

Critical Concepts

- CC1** Density and Layering in Fluids
- CC2** Isostasy, Eustasy, and Sea Level
- CC3** Convection and Convection Cells
- CC4** Particle Size, Sinking, Deposition, and Resuspension
- CC5** Transfer and Storage of Heat by Water
- CC6** Salinity, Temperature, Pressure, and Water Density
- CC7** Radioactivity and Age Dating
- CC8** Residence Time
- CC9** The Global Greenhouse Effect
- CC10** Modeling
- CC11** Chaos
- CC12** The Coriolis Effect
- CC13** Geostrophic Flow
- CC14** Phototrophy, Light, and Nutrients
- CC15** Food Chain Efficiency
- CC16** Maximum Sustainable Yield
- CC17** Species Diversity and Biodiversity
- CC18** Toxicity
- CC19** Metagenomics



The Critical Concepts are the heart of the science that you must learn in order to understand the basic processes of the Earth and its oceans. They are also concepts you will need to know before you can make sense of many issues that you will encounter in your future work or that will affect your life.

Some of these concepts are quite difficult to fully understand, especially without the use of mathematics at a level that is not appropriate for an introductory text like this one. Therefore, the Critical Concepts that follow are structured to allow easy learning of the essential nature of each concept. This essential nature is encapsulated in a section entitled “**Essential to Know**.” In this section, the essential aspects of each Critical Concept are listed as brief statements of all the key facts and characteristics of the concept. The information in these “**Essential to Know**” lists can be learned quickly and then used for easy reference while you study the areas of ocean sciences in which they are applied. This information is all you need to know to fully understand and enjoy the broad introduction to ocean sciences that this textbook provides. However, a fuller understanding of these Critical Concepts

might substantially enhance your introductory learning experience and will be necessary if you choose to study any aspect of ocean sciences at a more advanced level. Therefore, for each Critical Concept, the “**Essential to Know**” section is followed by a section entitled “Understanding the Concept,” which provides a more detailed explanation.

In the “Understanding the Concept” sections, the essential nature of the concept is explored in more depth. This section also describes some aspects of the scientific basis for, and general applications of, the concept in ocean sciences. To fully understand the “Understanding the Concept” section, you will need to read carefully and make constant reference to any figures noted. You will likely need to reread some of this material, perhaps several times, to fully understand it. However, you should remember that, as long as you learn the “**Essential to Know**” material, you will be able to understand the rest of this textbook and apply these concepts appropriately in class discussions, written projects, and exams.

CRITICAL CONCEPT 1

Density and Layering in Fluids

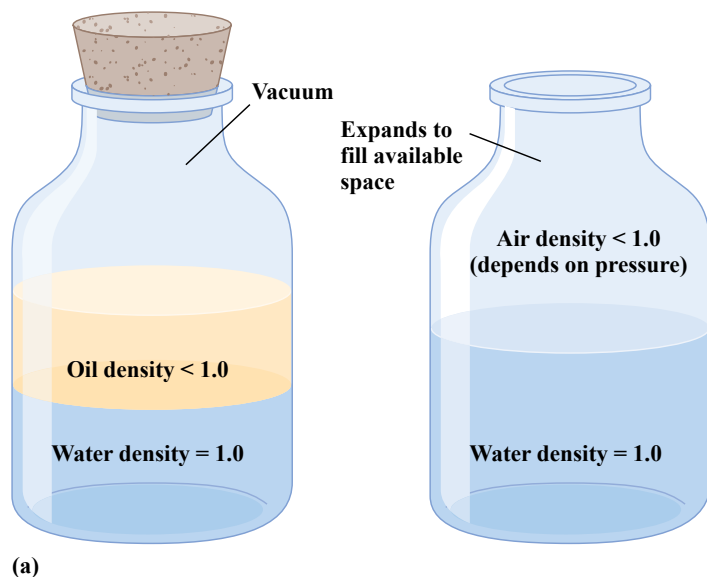
ESSENTIAL TO KNOW

- Fluids are separated by gravity into layers, each layer having a lower density than the one below it.
- Less dense fluids rise through more dense fluids.
- More dense fluids sink through less dense fluids.
- A fluid is neutrally buoyant when the fluid above has a lower density, and the fluid below has a higher density.
- A neutrally buoyant fluid will spread horizontally to form a layer.
- A fluid in which each higher layer is of lower density than the layer below it is stably stratified. There are no density-driven vertical motions within such a stably stratified fluid.

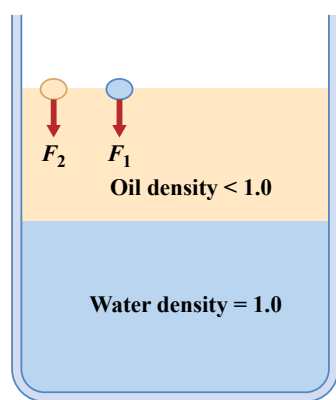
- The depth range in a fluid within which there is a marked vertical gradient of density is called a “pycnocline.” A pycnocline inhibits vertical mixing between fluid above and below the pycnocline layer.

UNDERSTANDING THE CONCEPT

All substances exist in one of three physical states: solid, liquid, or gas (**Chap. 5**). Both liquids and gases are fluids. Seawater is a liquid, and the atmosphere is a gas, so both are fluids that flow in response to forces such as **gravity**. Some parts of the Earth’s interior are liquid. Other parts, including the **asthenosphere** and **mantle** (**Chap. 4**), although they are solid, can flow extremely slowly at the high temperatures and pressures present



(a)



(b)

F_1 = Earth's gravitational attraction force on water droplet

F_2 = Earth's gravitational attraction force on oil droplet

$$F_1 = K \cdot \frac{m_w \cdot m_{\text{Earth}}}{r^2}$$

$$F_2 = K \cdot \frac{m_o \cdot m_{\text{Earth}}}{r^2}$$

Where: r = the distance from Earth's center to the center of the droplet

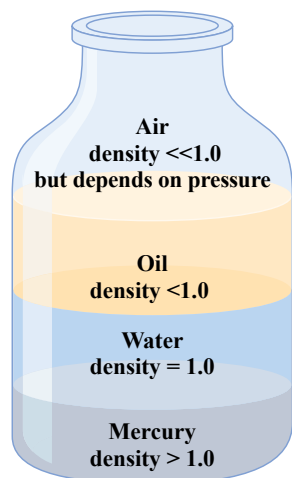
m_{Earth} = mass of Earth

K = gravitational constant

m_w = mass of water droplet

m_o = mass of equal volume oil droplet

Since $m_w > m_o$ then $F_1 > F_2$



(c)

FIGURE CC1-1. Fluids (liquids and gases) that do not mix with each other are separated into layers in such a way that the density of each layer is greater than that of the layer above. (a) Oil has a lower density than water and forms a separate layer that floats on the water. If this layer is very thin, it is called a “slick.” Air, which has a much lower density than either water or oil, forms a layer (the atmosphere) above the water. (b) Fluids are separated according to density because a more dense fluid has a greater mass and is subject to a greater gravitational force than the same volume of a less dense fluid. (c) Air, oil, water, and mercury separate into four layers, with density increasing toward the Earth's center (that is, with depth).

within the Earth. These solids behave like fluids over the million-year timescales in which geological processes occur.

When two fluids are brought into contact, gravity causes them to separate vertically according to their **densities**. The density of a fluid is the mass of the fluid divided by its volume. Pure liquid water has an absolute density (specific gravity) of $1000 \text{ kg} \cdot \text{m}^{-3}$ at 4°C and 1 atmosphere (atm) of pressure. Generally, density is stated as a relative density (r), which is the ratio of the fluid's specific gravity to the specific gravity of pure water at 4°C and 1 atm pressure. Thus, pure water at 4°C and 1 atm has a relative density of 1.

The effects of gravity and other forces on fluids of different densities account for many of the processes studied in this text, including the circulation of ocean water and atmosphere, and the formation of ocean basins and continents.

Many fluids do not mix with one another. Examples include oil and water, and water and air. When two such fluids are placed in a container, the denser fluid sinks to the bottom, leaving the less dense fluid above. If the less dense fluid is a gas, such as air, it expands to fill the upper part of the container not occupied by the denser fluid. If the less-dense fluid is a liquid such as oil, it forms a separate layer on top of the denser fluid (**Fig. CC1-1a**). Even if we vigorously shake such mixtures, they quickly separate again into layers after we stop shaking.

Gravity is the principal force that separates fluids of different densities. Consider a container full of water with an oil layer on top, into which we introduce a small drop of water (**Fig. CC1-1b**). Because water is denser than oil, the drop of water has a greater mass than the same volume of oil. The gravitational force between any object on the Earth's surface and the Earth increases in proportion to the object's mass. Therefore, the gravitational attraction force between the Earth and the water drop is greater than that between the Earth and the oil, and the water droplet is pulled down by the Earth's gravity more than the oil is. Because the oil can flow, the water droplet sinks, and the oil flows around it. When the water droplet reaches the top of the water layer, it is surrounded by a fluid of the same density, and the gravitational attraction on the water in the drop is now equal to the gravitational attraction on all the other water at its level. Therefore, the droplet sinks no farther. However, it does mix with the surrounding water by diffusion.

Gravitational sorting of fluids into layers according to density is called **stratification**. This process works equally well with more than two fluids. For example, if we mix mercury, water, oil, and air, the mixture will separate with water on top of the mercury, oil on top of the water, and air above the oil (**Fig. CC1-1c**). This process does not work quite the same way with solids because they cannot flow.

Density stratification can occur not only between fluids that do not mix, but also between fluids that do mix. For example, if we carefully add cream to a cup of cold coffee, it remains as a floating layer on top of the coffee until we stir the coffee and thoroughly mix the two layers. After two liquids are thoroughly mixed, the mixture has uniform density, and the two layers do not re-form.

