the ways that sediments are transported and deposited. The Bering Glacier is the largest in North America, and although most of it is in Alaska, it flows from an icefield that extends into southwestern Yukon. The surface of the ice is partially, or in some cases completely covered with rocky debris that has fallen from surrounding steep rock faces. There are muddy rivers issuing from the glacier in several locations, depositing sediment on land, into Vitus Lake, and directly into the ocean. There are dirty icebergs shedding their sediment into the lake. And, not visible in this view, there are sediments being moved along beneath the ice.

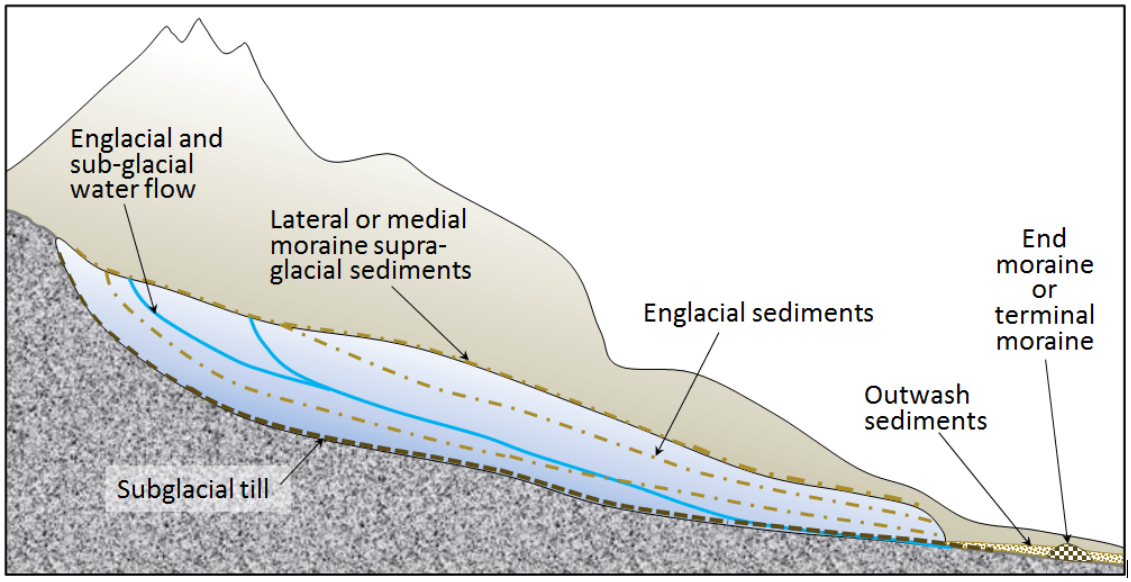


Figure 18.3.2 A depiction of the various types of sediments associated with glaciation. The glacier is shown in cross-section.

The formation and movement of sediments in glacial environments is shown diagrammatically in Figure 18.3.2. There are many types of glacial sediment generally classified by whether they are transported on, within, or beneath the glacial ice. The main types of sediment in a glacial environment are described below.

Supraglacial (on top of the ice) and **Supraglacial** (within the ice) sediments that slide off the melting front of a stationary glacier can form a ridge of unsorted sediments called an **englacial**. The end moraine that represents the farthest advance of the glacier is a **end moraine**. Supraglacial and englacial sediments can also be deposited when the ice melts. Sediments transported and deposited by glacial ice are known as **till**.

Subglacial sediment (e.g., **lodgement till**) is material that has been eroded from the underlying rock by the ice, and is moved by the ice. It has a wide range of grain sizes (in other words it is poorly sorted), including a relatively high proportion of silt and clay. The larger clasts (pebbles to boulders in size) tend to become partly rounded by abrasion. Lodgement till forms as a sheet of well-compacted sediment beneath a glacier, and ranges from several centimeters to many meters in thickness. Lodgement till is normally unbedded. An example is shown in Figure 18.3.3a.

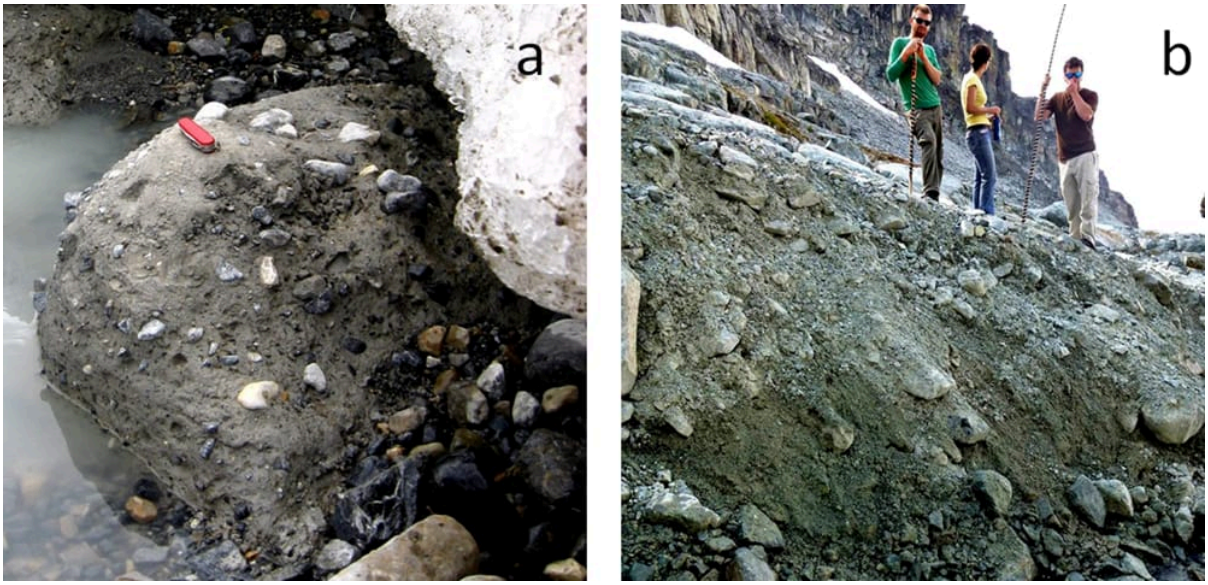


Figure 18.3.3 Examples of glacial till: a: lodgement till from the front of the Athabasca Glacier, Alberta; b: ablation till at the Horstman Glacier, Blackcomb Mountain, B.C.