Density stratification can also occur in a single fluid, such as pure water or air. Just as cream forms a layer on coffee, a less dense layer of any fluid will remain on top of a denser layer of the same fluid until they are mixed (**Fig. CC1-2**). In the natural environment, we often find fluids with multiple layers, each of

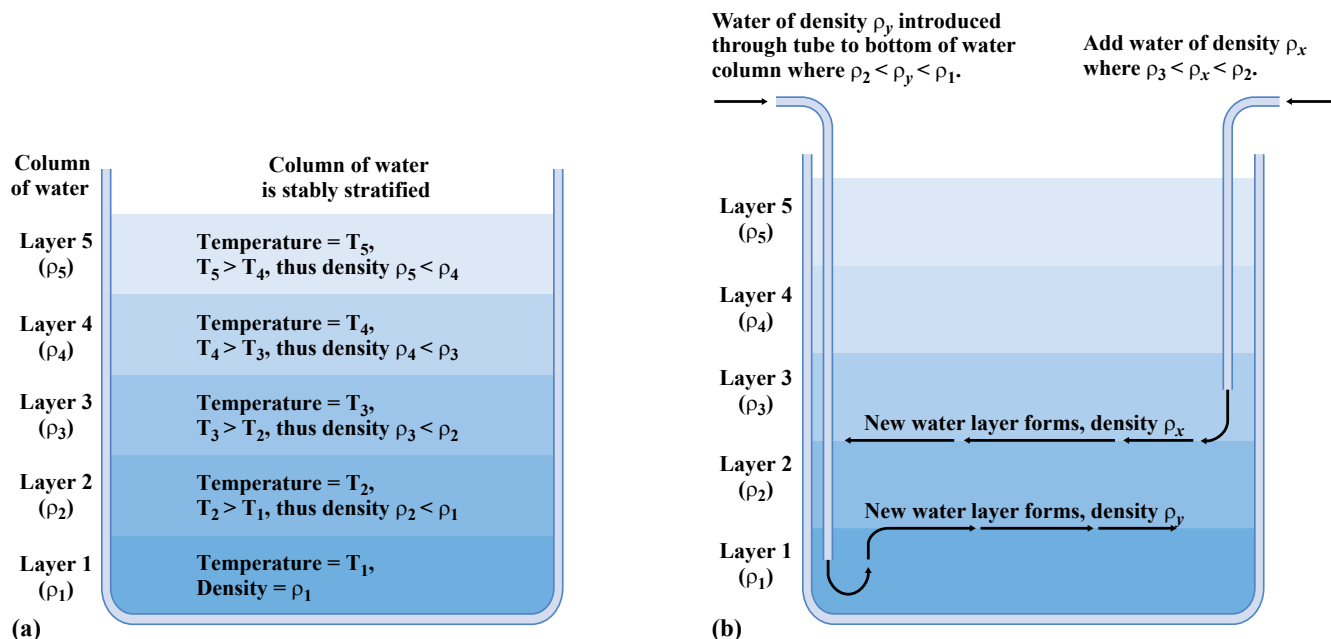


FIGURE CC1-2 Water may form distinct layers, each of which has a different density because it has a different temperature or salinity. (a) A water column is stably stratified if each successive layer has a higher density than the one above it and, thus, no water sinks or rises. (b) If water is introduced to a stratified water column, it will sink or rise until it reaches an equilibrium level at which the density of the water above it is lower and the density of the water below it is higher. It will then spread out to form a new layer at this depth. Some mixing of the introduced water with other water layers will occur as it sinks through layers of lower-density water or rises through layers of higher-density water. However, when the vertically moving volumes of water are very large (as they are in the oceans), such mixing is limited.

slightly higher density from the shallowest to the deepest layer (**Fig. CC1-2**). When the density of the fluid increases progressively with depth, no layer has a tendency to sink or rise, and the fluid is said to exhibit stable stratification. Ocean waters are arranged in layers of different density in this way (**Chap. 8**) because the density of water varies with temperature and salinity (**CC6**). The atmosphere is also vertically stratified (**Chap. 7**), and the components of the Earth's interior are arranged similarly (**Chap. 4**).

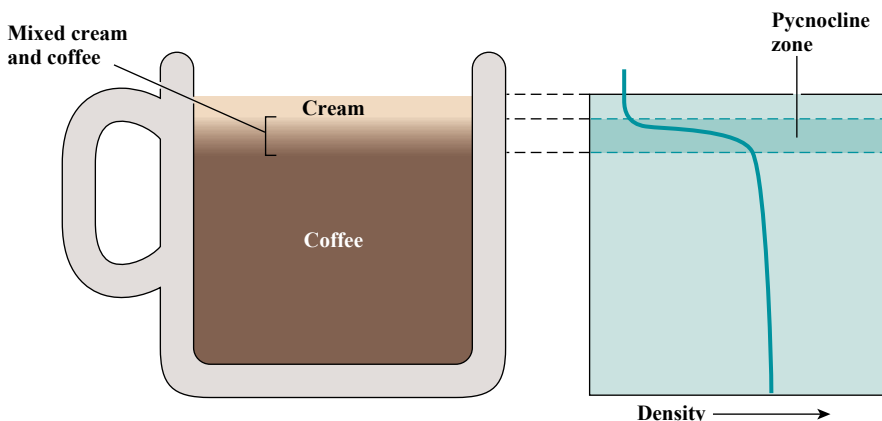
If we add a fluid to a column of other fluids, each with a different density and none of which mix, the added fluid will sink or rise until it reaches a depth where the fluid immediately above it is of a lower density and the fluid immediately below it is of a higher density. At this equilibrium level, the introduced fluid is neutrally buoyant, so it will neither sink nor rise, but will spread out to form its own layer (**Fig. CC1-2**). For example, if we introduce oil underwater to an air-water layered system, the oil will rise to the surface of the water and spread out. This is exactly what happens when sunken ships release their oil.

When a fluid of slightly different density is added to a

density-stratified column of the same fluid or of a fluid with which it mixes, some mixing will occur as the added fluid sinks or rises to its equilibrium level. We can see this effect if we pour cream from a height into a cup of coffee. Both the cream and the coffee are primarily water, but different chemicals are dissolved or suspended in each, so they have different densities. If we pour the cream in carefully, it will be partially mixed as its momentum carries it down into the coffee and as it subsequently rises because of its lower density. The cream, mixed with some coffee, will rise to form a layer on top of the coffee and will then mix only very slowly (by molecular diffusion) unless we stir the mixture (and create turbulent diffusion). When the volumes of fluid in each layer are very large and vertical motions are slow, as they are in the oceans, mixing is very limited. The oceans are vertically stratified (**Chap. 8**), and density-driven vertical motions of water masses move large volumes of water vertically to find their equilibrium level with little mixing between the ascending or descending water masses and the layers of water through which they rise or sink.

Our cream and coffee experiment can illustrate another

FIGURE CC1-3. If cream (which has a lower density than coffee) is introduced at the surface of a cup of cold coffee, the cream forms a surface layer overlying the coffee. Between the cream and coffee layers, there is a zone within which the cream and coffee are mixed in proportions that change progressively with depth from pure cream to pure coffee. In this zone between the coffee and cream layers, there is a rapid change in their respective proportions, which results in a sharp increase in density with depth. This zone, in which density changes rapidly with depth, is called a pycnocline zone. Vertical mixing between the two layers takes place only slowly across a pycnocline because the lower-density upper layer (cream) has no tendency to sink, and the higher-density layer (coffee) has no tendency to rise.



important concept. After we carefully place a layer of cream on the coffee, there is a distinct cream layer of lower density and a distinct lower coffee layer of higher density. Between these layers is a thin region where coffee and cream are mixed in varying amounts. The vertical distribution of density in this experiment is shown in **Figure CC1-3**. Density is uniform in the surface layer of cream, becomes progressively higher (this is referred to as a density gradient) in the transition layer of mixed cream, and is uniform in the lower layer of coffee. The transition region in which there is a vertical density gradient is called a pycnocline. More energy would be needed to move a molecule of water vertically through this **pycnocline** than would be needed to move a molecule the same vertical distance in either the cream or the coffee. Thus, pycnoclines act as barriers to vertical mixing.

tendency to sink, and the higher-density layer (coffee) has no tendency to rise.

The strength of the density gradient in a pycnocline determines how strong a barrier to vertical mixing it is. Try the experiment again, but use milk instead of cream. Milk has a slightly

higher density than cream (because it contains less fat). It is almost impossible to pour the milk into the coffee without the two mixing, even though milk does have a lower density than water. In this case, the density gradient, and therefore the pycnocline between milk and coffee, is very weak.

Pycnoclines form in many parts of the oceans, particularly where warm surface water overlies colder deep water, and they are important because they inhibit vertical mixing. The density of a fluid is altered as heat is gained from the sun at its surface and as heating or cooling occurs by conduction and radiation at its interfaces with its container or other fluids surrounding it. Density is also altered by changes in pressure and by changes in fluid composition, such as the addition of salts to water (**Chap. 5**) and of water vapor to air (**Chap. 7**). The factors that affect the density of air are discussed in **CC5** and **Chapter 7**, and the factors that affect the density of water are examined in **CC6** and **Chapter 5**. **CC3** explains more about the vertical motions created by such density changes.

CRITICAL CONCEPT 2

Isostasy, Eustasy, and Sea Level

ESSENTIAL TO KNOW

- The level at which a solid floats in a liquid is determined by the relative densities of the solid and the liquid.
- Solids of equal density float in a liquid with the same percentage of their volume submerged.
- Two solids of different densities will float in a liquid in such a way that the solid with the greater density has a greater fraction of its volume submerged than the less dense solid.
- Lithospheric plates float on the asthenosphere. Because oceanic crust is denser and thinner than continental crust, the surface of the oceanic crust is always lower than the surface of the continental crust.
- Changes occur in the density and thickness of lithospheric plate sections as a result of various processes. The affected plate section rises or falls in response to such changes in a process called “isostatic leveling.”
- Isostatic leveling is slow. Equilibrium may not be achieved for millions of years after an event, such as the start or end of an ice age, that changes the distribution of glaciers.
- Sea level changes on a section of coast as the section of the plate rises or falls isostatically. Isostatic sea-level changes occur at different rates on different coasts.
- Eustatic sea-level changes are caused by changes in the volume of ocean water or in the volume of the ocean basins that occur as a result of a variety of plate tectonic and climatic change processes.
- Eustatic changes of sea level take place simultaneously and uniformly throughout the world. Eustatic equilibrium is easily and quickly attained.
- Changes of sea level on different sections of the world’s coasts occur at different rates, and even in different directions, because of the interaction of eustatic and isostatic changes.

UNDERSTANDING THE CONCEPT

Solid objects will sink through a fluid in which they are placed if the **density** of the solid is higher than that of the fluid. For example, a rock thrown into a pond sinks to the bottom. Solid objects whose density is less than that of the fluid will float. For example, wood or Styrofoam will float on water, and helium balloons will float on air.

When a solid object floats on a liquid, the depth to which the object is immersed is determined by its density and the density of the liquid. A simple principle describes this relationship. Called Archimedes’ principle, it states that the floating solid will be immersed in the liquid deep enough that the water it displaces has exactly the same total mass as the solid.

We can understand Archimedes’ principle by considering a rectangular block of wood floating on water (**Fig. CC2-1**). If the wood has a density of 0.5, or half the density of water, the wood floats with exactly half of its volume underwater and half exposed. If the wood has a density of only 0.1, or one-tenth the density of water, the wood floats with only one-tenth of its volume underwater and nine-tenths exposed (**Fig. CC2-1a**). If we place a weight on top of the less dense piece of wood, the wood is forced to sink into the water. It sinks until the mass of the total volume of water displaced is equal to the total mass of the wood plus the weight (**Fig. CC2-1b**).

If two pieces of wood have the same density but different thicknesses, the thicker piece will float with its upper surface higher above the water surface than the thinner piece, and it will extend deeper into the water (**Fig. CC2-2**). The two pieces have the same percentage of their individual volumes submerged. If we have two pieces of wood of identical dimensions but different densities, the wood with higher density will float lower in the water (**Fig. CC2-1a**). This effect can be observed in a river, lake, or ocean. For example, most logs or driftwood pieces float with much of their volume immersed below the waterline. In contrast,

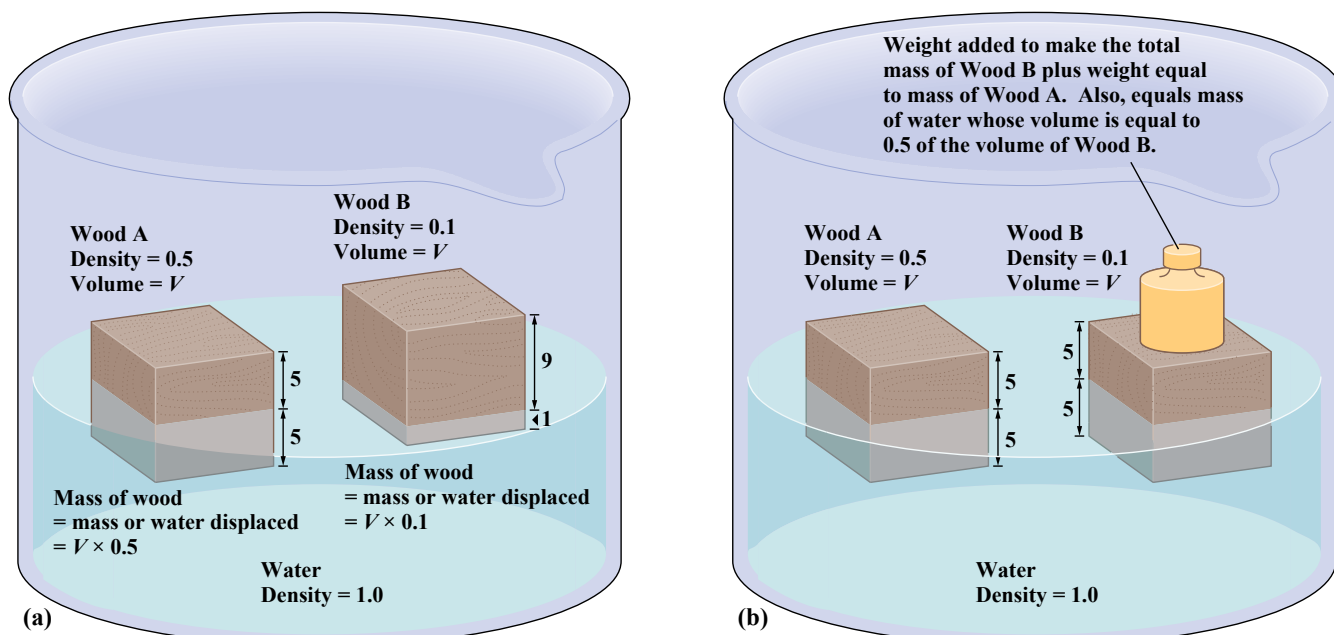


FIGURE CC2-1. Archimedes' principle determines how much of a floating solid's volume is below the surface. The solid floats at a depth such that the weight of the volume of water or other fluid that it displaces (that is, the volume of the part of the solid that is below the water or fluid surface level) equals the total weight of the solid. (a) Lower-density materials will float higher than higher-density materials. (b) Adding weight to a floating, low-density solid increases its mass and causes it to float lower and thus displace more water. The volume of water displaced will have a weight that is equal to the total weight of the floating solid plus the added weight that it supports.

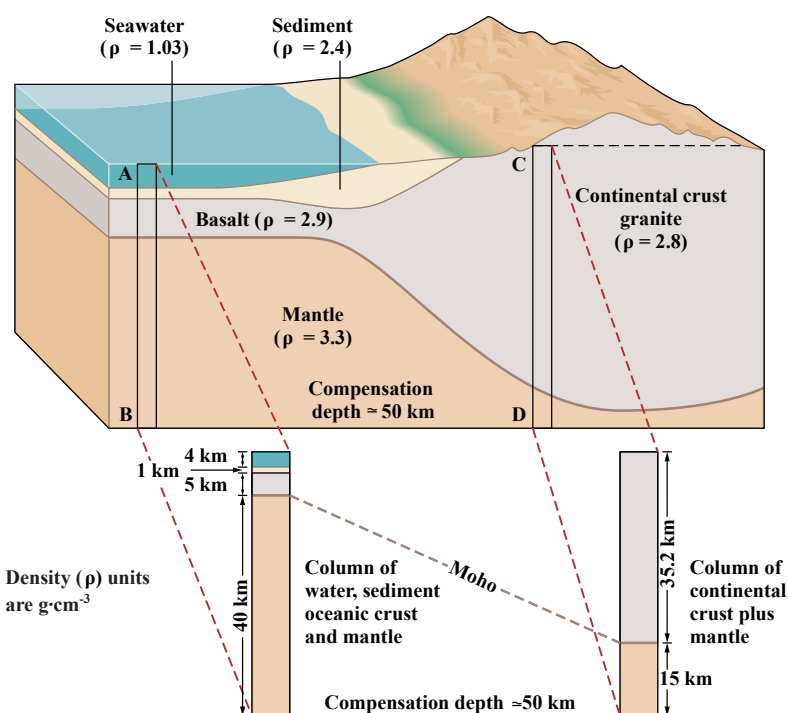
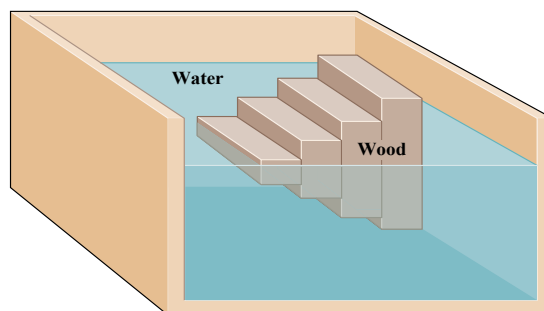
Styrofoam and balsa wood seem almost to sit on the water, because their low density means that only a small percentage of their volume is immersed.

Lithospheric plates float on the fluid surface of the **asthenosphere** in the same way that wood floats on water. Therefore, a **buoyancy** equilibrium is established in which the plates float freely on the material beneath them at a level that is determined by their density and thickness. This level is called **isostasy**, or "isostatic equilibrium."

Continental crust has a lower density than oceanic crust (**Chap. 4**). Therefore, lithospheric plates covered with continental crust float with a smaller percentage of their volume "submerged" than plates covered with oceanic crust. If the plates with continental crust and oceanic crust were the same thickness (**Fig. CC2-3**), the upper surface of the continental crust plate (the land surface) would be higher than the surface of the oceanic crust plate (the seafloor). In addition, the lithosphere-asthenosphere boundary would be shallower (farther from the Earth's center) under the continents than under the seafloor. However, lithospheric plates with oceanic crust are thinner than those with continental crust. Because the two types of crust are of different thicknesses, the lithosphere-asthenosphere boundary is shallower below ocean crust than it is below continental crust (**Fig. CC2-3**), and the continent

FIGURE CC2-3. Isostasy is the condition in which blocks of lithosphere float on the asthenosphere at the equilibrium level determined by Archimedes' principle. Above the compensation level (the depth at which the asthenosphere behaves as a fluid such that it can be displaced by floating lithospheric plates), the total weight of a column of continental crust plus mantle at isostasy will equal the total weight of a column of water, sediment, oceanic crust, and mantle. They do not quite do so in this figure, because the numbers are rounded.

FIGURE CC2-2. Solids of the same density but of different thicknesses always float with the same proportion of their volume below the surface. Thicker blocks float with their upper surface higher above the surface of the fluid on which they float, and their lower surface deeper below.