Supraglacial sediments are primarily derived from freeze-thaw eroded material that has fallen onto the ice from rocky slopes above. These sediments form **lateral moraines** (Figure 18.3.1) and, where two glaciers meet, **medial moraines**. (Medial moraines are visible on the Aletsch Glacier in Figure 18.3.4.) Most of this material is deposited on the ground when the ice melts, and is therefore called **ablation till**, a mixture of fine and coarse angular rock fragments, with much less sand, silt, and clay than lodgement till. An example is shown in Figure 18.3.3b. When supraglacial sediments become incorporated into the body of the glacier, they are known as englacial sediments (Figure 18.3.2).

Massive amounts of water flow on the surface, within, and at the base of a glacier, even in cold areas and even when the glacier is advancing. Depending on its velocity, this water is able to move sediments of various sizes and most of that material is washed out of the lower end of the glacier and deposited as outwash sediments. These sediments accumulate in a wide range of environments in the **proglacial** region (the area in front of a glacier), most in fluvial environments, but some in lakes and the ocean. **Glaciofluvial sediments** are similar to sediments deposited in normal fluvial environments, and are dominated by silt, sand, and gravel. The grains tend to be moderately well rounded, and the sediments have similar sedimentary structures (e.g., bedding, cross-bedding, clast imbrication) to those formed by non-glacial streams (Figure 18.3.4a and 16.4.4b).

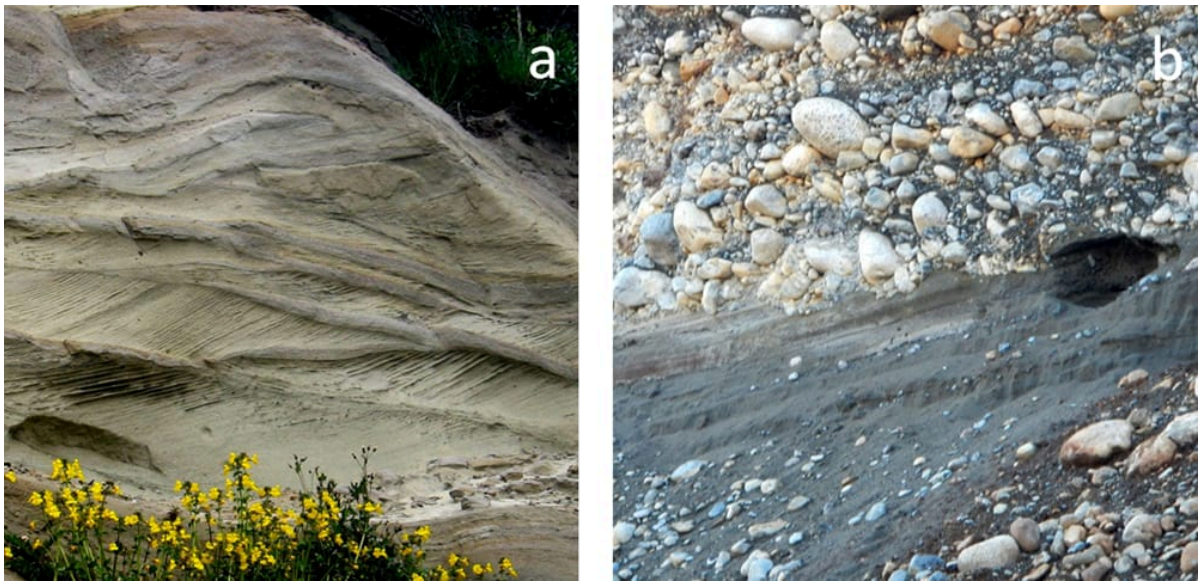


Figure 18.3.4 Examples of glaciofluvial sediments: a: glaciofluvial sand of the Quadra Sand Formation at Comox, B.C.; b: glaciofluvial gravel and sand, Nanaimo, B.C.

A large proglacial plain of sediment is called a **sandur** (a.k.a. an **outwash plain**)**,** and within that area, glaciofluvial deposits can be tens of meters thick (Figure 18.3.5). In situations where a glacier is receding, a block of ice might become separated from the main ice sheet and then get buried in glaciofluvial sediments. When the ice block eventually melts, a depression forms, known as a **kettle**, and if this fills with water, it is known as a **kettle lake** (Figure 18.3.6).



Figure 18.3.5 Part of a sandur in front of the Vatnajokull Glacier in Iceland,. The Sandur extends for many tens of km. Here is has been partially eroded by a stream (not visible).



Figure 18.3.6 A kettle lake amid vineyards and orchards in the Osoyoos area of B.C.



Figure 18.3.7 Part of an esker that formed beneath the Laurentide Ice Sheet in northern Canada.

A subglacial stream will create its own channel within the ice, and sediments that are being transported and deposited by the stream will build up within that channel. When the ice recedes, the sediment will remain to form a long sinuous ridge known as an **esker**. Eskers are most common in areas of continental glaciation. They can be several meters high, tens of meters wide, and tens of kilometers long (Figure 18.3.7).

Outwash streams commonly flow into proglacial lakes where **glaciolacustrine sediments** are deposited. These are dominated by silt- and clay-sized particles and are typically laminated on the millimetre scale. In some cases, **varves** develop; varves are series of beds with distinctive summer and winter layers: relatively coarse in the summer when melt discharge is high, and finer in the winter, when discharge is very low. Icebergs are common on proglacial lakes, and most of them contain englacial sediments of various sizes. As the bergs melt, the released clasts sink to the bottom and are incorporated into the glaciolacustrine layers as **drop stones** (Figure 18.3.8a).

The processes that occur in proglacial lakes can also take place where a glacier terminates at the ocean. The sediments deposited there are called **glaciomarine sediments** (Figure 18.3.8b).

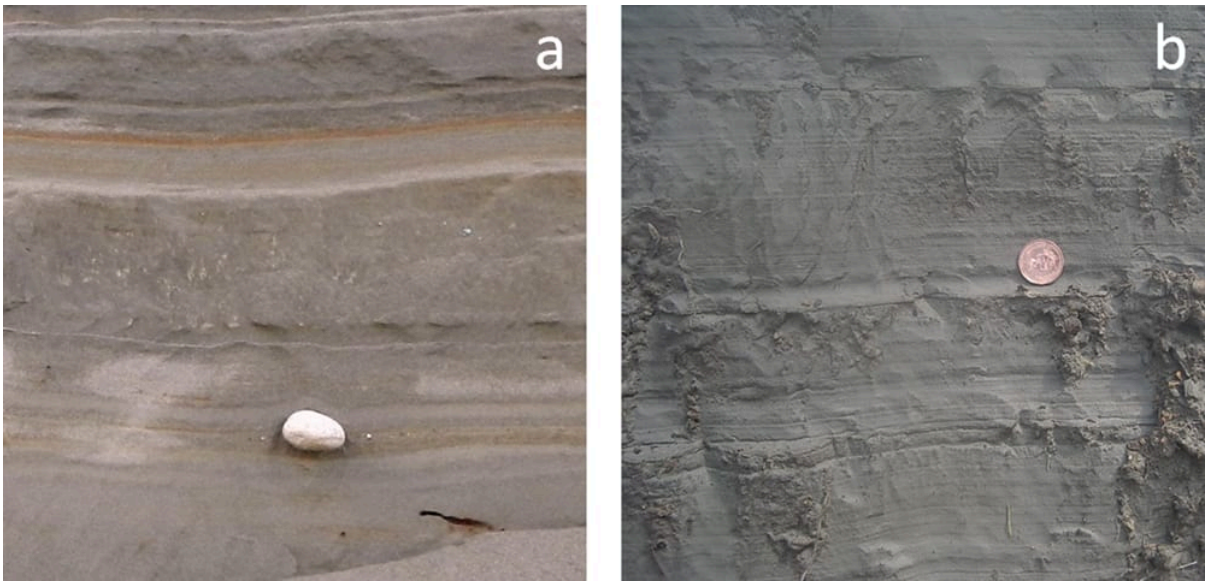


Figure 18.3.8 Examples of glacial sediments formed in quiet water: a: glaciolacustrine sediment with a drop stone, Nanaimo, B.C.; and b: a laminated glaciomarine sediment, Englishman River, B.C. Although not visible in this photo, the glaciomarine sediment has marine shell fossils.

Exercise 16.4 Identify glacial depositional environments



Figure 18.3.9

This photo shows the Bering Glacier in Alaska (same as Figure 18.3.1).

Glacial sediments of many different types are being deposited throughout this area. Identify where you would expect to find the following types of deposits:

1. Glaciofluvial sand
2. Lodgement till
3. Glaciolacustrine clay with drop stones
4. Ablation till
5. Glaciomarine silt and clay

See Appendix 3 for Exercise 16.4 answers.

Media Attributions

- Figure 18.3.1: “Bering Glacier” by NASA Earth Observatory. Public domain.
- Figures 16.4.2, 16.4.3, 16.4.4, 16.4.5, 16.4.6, 16.4.8: © Steven Earle. CC BY.
- Figure 18.3.7: “Esker on a morainal plain” © Canadian Soil Information Service (CANSIS). Approved for non-commercial reproduction.
- Figure 18.3.9: “Bering Glacier” by NASA Earth Observatory. Public domain.

This page titled 18.3: Glacial Deposition is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Steven Earle (BCCampus) via source content that was edited to the style and standards of the LibreTexts platform.

- 16.4: Glacial Deposition by Steven Earle is licensed CC BY 4.0. Original source: https://opentextbc.ca/physicalgeology2ed.

18.4: Earth's First 2 Billion Years

If you were to get into a time machine and visit Earth shortly after it formed (around 4.5 billion years ago), you would probably regret it. Large patches of Earth's surface would still be molten, which would make landing your time machine very dangerous indeed. If you happened to have one of the newer time-machine models with hovering capabilities and heat shields, you would still face the inconvenience of having nothing to breathe but a tenuous wisp of hydrogen and helium gas, and depending on how much volcanic activity was going on, volcanic gases such as water vapor and carbon dioxide. Some ammonia and methane might be thrown in just to make it interesting, but there would be no oxygen. Assuming you had the foresight to purchase the artificial atmosphere upgrade for your time machine, it would all be for naught if you materialized just in time to see an asteroid, or worse yet another planet, bearing down on your position. The moral of the story is that early Earth was a nasty place, and a time machine purchase is not something to take lightly.

Why was early Earth so nasty?

The early Earth was hot

Chapter 9 explains that Earth's heat comes from the decay of radioactive elements within Earth, as well as from processes associated with Earth's formation. Let's look more closely at how those formation processes heated up Earth:

- Heat came from the thermal energy already contained within the objects that accreted to form the Earth.
- Heat came from collisions. When objects hit Earth, some of the energy from their motion went into deforming Earth, and some of it was transformed into heat. Clap your hands vigorously to experience this on a much smaller (and safer!) scale.
- As Earth became larger, its gravitational force became stronger. This increased Earth's ability to draw objects to it, but it also caused the material making Earth to be compressed, rather like Earth giving itself a giant gravitational hug. Compression causes materials to heat up.