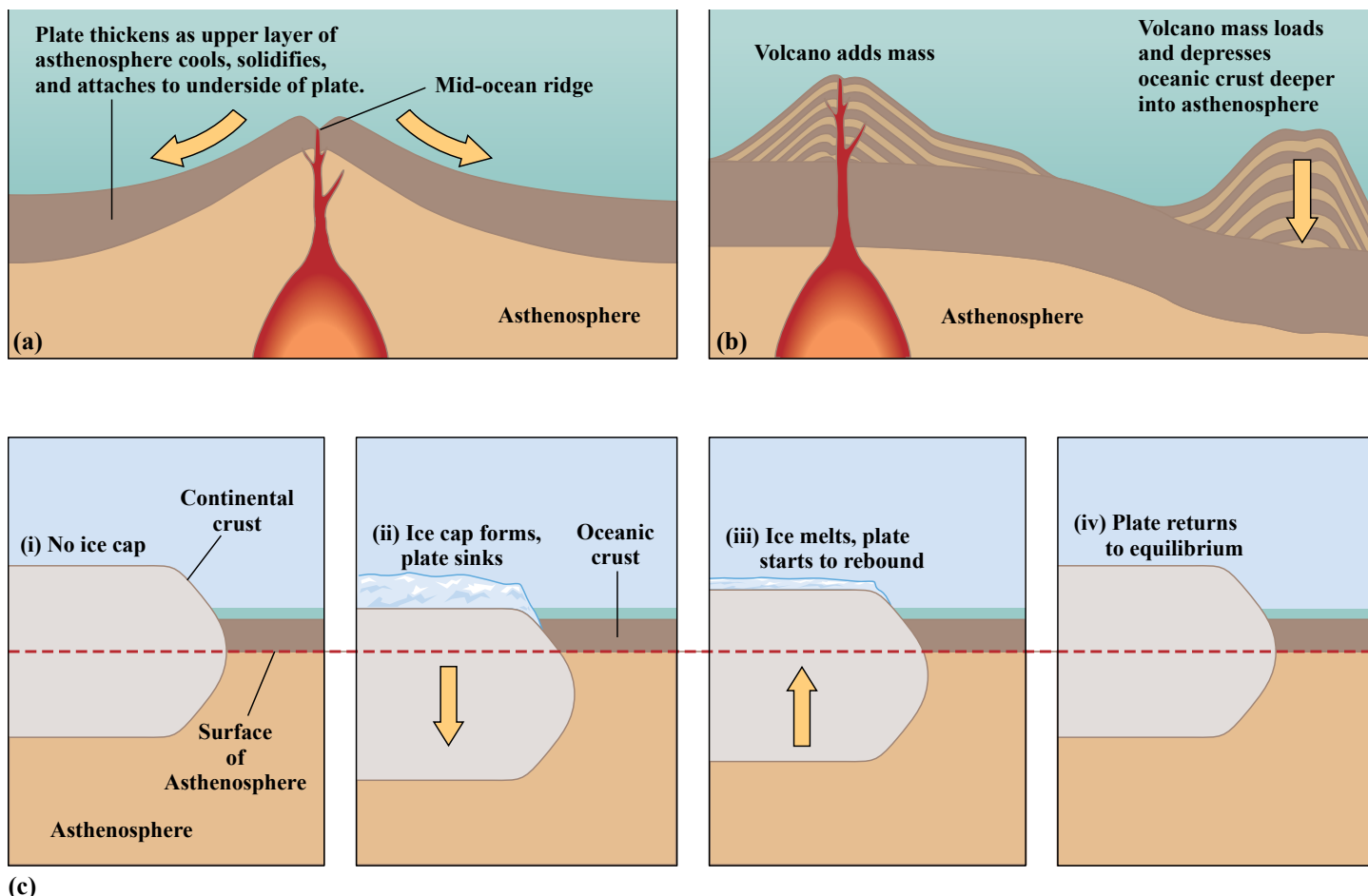


FIGURE CC2-4. Isostatic leveling takes place in response to crustal density changes, plate thickening, and plate loading processes. (a) As oceanic crust moves away from an oceanic ridge, it cools, its density increases, and it sinks lower in the asthenosphere. The plate also thickens as mantle material solidifies and is accreted to the cooling underside of the plate. (b) Volcanoes formed at hot spots add mass to the crust. The increased mass causes the plate to bend and sink lower under the volcano, especially after the old volcano moves away from a hot spot and cools, thus increasing its density. (c) A section of continental crust on which glaciers are formed will sink to a new isostatic equilibrium because of the additional weight of the glacier (like the wood in [Figure CC2-1b](#)). If the ice melts, the continent will rise in a process called “isostatic rebound” until it reaches its new equilibrium level. Isostatic changes take place very slowly compared to the climate changes that can cause large variations in the area of the continents that is covered by glaciers.

surface is more elevated above the oceanic crust surface than it would be if ocean and continental crust were the same thickness.

If the density and thickness of the lithospheric plates were invariable, the continents and seafloor would always remain at a fixed equilibrium height above the asthenosphere. However, both the density and the thickness of lithospheric plates are altered by a variety of processes. First, the crust can be heated, which reduces its density; or cooled, which increases its density. Second, the thickness of the crust can be increased by the formation of mountains or volcanoes or by the **deposition** of large amounts of **sediment**, or it can be thinned by the **erosion** of mountains or stretching of the lithospheric plate. Third, the thickness of the lithospheric plate can be increased by the cooling and solidification of upper mantle material. This material is added to the bottom of the plate. Conversely, the plate can be thinned by the heating and melting of mantle material on the underside of the plate. Finally, the volume and mass of the continental crust can be increased by the development of **glaciers**, or decreased if glaciers melt, or altered by several other means, such as the growth of **coral reefs**.

If the density of a section of lithospheric plate is increased, it sinks until it reaches its new equilibrium level in a process known as isostatic leveling ([Fig. CC2-4](#)). Similarly, crust whose density

has decreased will rise isostatically. Changes in thickness of the crust will also cause isostatic leveling ([Fig. CC2-4](#)).

Isostatic leveling is very slow because the asthenosphere is very **viscous** and flows extremely slowly to accommodate the rising or sinking lithospheric plate. The processes that cause changes in isostatic level are often localized to certain sections of a lithospheric plate and its overlying crust. Therefore, one section of a plate may be rising while another is sinking. Lithospheric plates can bend to accommodate this process. When the section of a lithospheric plate that supports the edge of a continent changes its isostatic level, sea level rises or falls along this **coast** but does not necessarily change on other coasts.

Sea level can be altered not only by isostatic leveling of the continents, but also by processes that increase or decrease either the volume of ocean water or the volume of the ocean basins. Changes in water volume or in ocean basin volume cause changes in sea level that occur simultaneously and almost uniformly along all the world’s coasts. Simultaneous worldwide changes in sea level are called **eustatic** changes ([Fig. CC2-5b](#)).

Eustatic sea-level changes can be caused by several processes. For example, the volume of water in the oceans can be increased by the addition of water from melting glaciers or by elevation of the average temperature of ocean water that causes the water to

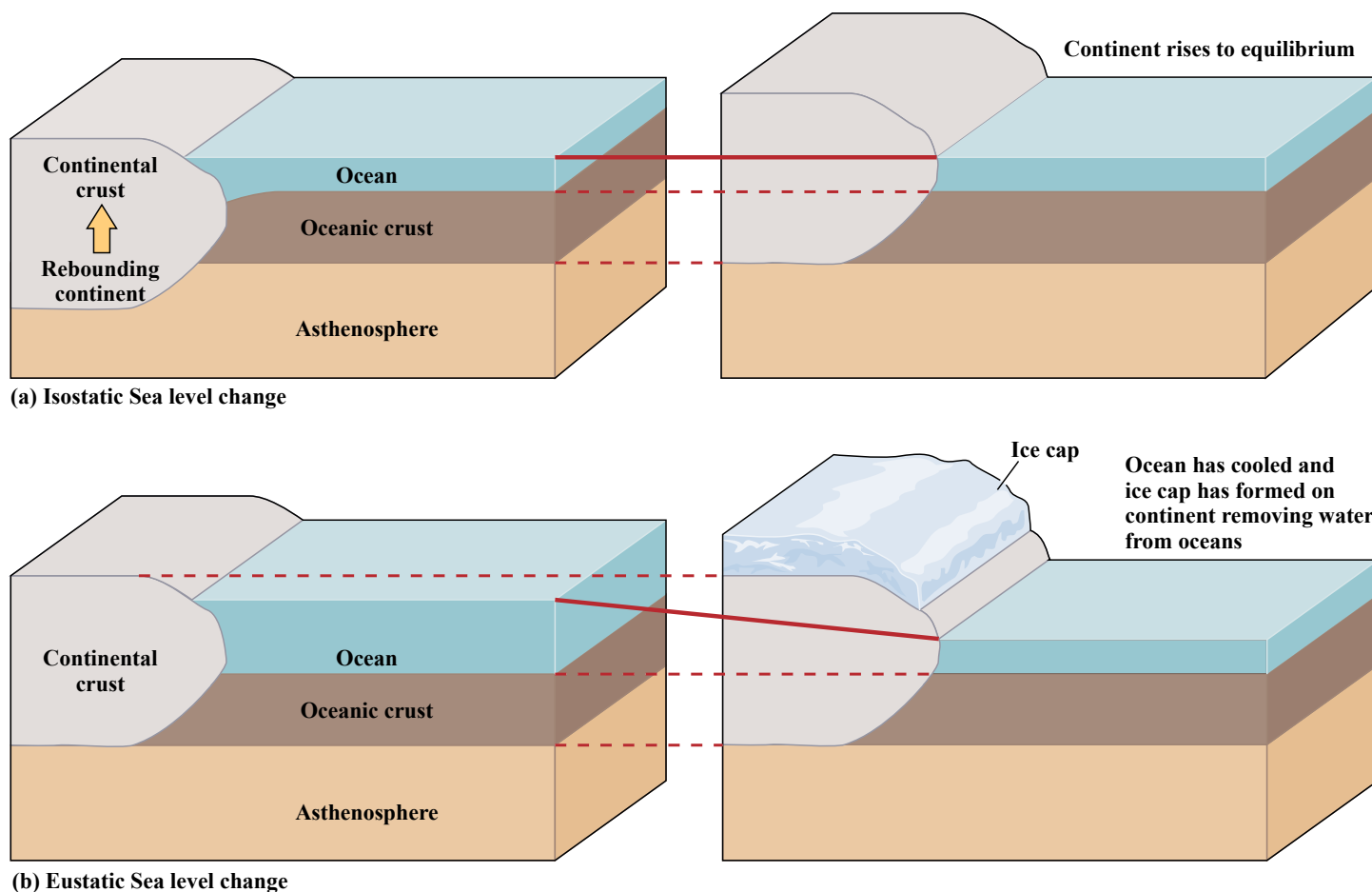


FIGURE CC2-5. (a) Isostatic sea-level changes take place when the land rises or falls while the seafloor and ocean depth remain the same. In this case, a continent that was previously depressed by the weight of an ice sheet rises slowly, exposing more of the continent to the atmosphere. (b) Eustatic sea-level changes take place when the volume of water in the oceans changes. In this case, the ocean has cooled and the water has been transferred to an ice cap on the land, thus lowering the sea level. Notice that the continent has not sunk lower in the asthenosphere in this diagram. In the situation depicted, the continent would eventually sink isostatically, but isostatic changes are much slower than eustatic changes.

expand. Correspondingly, the volume of ocean water can be decreased by increased **glaciation** or by a decline in average ocean water temperatures.

The volume of the ocean basins (actually the volume available for seawater to fill below a fixed level, independent of vertical movements of the continents) can be decreased if a larger percentage of the ocean basin floor is occupied by **oceanic ridges**. During the parts of **spreading cycles** when continents are being broken up, such as at present, there are many continents moving apart and many oceanic ridges between them. Hence, the percentage of the ocean floor covered by oceanic ridge is larger than it is during periods of reassembly of the continents. As a result, the ocean basin volume tends to be diminished, and sea level is correspondingly high.

Changes in the rate of **seafloor spreading** within a spreading cycle also affect sea level. When seafloor spreading is fast, the young, warm oceanic crust extends far from each oceanic ridge. Therefore, the volume of the ocean basin is reduced, sea level is higher, and more continental crust is covered by oceans. When seafloor spreading is slow, a smaller amount of young, warm oceanic crust is produced, the volume of the ocean basins is increased, and sea level is lower. During periods when the Earth's continents are assembled in supercontinents such as Pangaea, the number of oceanic ridges and the rate of seafloor spreading are reduced. Therefore, young, warm oceanic crust covers a smaller percentage of the ocean floor, the ocean basin volume is in-

creased, and sea level is lower. The processes, just described, that affect the volume of the ocean basins are complicated and interact with each other as the various **plate tectonic** processes simultaneously change the thickness, relative abundance, and isostatic level of the continental and oceanic crusts. At the same time, **climate** changes, which themselves may be linked to tectonic processes, also change the volume of water in the oceans.

Processes that alter sea level occur on a variety of different timescales, from centuries to millennia. Eustatic equilibrium is reached quickly in response to such changes, whereas isostatic equilibrium is attained very slowly. Sea level may be rising, static, or falling on any individual section of the world's coasts as various eustatic and isostatic changes work in concert or in opposition in different regions.

At present, there is concern that human enhancement of the **greenhouse effect** (Chaps. 1, 7, CC9) will lead to substantial additional global warming. One fear is that if sufficient warming occurs, it will increase the volume of ocean waters as the average temperature increases and the water expands, and as partial melting of the polar ice sheets contributes additional water. The result is likely to be an acceleration in the eustatic rise in sea level, which would have severe consequences because many coastal regions and cities would be inundated by the sea. Indeed, the acceleration of the rate of sea level rise has been observed in currently available data. Therefore, measuring the current rate of eustatic sea-level change is considered to be critically important.

Measurement of the rate of eustatic change of sea level is complicated by the interaction of isostatic and eustatic changes that occur simultaneously on any given section of coast. For example, the sea-level change measured on the northeast coast of the United States will be the net result of several processes: rising sea level due to any greenhouse warming effect, rising sea level due to progressive cooling and isostatic sinking of the continental crust, and falling sea level due to isostatic rise caused by the

geologically recent melting of the **ice age** glaciers that once lay on this crust. As a further complication, human activities such as **groundwater** and oil withdrawal cause the subsidence of certain sections of the coast. Because of these complicated interactions, we are able to measure any greenhouse-induced sea-level rise only by studying sea level change on many different coasts or by extremely precise and difficult satellite measurements.

CRITICAL CONCEPT 3

Convection and Convection Cells

ESSENTIAL TO KNOW

- Vertical motions in a fluid can be caused by temperature or compositional changes that alter the density of parts of the fluid.
- When density continuously decreases within one layer of a fluid (often at its lower boundary with another fluid or solid) and continuously increases in a higher layer (often the fluid surface), a convection cell is established.
- In a convection cell, a plume of lower-density fluid rises within the fluid until it reaches an equilibrium level where the surrounding fluid is of equal density. If the rising fluid has a lower density than all fluid layers above it, it rises to the surface. At its equilibrium level (or the surface), the fluid spreads out. If it is cooled or otherwise altered so that its density increases, the fluid eventually becomes dense enough to sink back to its original level, where it replaces rising fluid and reenters the cycle.
- Some convection cells have toroidal (doughnut-shaped) circulation, in which a rising column of fluid spreads out laterally in all directions at the top of the convection cell and returns by sinking in a ring-shaped band surrounding the rising plume. This pattern may be reversed with a column of sinking fluid surrounded by a ring of rising fluid.
- In other convection cells, both the rising fluid and the sinking fluid form an elongated curtain. The simplest example of such a convection cell is cylindrical.
- The rising plume is called an “upwelling plume,” and the sinking plume is called a “downwelling plume.”
- At the top or bottom of a convection cell, the fluid flows horizontally away from a location at a divergence. The fluid flows toward a location at a convergence.
- At the top of a convection cell, divergences are regions of upwelling, and convergences are regions of downwelling. At the bottom of the convection cell, divergences are regions of downwelling, and convergences are regions of upwelling.
- Several convection cells may be formed side by side within a fluid. Convergences and divergences always alternate across the top or bottom of the cells.

UNDERSTANDING THE CONCEPT

CC1 explains how density differences cause stratification in fluids, in which successive layers have higher density with depth. If the density of the layers in a vertically stratified fluid does not change, the system is stable and vertical motions do not occur.

Three stratified fluids are important to oceanographers: the Earth’s interior layers, the oceans, and the atmosphere. Heat is introduced into the Earth’s layers by energy released during **radioactive** decay in the core and **mantle** (**Chap. 4**). In the oceans, a small amount of heat is introduced by conduction from the mantle through the seafloor, and a much larger amount by solar heating of surface water. Heat is also exchanged (gained in some areas and lost in others) between oceans and atmosphere by **conduction**, radiation, evaporation, and precipitation (**Chaps. 5, 7**). The atmosphere exchanges heat with the oceans and land, gains heat from solar radiation, and loses heat by radiation to space (**Chap. 7**).

The heat transfer processes between sun, atmosphere, ocean, and land all vary with time and with location on the Earth. Variable heat transfers produce density changes in the mantle, ocean, and atmosphere and cause the stratification to become unstable. Changes in composition can also alter density. For example, changes in the salinity of ocean water (**Chap. 5**) or in the water vapor pressure of air (**Chap. 7**) alter the fluid density. When

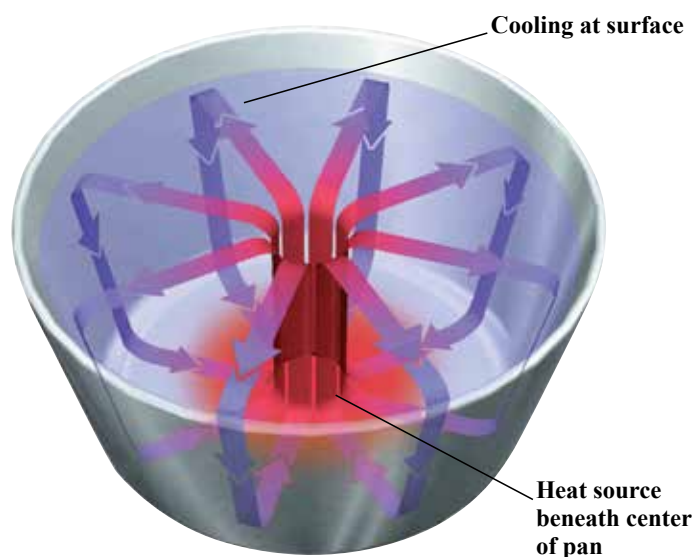


FIGURE CC3-1. In a saucepan heated at the center of its bottom, water will establish a circulation in which heated water rises to the surface, spreads toward the pan’s sides, cools, and then sinks back to the bottom. This toroid-shaped convection cell is established because the water’s density decreases when it is heated and increases when it is cooled. The idealized toroidal convection cell depicted would, in practice, normally be distorted because saucepans are usually heated across their entire base, not just at the center, and by turbulence in the flow patterns caused by variations in the rate of heating and cooling.

stratification becomes unstable, vertical motions called convection occur.

We can best understand convection using a simple kitchen analogy: the motion of water in a saucepan heated on the stove (Fig. CC3-1). Initially, when we place the saucepan on the burner, all the water within the pan is at room temperature. No vertical motion occurs because the water is uniform in density. When we begin heating the base of the saucepan, the water at the bottom is heated, while the water above remains at room temperature. As bottom water is heated, its density decreases, and it rises through the overlying water to the surface. If you watch a saucepan of water carefully as you start to heat it, you can see the vertical water movements as swirls of motion.

When the water is heated above room temperature, it begins to lose heat through the saucepan's surface and sides (Fig. CC3-1). Heated water rises to the surface, spreads out, and cools. As more warmed water rises to the surface, it forces the cooler surface water toward the sides of the saucepan. Because heat is lost through the water surface, the surface water that is displaced to the edges of the pan is cooled slightly and has a slightly higher density than the newly warmed water rising to the surface at the center of the pan. Accordingly, warmer water continuously flows toward the edge of the pan, cooling as it flows. As we continue to heat the saucepan, a continuous recycling motion is established. Water heated at the bottom of the water column rises at the center of the pan and is replaced by cooler water flowing in from the sides. The heated water rises to the surface, then spreads out, cools, and sinks down the sides of the pan to be heated again (Fig. CC3-1). The closed cycle of fluid circulation, driven by warming at depth that causes the fluid to rise, cooling at a lower-density level, and sinking back to its original depth to rejoin the cycle, is called a **convection cell**.