Heating had a very important consequence for Earth's structure. As Earth grew, it collected a mixture of silicate mineral grains as well as iron and nickel. These materials were scattered throughout Earth. That changed when Earth began to heat up: it got so hot that both the silicate minerals and the metals melted. The metal melt was much denser than the silicate mineral melt, so the metal melt sank to Earth's centre to become its core, and the silicate melt rose upward to become Earth's crust and mantle. In other words, Earth unmixed itself. The separation of silicate minerals and metals into a rocky outer layer and a metallic core, respectively, is called **differentiation**. The movement of silicate and metal melts within Earth caused it to heat up even more.

Earth's high temperature early in its history also means that early tectonic processes were accelerated compared to today, and Earth's surface was more geologically active.

Earth was heavily bombarded by objects from space

Although Earth had swept up a substantial amount of the material in its orbit as it was accreting, unrest within the solar system caused by changes in the orbits of Saturn and Jupiter was still sending many large objects on cataclysmic collision courses with Earth. The energy from these collisions repeatedly melted and even vaporized minerals in the crust, and blasted gases out of Earth's atmosphere. Very old scars from these collisions are still detectable, although we have to look carefully to see them. For example, the oldest impact site discovered is the 3 billion year old Maniitsoq "crater" in west Greenland, although there is no crater to see. What is visible are rocks that were 20 km to 25 km below Earth's surface at the time of the impact, but which nevertheless display evidence of deformation that could only be produced by intense, sudden shock.

The evidence of the very worst collision that Earth experienced is not subtle at all. In fact, you have probably looked directly at it hundreds of times already, perhaps without realizing what it is. That collision was with a planet named Theia, which was approximately the size of Mars (Figure 18.4.1). Not long after Earth formed, Theia struck Earth. When Theia slammed into Earth, Theia's metal core merged with Earth's core, and debris from the outer silicate layers was cast into space, forming a ring of rubble around Earth. The material within the ring coalesced into a new body in orbit around Earth, giving us our moon. Remarkably, the debris may have coalesced in 10 years or fewer! This scenario for the formation of the moon is called the **giant impact hypothesis.**

Earth's atmosphere as we know it took a long time to develop

Earth's first experiment with having an atmosphere didn't go well. It started out with a thin veil of hydrogen and helium gases that came with the material it accreted. However, hydrogen and helium are very light gases, and they bled off into space.

Earth's second experiment with having an atmosphere went much better. Volcanic eruptions built up the atmosphere by releasing gases. The most common volcanic gases are water vapor and carbon dioxide (CO_2), but volcanoes release a wide variety of gases. Other important contributions include sulfur dioxide (SO_2), carbon monoxide (CO), hydrogen sulfide (H_2S), hydrogen gas, and methane (CH_4). Meteorites and comets also brought substantial amounts of water and nitrogen to Earth. It is not clear what the exact composition of the atmosphere was after Earth's second experiment, but carbon dioxide, water vapor, and nitrogen were likely the three most abundant components.

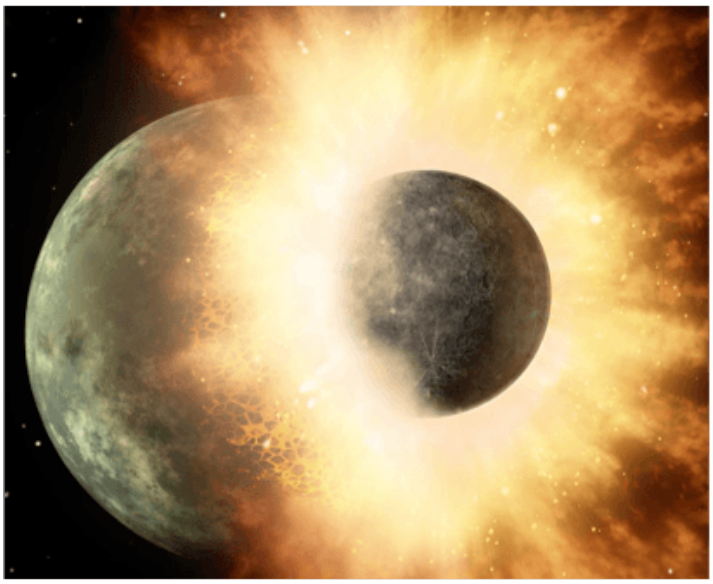


Figure 18.4.1 Artist's impression of a collision between planets. A similar collision between Earth and the planet Theia might have given us our Moon. Fortunately for us, the collision that gave us the moon was a glancing blow rather than the direct hit shown here. Earth might not have survived a direct hit.

One thing we can say for sure about Earth's second experiment is that there was effectively no free oxygen (O_2, the form of oxygen that we breathe) in the atmosphere. We know this in part because prior to 2 billion years ago, there were no sedimentary beds stained red from oxidized iron minerals. Iron minerals were present, but not in oxidized form. At that time, O_2 was produced in the atmosphere when the Sun's ultraviolet rays split water molecules apart; however, chemical reactions removed the oxygen as quickly as it was produced.

It wasn't until well into Earth's third experiment—life—that the atmosphere began to become oxygenated. Photosynthetic organisms used the abundant CO_2 in the atmosphere to manufacture their food, and released O_2 as a by-product. At first all of the oxygen was consumed by chemical reactions, but eventually the organisms released so much O_2 that it overwhelmed the chemical reactions and oxygen began to accumulate in the atmosphere, although present levels of 21% oxygen didn't occur until about 350 Ma. Today the part of our atmosphere that isn't oxygen consists largely of nitrogen (78%).