Convection is not as uniform as depicted in Figure CC3-1. The heating of the saucepan bottom and the heat loss through the water surface and the pan sides vary. Therefore, the location of the rising plume of the convection cell fluctuates, and more than one rising plume is often present. In addition, the water flow

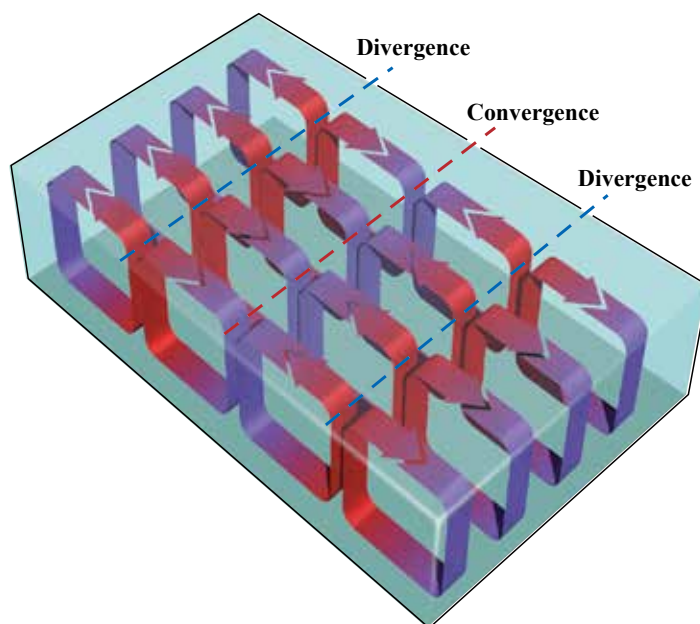


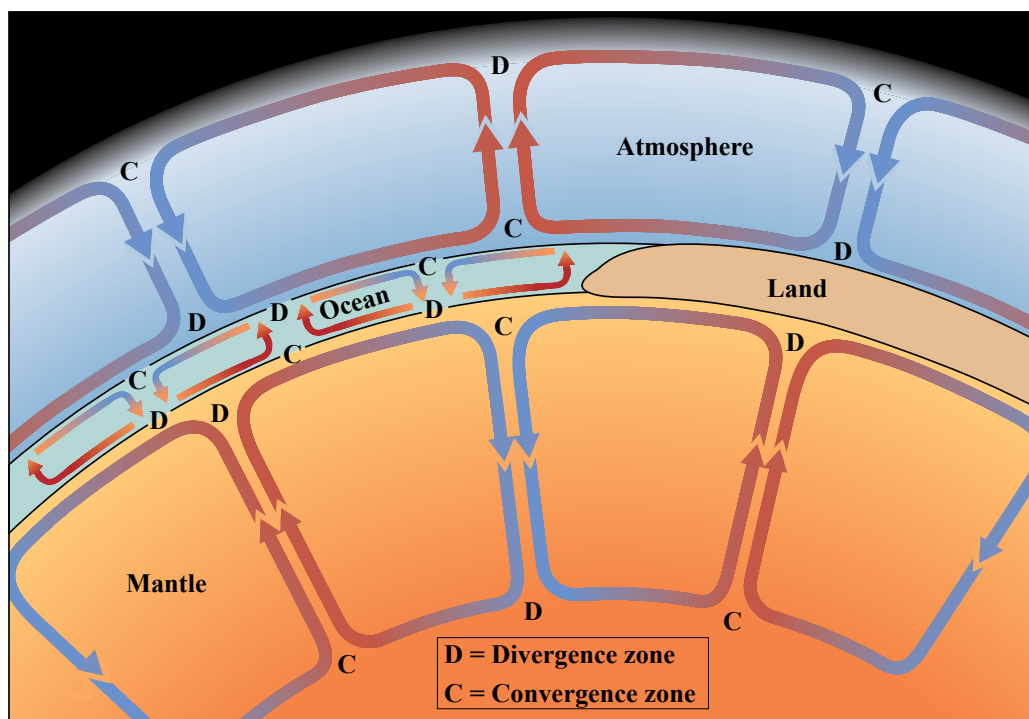
FIGURE CC3-2. Cylindrical convection cells. Note that convergences and divergences alternate between adjacent cells.

is **turbulent** rather than smooth. Turbulence can be seen in the internal swirling, churning motions in fast-running streams.

If we set our stove on low heat, the water in the saucepan does not boil, and convection cell motion continues indefinitely. Heat is added continuously and lost at the same rate that it is added. In nature, most convection cells operate in this way.

Convection cells transport heat vertically within a fluid. Convection also establishes areas on the surface and at the bottom of the cell where warmer fluid (in the center of the saucepan) and cooler fluid (at the edges of the saucepan) are concentrated. These features of convection cells are important in many ocean processes, including **plate tectonics** (Chap. 4), ocean water circulation (Chap. 8), and atmospheric circulation (Chap. 7). The upper and lower boundaries of convection cells are often the interfaces between two fluids (e.g. water and air) or between a fluid and a

FIGURE CC3-3. The mantle, oceans, and atmosphere each have convection cells with alternating divergences and convergences across their upper and lower boundaries. In the atmosphere, we are generally concerned with the lower (sea-level) boundary of the convection processes, where divergences are upwelling zones, and convergences are downwelling zones. In the mantle and oceans, we are generally concerned with the upper (sea surface and upper surface of the asthenosphere) boundaries of the convection processes, where divergences are upwelling zones, and convergences are downwelling zones.



solid (e.g., the ocean surface and the seafloor). However, convection cells may develop between any two density surfaces within a vertically stratified fluid.

The convection cell just described is toroidal (doughnut-shaped), with a concentrated rising plume of heated water at its center and a ring of sinking water surrounding this plume (Fig. CC3-1). In another configuration of convection cells, both the rising plume and the sinking plume are extended laterally. In its simplest form, this type of convection cell is cylindrical in shape (Fig. CC3-2). Because fluid cannot both rise and sink at the same location, convection cells must be arranged so that the rising and sinking plumes alternate (Fig. CC3-2). The areas where fluid rises are upwelling zones, and areas where it sinks are downwelling zones. Upwelling and downwelling areas must alternate across the top or bottom of the fluid. We cannot go from one upwelling zone to another upwelling zone without passing through a down-

welling zone. The locations of upwelling and downwelling in the atmosphere, ocean, and Earth's mantle are important to many aspects of Earth's geology, chemistry, climate, and biology.

At the top of a convection cell, fluid flows horizontally away from upwelling areas as it is displaced by upwelled fluid. Such areas are called divergences (Fig. CC3-3). Similarly, fluid flows into downwelling areas to replace downwelled fluid. Such areas are called convergences (Fig. CC3-3).

This text primarily considers processes that occur at the tops of convection cells (e.g., in the upper mantle and ocean surface waters), where upwelling zones are divergent, and downwelling zones are convergent (Fig. CC3-3). When we consider atmospheric circulation, we are concerned primarily with the bottom of the convection cell at the Earth's surface, where upwelling zones are convergences and downwelling zones are divergences (Fig. CC3-3).

CRITICAL CONCEPT 4

Particle Size, Sinking, Deposition, and Resuspension

ESSENTIAL TO KNOW

- The sinking rate of suspended particles in the ocean is determined primarily by particle size. Large particles sink quickly, and small particles sink more slowly.
- Large particles are deposited near where they enter the oceans (unless transported by turbidity currents). Very fine particles may be transported long distances by currents before they eventually settle on the seafloor.
- Particles can be resuspended after they reach the seafloor if the current speed is sufficiently high.
- Because fine-grained particles are cohesive, a high current speed is necessary to resuspend them from some fine-grained sediments.
- Particles may be alternately deposited and resuspended many times, where current speeds are variable.
- The particle size of grains within deposited sediment is determined by the range of current speeds and the size range of particles transported into the area. Areas with high maximum current speed generally have coarse-grained sediments. Fine-grained sediments are present only where the minimum current speed is low, and the maximum current speed is not extremely high.

UNDERSTANDING THE CONCEPT

Solid particles, or **suspended sediments**, are transported through the water by **currents** in the same way that dust particles are carried through the air by winds. Just as dust particles settle when the air is calm, waterborne particles (suspended sediments) sink to the seafloor when currents are slow or absent.

Particles sink through the water in response to **gravity**. However, particles do not all sink at the same rate, because of **friction** between them and the water molecules they must push aside. Higher-**viscosity** fluids have higher frictional resistance to sinking particles. The viscosity of water is low compared with, for example, molasses or motor oil.

Large, dense particles are not slowed significantly by friction as they fall through the water column. However, viscosity becomes more important as particle size decreases. To understand

why smaller particles are more affected by viscosity, we must consider three factors. First, the gravitational attraction on a particle falling through the water column is directly proportional to its mass (for particles of the same **density**, it would be directly proportional to the volume). Second, because viscous friction occurs at the particle surface, particles with larger surface area are subject to greater friction. Third, the ratio of the surface area of a particle to its volume generally increases as particle size is reduced. The ratio of viscous friction to gravitational force, therefore, increases as particle size is reduced. Consequently, for small particles, such as those in suspended sediment, the particle's settling rate decreases progressively (the particle sinks more slowly) as the particle size decreases.

There are some exceptions to this rule. First, the density of the particle is important. Less dense particles sink more slowly than denser particles. This fact explains how some marine organisms avoid sinking (Chap. 14). Second, the shape of the particle is important. A parachute or feather falls more slowly through the atmosphere than a rock or rice grain because the surface area of a parachute or feather is large in relation to its volume and weight. Thus, air resistance or viscous friction is enhanced. Similarly, some marine organisms have body designs that maximize their surface area to prevent them from sinking (Chap. 14).