The oxygen-rich atmosphere on our planet is life's signature. If geologic process were the only processes controlling our atmosphere, it would consist mostly of carbon dioxide, like the atmosphere of Venus. It is an interesting notion (or a disconcerting one, depending on your point of view) that for the last 2 billion years the light reflected from our planet has been beaming a bar code out to the universe, similar to the ones in Figure 18.4.4, except ours says "oxygen." For 2 billion years, our planet has been sending out a signal that could cause an observer from another world to say, "That's odd... I wonder what's going on over there."

Media Attributions

- Figure 18.4.1: "Planetary Smash-Up" by NASA/JPL-Caltech. Public domain.

This page titled 18.4: Earth's First 2 Billion Years is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Steven Earle (BCCampus) via source content that was edited to the style and standards of the LibreTexts platform.

- 22.4: Earth's First 2 Billion Years by Steven Earle is licensed CC BY 4.0. Original source: https://opentextbc.ca/physicalgeology2ed.

18.5: Are There Other Earths?

If by that you mean, are there other planets where we could walk out of a spaceship with no equipment other than a picnic basket, and enjoy a pleasant afternoon on a grassy slope near a stream, then that remains to be seen. On the other hand, if you are asking if other planets exist that are rocky worlds approximately Earth's size, and orbiting within their star's **habitable zone** (the zone in which liquid water, and potentially life, can exist), then many planet hunters are cautiously optimistic that we have found at least 12 such worlds so far.

As of July 2015, NASA's Kepler mission has detected a total of 4,696 possible exoplanets. The Kepler spacecraft has an instrument to measure the brightness of stars, and looks for tiny variations in brightness that could be caused by a planet passing between the star it orbits and the instrument observing the star. Potential candidates are then examined in more detail to see whether they are in fact planets or not. So far 1,030 of those candidates have been confirmed as planets.[1] Of those, 12 satisfy the criteria of being one to two times the size of Earth, and orbiting their star within the habitable zone.[2]

The uncertainty about the 12 possible Earth-like worlds is related to their composition. We don't yet know their composition; however, it is tempting to conclude that they are rocky because they are similar in size to Earth. Remember the rules of the accretion game: you can only begin to collect gas once you are a certain size, and how much matter you collect depends on how far away from the Sun you are. Given how large our gas giant and ice giant planets are compared to Earth, and how far away they are from the Sun, we would expect that a planet similar in size to Earth, and a similar distance from its star, should be rocky.

It isn't quite as simple as that, however. We are finding that the rules to the accretion game can result in planetary systems very different from our own, leading some people to wonder whether those planetary systems are strange, or ours is, and if ours is strange, how strange is it?

Consider that in the Kepler mission's observations thus far, it is very common to find planetary systems with planets larger than Earth orbiting closer to their star than Mercury does to the Sun. It is rare for planetary systems to have planets as large as Jupiter, and where large planets do exist, they are much closer to their star than Jupiter is to the Sun. To summarize, we need to be cautious about drawing conclusions from our own solar system, just in case we are basing those conclusions on something truly unusual.

On the other hand, the seemingly unique features of our solar system would make planetary systems like ours difficult to spot. Small planets are harder to detect because they block less of a star's light than larger planets. Larger planets farther from a star are difficult to spot because they don't go past the star as frequently. For example, Jupiter goes around the Sun once every 12 years, which means that if someone were observing our solar system, they might have to watch for 12 years to see Jupiter go past the Sun once. For Saturn, they might have to watch for 30 years.

So let's say the habitable-zone exoplanets are terrestrial. Does that mean we could live there?

The operational definition of "other Earths," which involves a terrestrial composition, a size constraint of one to two times that of Earth, and location within a star's habitable zone, does not preclude worlds incapable of supporting life as we know it. By those criteria, Venus is an "other Earth," albeit right on the edge of the habitable zone for our Sun. Venus is much too hot for us, with a constant surface temperature of 465°C (lead melts at 327°C). Its atmosphere is almost entirely carbon dioxide, and the atmospheric pressure at its surface is 92 times higher than on Earth. Any liquid water on its surface boiled off long ago. Yet the characteristics that make Venus a terrible picnic destination aren't entirely things we could predict from its distance from the sun. They depend in part on the geochemical evolution of Venus, and at one time Venus might have been a lot more like a youthful Earth. These are the kinds of things we won't know about until we can look carefully at the atmospheres and compositions of habitable-zone exoplanets.

Exercise 22.2 How well do we know the size of exoplanets?

One of the techniques for finding exoplanets is to measure changes in the brightness of a host star as the planet crosses in front of it and blocks some of its light. This diagram shows how the brightness changes over time. The dip in brightness reflects a planet crossing between the star and the instrument observing the star.

Often the planet itself is too small to see directly. If all we know is how the planet affects the brightness of the star, and we can't even see the planet, then how do we know how big the planet is? The answer is that the two are related. We can write an equation for this relationship using the radius of the planet and the radius of the star.

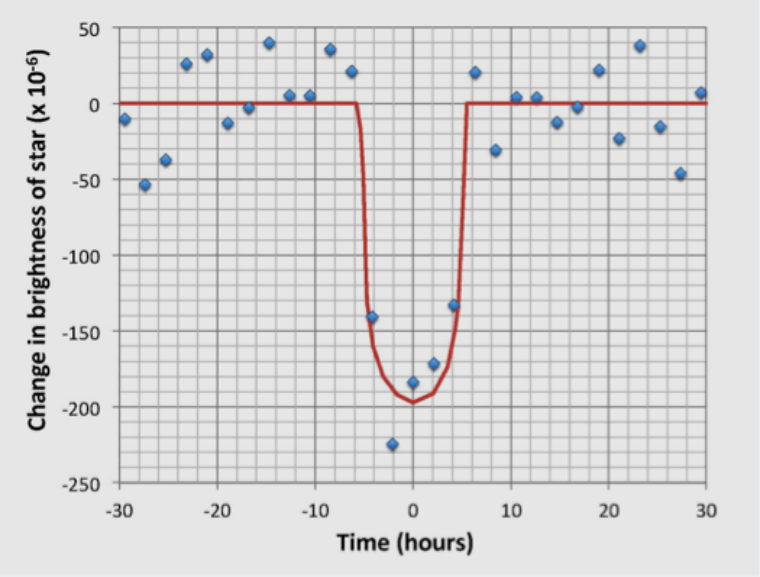


Figure 18.5.1 Plot showing how the star Kepler-452 dims as the planet Kepler-452b moves in front of it.

$$decrease\ in\ brightness = \frac{planet\ radius^2}{star\ radius^2}$$

Equation 1: Calculate decrease in brightness

Let's try this out for the Earth-like exoplanet called Kepler-452b. The first thing we need to know is the size of the host star Kepler-452. We can get that information by comparing its surface temperature and brightness to that of the sun. Start by calculating the ratios of the sun's temperature to the star's temperature, and the star's luminosity to the sun's luminosity using the data in Table 22.5. Record your answers in the table. Then find the star's radius using the following equation, and record your result:

$$star\ radius = sun\ radius \times \left(\frac{sun\ temperature}{star\ temperature}\right)^2 \times \sqrt{\frac{star\ luminosity}{sun\ luminosity}}$$

Equation 2: Calculate star radius.

Table 22.5 Calculating the radius of star Kepler-452

Description	Sun	Kepler-452	Ratio
Temperature (degrees Kelvin)	5,778	5,757	
Luminosity (× 1026 watts)	3.846	4.615	
Radius (km)	696,300		

The second thing we need to know is how the brightness of Kepler-452 changes as planet Kepler-452b moves in front of it. Use the plot shown in this exercise box to find this information. Find the value on the y-axis where the red curve shows the most dimming from the planet and record your result in Table 22.6.

Table 22.6 Calculating the radius of planet Kepler-452b

Decrease in brightness*	Earth radius in km	Kepler-452b radius in km	Kepler-452b radius/Earth radius
x 10^{-6}	6,378		

* Because we know this is a decrease, you don't need to keep the negative sign.

Use the following equation to find the radius of Kepler-452b:

$$planet\ radius = \sqrt{star\ radius^2 \times decrease\ in\ brightness}$$

Equation 3: Calculate planet radius.

To put the size of Kepler-452b in perspective, divide its radius by that of Earth and record your answer.

See Appendix 3 for Exercise 22.2 answers.

Image Descriptions

Equation 2 image description: Star radius equals sun radius times begin fraction sun temperature over star temperature end fraction squared times begin square root begin fraction star luminosity over sun luminosity end fraction end square root. [Return to Equation 2]

Media Attributions

- Figure 18.5.1: © Karla Panchuk. CC BY. Based on data from Jenkins, J. et al, 2015, Discovery and validation of Kepler-452b: a 1.6REarth super Earth exoplanet in the habitable zone of a G2 star, Astronomical Journal, V 150, DOI 10.1088/0004-6256/150/2/56.

---

1. You can access a catalogue of confirmed exoplanets found by NASA and other planet-hunting organizations at http://exoplanet.eu/catalog/↵
2. Read more about habitable-zone planets discovered so far at http://www.nasa.gov/jpl/finding-another-earth↵

This page titled 18.5: Are There Other Earths? is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Steven Earle (BCCampus) via source content that was edited to the style and standards of the LibreTexts platform.

- 22.5: Are There Other Earths? by Steven Earle is licensed CC BY 4.0. Original source: https://opentextbc.ca/physicalgeology2ed.

INDEX

No index terms found.

This page titled Index is shared under a license and was authored, remixed, and/or curated by LibreTexts.

GLOSSARY

S

Sample Word 1 | Sample Definition 1

This page titled Glossary is shared under a license and was authored, remixed, and/or curated by LibreTexts.

DETAILED LICENSING

Overview

Title: Introduction to Earth and Space Sciences

Webpages: 168

Applicable Restrictions: Noncommercial

All licenses found:

- CC BY 4.0: 50% (84 pages)
- CC BY-NC-SA 4.0: 25% (42 pages)
- CC BY-SA 4.0: 20.8% (35 pages)
- Public Domain: 1.8% (3 pages)
- Unknown License: 1.2% (2 pages)
- Undeclared: 1.2% (2 pages)
- Unknown License: 0% (0 page)