Most suspended particles (mineral grains) in the oceans, other than organic **detritus**, are of similar density. Hence, they sink at rates that are determined primarily by their diameter. Table CC4-1 reports typical sinking or settling rates for particles of varying sizes. Sand-sized particles sink rapidly, and smaller particles sink much more slowly. Organic detritus particles are generally of lower density than mineral grains and sink more slowly.

Particle-sinking rates are modified by the presence of currents. As current speed increases, **turbulence** increases and particle sinking slows. This effect can be seen in a glass of orange juice. If the glass sits undisturbed for a while, the particles of orange pulp settle to the bottom of the glass. If the orange juice in the glass is gently stirred to keep the liquid in motion, the pulp will not settle. The effect of currents on particle-sinking velocities

TABLE CC4-1 Settling Velocity of Particles of Average Density and Approximately Spherical Shape When No Current Is Present

	<i>Sand</i>	<i>Silt</i>	<i>Clay</i>
Particle diameter	0.1 mm	0.01 mm	0.001 mm
Settling velocity	2.5 cm•s ⁻¹	0.025 cm•s ⁻¹	0.00025 cm•s ⁻¹
Time to settle 4 km (average ocean depth)	1.8 days	6 months (185 days)	50 years

depends on both current speed and particle size. For a given particle size, sinking rate is reduced as current speed increases until the turbulence is sufficient to prevent the particle from sinking at all. The current speed at which particles no longer sink varies with particle size. This relationship is shown in **Figure CC4-1a**. At high current speeds, only large particles settle to the seafloor, and smaller particles remain in suspension. At lower current speeds, smaller particles also settle to the seafloor.

Once a particle has settled to the seafloor, it can be **resuspended** if the current speed at the seafloor is sufficiently high. **Figure CC4-1b** shows the current speeds needed to resuspend sediment particles of different sizes from sediments that consist primarily of grains of that size range. For all particle sizes, the speed needed to resuspend them is higher than the speed needed to prevent them from sinking (**Fig. CC4-1a,b**). High current speeds are necessary to resuspend large particles, and somewhat lower speeds are needed to prevent them from sinking once resuspended.

As particle size decreases, both the resuspension speed and the speed required to prevent sinking decrease at approximately the same rate until silt- or clay-size particles are reached. The current speed needed to resuspend these smaller particles increases with decreasing particle size, so the speed needed to resuspend the finest sediment particles is higher than that required to resuspend sand-sized sediments. The reason is that fine sediment particles are irregularly shaped and have a very large surface area

relative to their volume (size). The grains are so close to each other that they are held by **electrostatic** attraction between the individual particle surfaces. In addition, they tend to lock together because of their irregular shapes. Organic matter may also help the particles stick to each other. Because they cling together so tightly, fine-grained sediments are considered **cohesive**.

Figure CC4-2 shows how suspended sediments of various grain sizes behave at different current speeds. The settling and resuspension threshold lines on this chart are the same as those in **Figure CC4-1**. Three areas are shown in **Figure CC4-2**: one where conditions are such that a particle would sink to the seafloor; another where conditions are such that particles would remain suspended; and a third where conditions are such that particles would not only remain suspended, but would also be resuspended if they were deposited on the seafloor. Referring to **Figure CC4-2**, consider an ocean area with no local sources of particles, where the current speed ranges from A to B. No sediment of grain size larger than d_2 is brought into this area by currents. Particles of grain diameters between d_3 and d_4 would be deposited, and sediments would consist of these grain sizes. Particles smaller than d_1 would be transported through this area and would not settle.

If the current speed in **Figure CC4-2** ranged from A to D, particles between sizes d_1 and d_5 could enter the area and be deposited, but finer-grained particles would not be deposited. Particles between sizes d_6 and d_7 could be resuspended and removed,

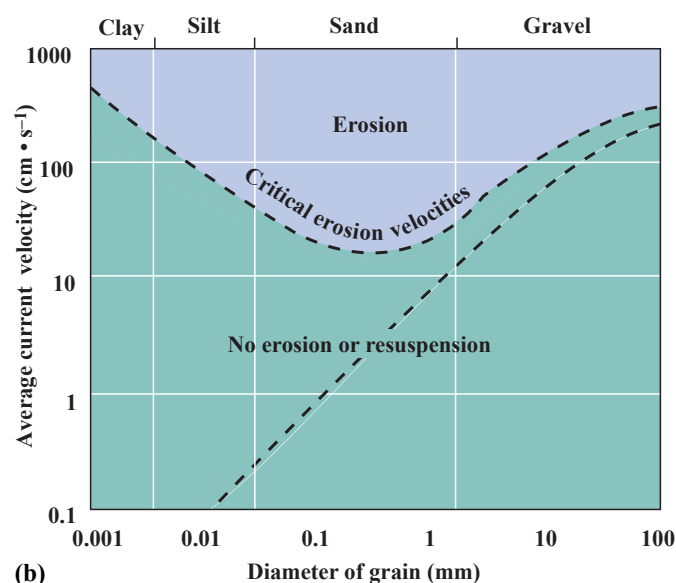
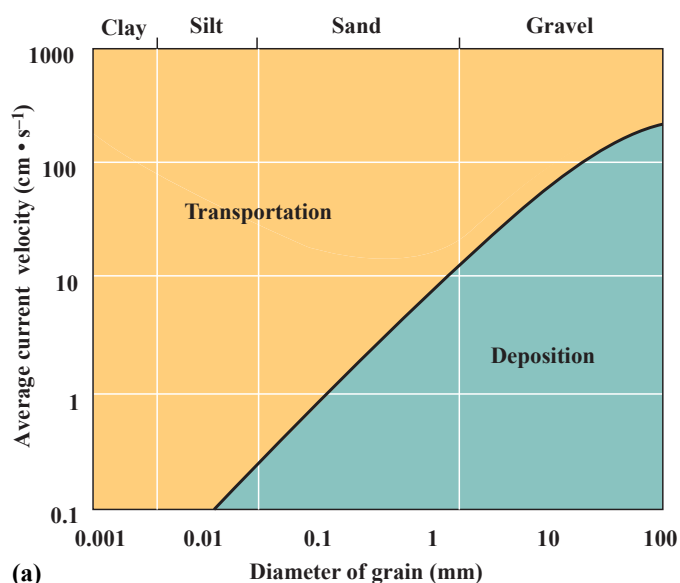


FIGURE CC4-1. Particle settling, resuspension, and current velocity. (a) Settling rate changes with current velocity. (b) Particle size and current velocity affect particle resuspension.

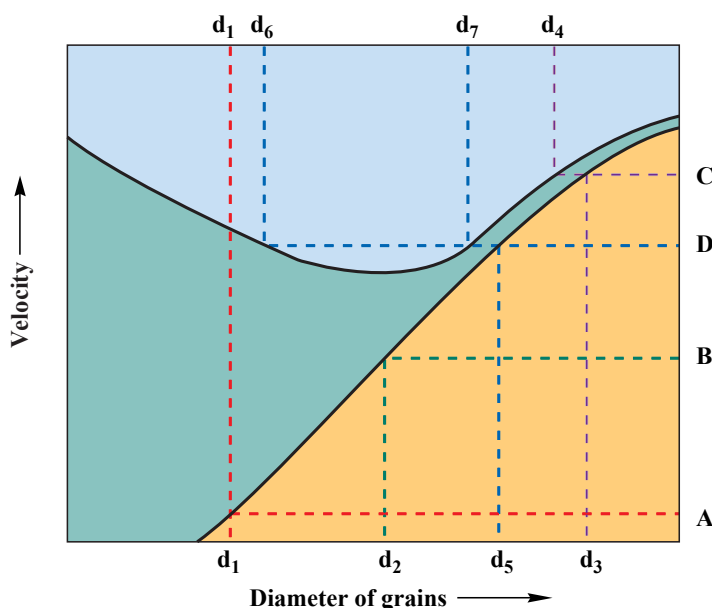


FIGURE CC4-2. Depiction of why most sediments are sorted by grain size, as explained in the text. Note how grain size sorting is determined by the range of wave and/or current speeds at a particular point on the seafloor.

but fine-grained sediments between sizes d_6 and d_7 , once deposited, could not be resuspended. Particles of size range d_4 to d_7 also would accumulate. However, in most locations, fine-grained particles far outnumber larger particles in the suspended sediment. Therefore, at this site the sediments would be fine-grained muds consisting primarily of grains between d_1 and d_6 in size.

From **Figure CC4-2**, we can conclude that large-grained particles cannot be carried far from areas where they are introduced to the oceans unless maximum current speeds are very high (but they could be transported by **turbidity currents**; **Chap. 6**). In contrast, fine-grained particles tend to be carried long distances before settling in an area where the minimum current speed is low, and they can not accumulate in any area where the minimum current speed is high. The diagram further demonstrates that the sediment particles that accumulate in a given location are restricted to a range of sizes determined by the current regime

CRITICAL CONCEPT 5

Transfer and Storage of Heat by Water

ESSENTIAL TO KNOW

- Water has a high heat capacity. More heat is needed to raise the temperature of water by 1°C than is required to do the same for almost any other substance.
- Large amounts of heat can be absorbed and stored in the waters of the mixed layer (approximately the upper 100 m) of the oceans without causing major water temperature changes.
- Heat can be transported with water masses and released at another location by radiation, conduction, and evaporation. This is a mechanism by which heat is transported from the tropics to higher latitudes.
- Water has a high latent heat of fusion. More heat is needed to melt ice than is needed to melt almost any other substance.
- Ice in the polar regions acts as a thermostat because of water's high latent heat of fusion. Large amounts of heat can be lost or gained seasonally by the conversion of ice to water, or vice versa, without changing the temperature of the ocean surface water.
- Water has a high latent heat of vaporization. More heat is needed to vaporize or evaporate water than is needed to vaporize any other substance.
- Water evaporated from the ocean surface contains large quantities of heat as latent heat of vaporization. This heat energy is transported with the air mass and then released to the atmosphere when the water vapor condenses. This is another mechanism by which heat is transported from the tropics to higher latitudes.