By Page

- Introduction to Earth and Space Sciences - *CC BY-NC-SA 4.0*
  - Front Matter - *CC BY-NC-SA 4.0*
    - TitlePage - *CC BY-NC-SA 4.0*
    - InfoPage - *Public Domain*
    - Table of Contents - *Public Domain*
    - Licensing - *Public Domain*
  - 1: Studying the Universe - *CC BY-NC-SA 4.0*
    - 1.1: The Nature of Astronomy - *CC BY 4.0*
    - 1.2: The Nature of Science - *CC BY 4.0*
    - 1.3: Consequences of Light Travel Time - *CC BY 4.0*
    - 1.4: A Tour of the Universe - *CC BY 4.0*
  - 2: Gravity and Planetary Motion - *CC BY 4.0*
    - 2.1: The Laws of Planetary Motion - *CC BY 4.0*
    - 2.2: Newton's Great Synthesis - *CC BY 4.0*
    - 2.3: Newton's Universal Law of Gravitation - *CC BY 4.0*
  - 3: Our Solar System - *CC BY-NC-SA 4.0*
    - 3.1: Overview of Our Planetary System - *CC BY 4.0*
    - 3.2: Composition and Structure of Planets - *CC BY 4.0*
    - 3.3: Formation of Planetary Systems - *CC BY-NC-SA 4.0*
      - 3.3.1: Origin of the Solar System - *CC BY 4.0*
      - 3.3.2: Formation of the Solar System - *CC BY 4.0*
      - 3.3.3: Comparison with Other Planetary Systems - *CC BY 4.0*
  - 4: Light and Matter - *CC BY-NC-SA 4.0*
    - 4.1: The Behavior of Light - *CC BY 4.0*
    - 4.2: The Electromagnetic Spectrum - *CC BY 4.0*
    - 4.3: Spectroscopy in Astronomy - *CC BY 4.0*
    - 4.4: The Structure of the Atom - *CC BY 4.0*
    - 4.5: Formation of Spectral Lines - *CC BY 4.0*
    - 4.6: The Doppler Effect - *CC BY 4.0*
  - 5: Stars - *CC BY-NC-SA 4.0*
    - 5.1: The Structure and Composition of the Sun - *CC BY 4.0*
    - 5.2: Understanding Stars - *CC BY-NC-SA 4.0*
      - 5.2.1: The Brightness of Stars - *CC BY 4.0*
      - 5.2.2: Colors of Stars - *CC BY 4.0*
      - 5.2.3: The Spectra of Stars (and Brown Dwarfs) - *CC BY 4.0*
      - 5.2.4: Using Spectra to Measure Stellar Radius, Composition, and Motion - *CC BY 4.0*
      - 5.2.5: The H-R Diagram - *CC BY 4.0*
    - 5.3: Stellar Evolution - *CC BY-NC-SA 4.0*
      - 5.3.1: Star Formation - *CC BY 4.0*
      - 5.3.2: The H-R Diagram and the Study of Stellar Evolution - *CC BY 4.0*
      - 5.3.3: Evolution from the Main Sequence to Red Giants - *CC BY 4.0*
      - 5.3.4: The Death of Low-Mass Stars - *CC BY 4.0*
      - 5.3.5: Evolution of Massive Stars-An Explosive Finish - *CC BY 4.0*
  - 6: Galaxies and the Universe - *CC BY-NC-SA 4.0*
    - 6.1: The Architecture of the Galaxy - *CC BY 4.0*
    - 6.2: Types of Galaxies - *CC BY 4.0*
    - 6.3: The Expanding Universe - *CC BY 4.0*
    - 6.4: The Beginning of the Universe - *CC BY 4.0*
    - 6.5: The Cosmic Microwave Background - *CC BY*

4.0
- 7: Earth's Atmosphere - *CC BY-NC-SA 4.0*
  - 7.1: Overview of Earth's Atmosphere - *CC BY-SA 4.0*
  - 7.2: Layers of the Atmosphere - *CC BY-SA 4.0*
  - 7.3: Weather and Climate: What's the Difference? - *CC BY-SA 4.0*
  - 7.4: Radiation - *CC BY-SA 4.0*
  - 7.5: Earth-Sun Energy Balance - *CC BY-NC-SA 4.0*
    - 7.5.1: Albedo - *CC BY-SA 4.0*
    - 7.5.2: Surface Energy Balance - *CC BY-SA 4.0*
    - 7.5.3: Radiation Interaction with the Atmosphere - *CC BY-SA 4.0*
  - 7.6: Air Pollution - *CC BY-NC-SA 4.0*
- 8: Energy & Humidity - *CC BY-NC-SA 4.0*
  - 8.1: Energy Transfer - *CC BY-SA 4.0*
  - 8.2: Latent Heat - *CC BY-SA 4.0*
  - 8.3: Saturation - *CC BY-SA 4.0*
  - 8.4: Humidity - *CC BY-SA 4.0*
  - 8.5: Why Do We Care So Much About Moisture? - *CC BY-SA 4.0*
  - 8.6: Atmospheric Stability and Lapse Rates - *CC BY-SA 4.0*
- 9: Clouds and Precipitation - *CC BY-NC-SA 4.0*
  - 9.1: Cloud Condensation Nuclei - *CC BY-SA 4.0*
  - 9.2: Cloud Naming Conventions - *CC BY-SA 4.0*
  - 9.3: Cloud Droplets - *CC BY-SA 4.0*
  - 9.4: Precipitation Types - *CC BY-SA 4.0*
- 10: Weather Observation and Maps - *CC BY-NC-SA 4.0*
  - 10.1: Meteorological Reports and Observations - *CC BY-SA 4.0*
  - 10.2: Synoptic Weather Maps - *CC BY-SA 4.0*
  - 10.3: Map Analysis - *CC BY-SA 4.0*
- 11: Atmospheric Circulation & Air Masses - *CC BY-NC-SA 4.0*
  - 11.1: Atmospheric Circulation - *CC BY-NC-SA 4.0*
    - 11.1.1: Main Forces - *CC BY-SA 4.0*
    - 11.1.2: Differential Heating - *CC BY-SA 4.0*
    - 11.1.3: Single-Cell Model - *CC BY-SA 4.0*
    - 11.1.4: Three-Cell Model - *CC BY-SA 4.0*
    - 11.1.5: Global Surface Winds - *CC BY-SA 4.0*
    - 11.1.6: Jet Streams - *CC BY-SA 4.0*
  - 11.2: Air Masses and Fronts - *CC BY-NC-SA 4.0*
    - 11.2.1: Air Masses - *CC BY-SA 4.0*
    - 11.2.2: Surface Fronts - *CC BY-SA 4.0*
- 12: Weather Systems and Severe Weather - *CC BY-NC-SA 4.0*
  - 12.1: Mid-latitude Frontal Cyclones - *CC BY-SA 4.0*
  - 12.2: Tropical Cyclones - *CC BY-NC-SA 4.0*
    - 12.2.1: Tropical Cyclone Structure - *CC BY-NC-SA 4.0*
    - 12.2.2: Intensity and Geographic Distribution - *CC BY-NC-SA 4.0*
    - 12.2.3: Evolution - *CC BY-NC-SA 4.0*
    - 12.2.4: Hazards - *CC BY-NC-SA 4.0*
  - 12.3: Thunderstorms and Severe Weather - *CC BY-NC-SA 4.0*
    - 12.3.1: Thunderstorms - *CC BY-SA 4.0*
    - 12.3.2: Atmospheric Instability and Thunderstorms - *CC BY-SA 4.0*
    - 12.3.3: Precipitation and Hail - *CC BY-SA 4.0*
    - 12.3.4: Lightning - *CC BY-SA 4.0*
    - 12.3.5: Tornadoes - *CC BY-SA 4.0*
    - 12.3.6: Keeping Yourself Safe - *CC BY-SA 4.0*
- 13: Earth Materials - *CC BY-NC-SA 4.0*
  - 13.1: What is Geology? - *CC BY 4.0*
  - 13.2: Minerals and Rocks - *CC BY 4.0*
  - 13.3: Mineral Groups - *CC BY 4.0*
  - 13.4: Silicate Minerals - *CC BY 4.0*
  - 13.5: Mineral Properties - *CC BY 4.0*
- 14: Rocks and Geological Time - *CC BY-NC-SA 4.0*
  - 14.1: The Three Rock Types - *CC BY-NC-SA 4.0*
    - 14.1.1: Classification of Igneous Rock - *CC BY 4.0*
    - 14.1.2: Clastic Sedimentary Rocks - *CC BY 4.0*
    - 14.1.3: Chemical Sedimentary Rocks - *CC BY 4.0*
    - 14.1.4: Sedimentary Structures and Fossils - *CC BY 4.0*
    - 14.1.5: Classification of Metamorphic Rocks - *CC BY 4.0*
  - 14.2: The Rock Cycle - *CC BY 4.0*
  - 14.3: Reading Earth's History - *CC BY-NC-SA 4.0*
    - 14.3.1: Geological Time - *CC BY 4.0*
    - 14.3.2: The Geological Time Scale - *CC BY 4.0*
    - 14.3.3: Relative Dating Methods - *CC BY 4.0*
    - 14.3.4: Isotopic Dating Methods - *CC BY 4.0*
    - 14.3.5: Understanding Geological Time - *CC BY 4.0*
- 15: The Dynamic Earth - *CC BY-NC-SA 4.0*
  - 15.1: Understanding Earth Through Seismology - *CC BY 4.0*
  - 15.2: Plate Tectonics - *CC BY-NC-SA 4.0*
    - 15.2.1: Alfred Wegener- The Father of Plate Tectonics - *CC BY 4.0*
    - 15.2.2: Plate, Plate Motions, and Plate Boundary Processes - *CC BY 4.0*
    - 15.2.3: Mechanisms for Plate Motion - *CC BY 4.0*
  - 15.3: Earthquakes - *CC BY-NC-SA 4.0*
    - 15.3.1: What is an Earthquake? - *CC BY 4.0*
    - 15.3.2: Earthquakes and Plate Tectonics - *CC BY 4.0*
    - 15.3.3: Measuring Earthquakes - *CC BY 4.0*

        - 15.3.4: The Impacts of Earthquakes - *CC BY 4.0*
    - 15.4: Volcanoes - *CC BY-NC-SA 4.0*
        - 15.4.1: Plate Tectonics and Volcanism - *CC BY 4.0*
        - 15.4.2: Magma Composition and Eruption Style - *CC BY 4.0*
        - 15.4.3: Types of Volcanoes - *CC BY 4.0*
        - 15.4.4: Volcanic Hazards - *CC BY 4.0*
- 16: Weathering and Soils - *CC BY-NC-SA 4.0*
    - 16.1: Mechanical Weathering - *CC BY 4.0*
    - 16.2: Chemical Weathering - *CC BY 4.0*
    - 16.3: The Products of Weathering and Erosion - *CC BY 4.0*
    - 16.4: Weathering and the Formation of Soil - *CC BY 4.0*
- 17: Water and Landscapes - *CC BY-NC-SA 4.0*
    - 17.1: Topographic Maps - *CC BY-NC-SA 4.0*
    - 17.2: Surface Water - *CC BY-NC-SA 4.0*
        - 17.2.1: The Hydrological Cycle - *CC BY 4.0*
        - 17.2.2: Drainage Basins - *CC BY 4.0*
        - 17.2.3: Stream Erosion and Deposition - *CC BY 4.0*
        - 17.2.4: Stream Types - *CC BY 4.0*
        - 17.2.5: Flooding - *CC BY 4.0*
    - 17.3: Groundwater - *CC BY-NC-SA 4.0*
        - 17.3.1: Groundwater and Aquifers - *CC BY 4.0*
        - 17.3.2: Groundwater Flow - *CC BY 4.0*
        - 17.3.3: Groundwater Extraction - *CC BY 4.0*
        - 17.3.4: Groundwater Quality - *CC BY 4.0*
    - 17.4: Mass Wasting - *CC BY-NC-SA 4.0*
        - 17.4.1: Factors That Control Slope Stability - *CC BY 4.0*
        - 17.4.2: Classification of Mass Wasting - *CC BY 4.0*
        - 17.4.3: Preventing, Delaying, Monitoring, and Mitigating Mass Wasting - *CC BY 4.0*
- 18: Glaciers, Earth's History, and Other Worlds - *CC BY-NC-SA 4.0*
    - 18.1: How Glaciers Work - *CC BY 4.0*
    - 18.2: Glacial Erosion - *CC BY 4.0*
    - 18.3: Glacial Deposition - *CC BY 4.0*
    - 18.4: Earth's First 2 Billion Years - *CC BY 4.0*
    - 18.5: Are There Other Earths? - *CC BY 4.0*
- Back Matter - *Unknown License*
    - Index - *Undeclared*
    - Glossary - *Undeclared*
    - Detailed Licensing - *Unknown License*

This page titled Introduction to Earth and Space Sciences is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by LibreTexts.