UNDERSTANDING THE CONCEPT

Water has a very high **heat capacity** (**Chap. 5, Fig. 5-10**). This means that it takes much more heat energy to increase the temperature of water by 1°C than it does to cause the same temperature change in other substances, such as rocks or air. Consequently, in regions where solar radiation is intense, the surface layer of the oceans can absorb large quantities of heat while the water temperature undergoes little change.

Most of the solar heat absorbed by the oceans is initially absorbed in the upper few meters of water because solar energy is absorbed rapidly and does not penetrate far into the water column. The heat energy is distributed rapidly by **turbulence** throughout the ocean's upper **mixed layer**, stirred primarily by winds (**Chap. 8**). Thus, solar heat is distributed and stored in large volumes of water. In contrast, solar heat reaching the land is transferred downward only slowly by **conduction** (a very inefficient heat transfer mechanism) and is mostly radiated back into the atmosphere and space at longer **wavelengths**.

Because large quantities of heat can be stored in the surface layers of the oceans, heat can be transported with ocean **currents**. As the heated water travels, the heat energy can be released to the atmosphere. Thus, the high heat capacity of water facilitates the transfer of heat from tropical regions to colder regions near the poles (**Chap. 7**). The mild **climate** of parts of Europe in comparison with the climate of eastern North America at similar **latitudes** is partially the result of this mechanism. Heat is transported from the tropics into European seas by the warm Gulf Stream water and then released to the atmosphere (**Chap. 8**).

Water also has a high **latent heat of fusion** and a high **latent**

heat of vaporization. This means that a large amount of heat energy is needed to convert ice to water, or water to water vapor (**Chap. 5, Fig. 5-10**). Heat energy added to ice to form water, or to water to form water vapor, is released when water freezes, or water vapor condenses. Thus, heat can be stored and transferred from one location to another if the conversions between ice and water, or between water and water vapor, occur at different times or places. For example, heat from the ocean surface can be used to vaporize or evaporate water. Water vapor can be transported through the atmosphere for thousands of kilometers until it recondenses and releases its heat energy to the atmosphere. Also, icebergs can be transported to lower latitudes by currents, where they melt as they absorb heat from the sun or surrounding water.

The high latent heat of fusion is important in polar regions. If water had a lower latent heat of fusion, the extent of the Arctic Ocean ice and the Antarctic ice sheet would vary more between summer and winter than they do at present. The presence of ice moderates the climate in these regions. As heat is lost in winter and gained in summer, most of the loss or gain goes to the freezing or melting of ice, and the ocean water temperature remains at the freezing point. Thus, the heat stored in the water by the melting of ice in summer is returned to the atmosphere during the winter as the water refreezes. In this way, the polar ice acts virtually as a thermostat, preventing air temperatures from becoming much higher in summer and much lower in winter.

The high latent heat of vaporization is important globally because it allows large amounts of heat energy to be transferred

from the oceans to the atmosphere. Once in the atmosphere, the heat is redistributed by atmospheric circulation. Water need not boil for water molecules to be transferred from the liquid to the vapor phase (**Chap. 5**). At any temperature, water molecules are transferred from ocean to atmosphere by evaporation at the sea surface, but the rate of evaporation generally increases with increasing water temperature. When a water molecule evaporates from the oceans, it carries with it its latent heat of vaporization, which is more than 1000 times the amount of heat needed to raise the temperature of the same amount of water vapor by 1°C. Once evaporated, the water carries its latent heat through the atmosphere until the air cools, causing water vapor to condense to liquid water (clouds and rain).

When water molecules condense, they release their latent heat of vaporization to the surrounding air molecules. Because the heat capacity of all atmospheric gases is less than that of water, one water molecule can release enough heat to significantly increase the temperature of the air around it. Thus, evaporation and condensation of small amounts of water can transfer large amounts of heat energy from the oceans to the atmosphere. This mechanism is responsible for transporting heat from the equator toward the poles, thus moderating climate differences that otherwise would be extreme with increasing latitude (**Chap. 7**). In addition, this mechanism is responsible for the more moderate climate of coastal areas than of inland areas of the same latitude. It is also the driving energy of air movements in the atmosphere, including storms and **hurricanes** (**Chap. 7**).

CRITICAL CONCEPT 6

Salinity, Temperature, Pressure, and Water Density

ESSENTIAL TO KNOW

- The density of water increases as the salinity increases.
- The density of seawater (salinity greater than 24.7) increases as temperature decreases at all temperatures above 4°C
- The density of seawater increases with increasing pressure. Density changes about 2% because of the pressure difference between the surface and the deep seafloor. The effect of pressure on density usually can be ignored because most applications require density comparisons between water masses at the same depth.
- Water with a salinity less than 24.7 has an anomalous density maximum. Pure water has its maximum density at about 4°C, but the maximum density of water occurs at lower temperatures as salinity increases.
- Between 4°C and the freezing point, the density of pure water decreases as temperature decreases.
- The relative importance of changes in temperature and salinity in determining seawater density varies with water temperature. Temperature variations are more important in warm ocean waters, whereas salinity variations are more important in cold ocean waters.

UNDERSTANDING THE CONCEPT

Below the ocean surface layers, movements of water masses are caused predominantly by differences in density between water masses (**CC1, CC3**). For this reason, water density, which

depends on temperature, the concentration of dissolved salts, and pressure, is among the most important properties of seawater.

Increasing pressure compresses liquids, squeezing more mass into a smaller volume. Thus, density increases with increasing pressure. However, the molecules of water can be forced together only slightly, even by large pressure increases. Water, unlike air, is therefore almost incompressible, and the effect of pressure on water density is small. The density of seawater is only about 2% greater at the deepest depths of the oceans than it is at the surface. Consequently, the effects of pressure on density are generally ignored because oceanographers are usually interested in density differences between water masses at or near the same depth.

The density of water increases as **salinity** increases. This is because most **ions** have a higher density than water molecules do, and dissolved substances reduce the clustering of water molecules.

Decreasing temperature generally causes liquids to contract. Therefore, decreasing temperature generally increases density. However, the behavior of pure water is an exception to this rule. Between its boiling point and 4°C, water behaves normally: density increases as temperature decreases. However, between about 4°C and its freezing point, pure water decreases slightly in density as the temperature decreases. In other words, water density has a maximum at about 4°C (**Chap. 7**).

As salinity increases, the temperature at which the density maximum occurs decreases. At salinities of 24.7 and higher, the

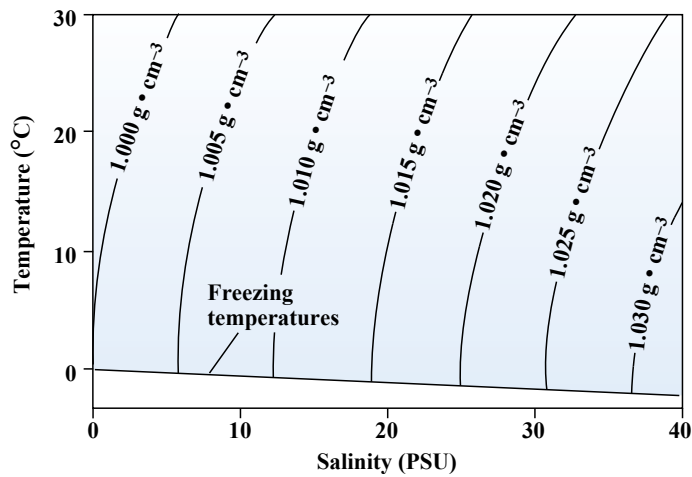


FIGURE CC6-1. The relationships among salinity, temperature, and density are complex. Generally, density increases as temperature decreases or as salinity increases. This figure is reproduced with additional information in **Chapter 5** (**Fig. 5-13a**).

water density maximum is at the freezing point. The salinity of open-ocean waters is generally above 24.7. Therefore, in contrast to pure water and water with salinity less than 24.7, open-ocean seawater increases in density as temperature decreases at all temperatures above its freezing point.

The relationships among seawater density and pressure, temperature, and salinity are complex and are discussed in more detail in **Chapter 5** and **Figure CC6-1**. However, they are illustrated by the following summary:

- Pure water at 4°C and 1 atmosphere (atm) pressure has, by definition, a relative density of exactly 1.
- Density varies almost linearly with salinity. Density increases by about 0.00080 (± 0.00004) for each unit of salinity increase, where salinity is measured in practical salinity units, abbreviated PSU. The density of “average” seawater is about 1.028 at salinity 35, 0°C, and 1 atm pressure, which is about 3% more than that of pure water at 4°C.
- Density varies almost linearly with pressure. Density increases by about 0.00045 for every 100 m of depth in the

oceans.

- The rate of change of density with temperature is a function of both temperature and salinity. Density increases as temperature decreases, as shown in **Table CC6-1**.

Table CC6-1. Density increase with salinity and temperature

Salinity	Temperature (°C)	Density Increase per 1°C Reduction
0	1	−0.00007 ^a
0	10	0.00009
0	20	0.00021
0	30	0.00030
20	1	0.00001
20	10	0.00014
20	20	0.00024
20	30	0.00033
40	1	0.00017
40	10	0.00018
40	20	0.00027
40	30	0.00034

^aThe density of pure water decreases with decreasing temperature below 4°C.

The relative importance of temperature and salinity in determining density varies with water temperature. In cold waters, variations in salinity are more important than variations in temperature. In warm waters, variations in temperature are more important.

The density of seawater in the surface layer is altered continuously by solar heating, which changes temperature, and in turn evaporation, precipitation, and ice formation and melting, which alter salinity. These processes take place at the ocean surface. Vertical motions of water masses in the oceans are caused primarily by the sinking of high-density surface water formed by cooling or evaporation (increased salinity).

The vertical motions resulting from changes in water density are discussed in **CC3** and in **Chapter 8**.

CRITICAL CONCEPT 7

Radioactivity and Age Dating

ESSENTIAL TO KNOW

- Elements usually have several isotopes that have a different number of neutrons and differ in atomic weight but are essentially identical in chemical properties.
- Some individual isotopes of many elements are radioactive. Others are stable.
- Radioisotopes called “parent isotopes” decay by losing part of the atom, becoming a different isotope, called a “daughter isotope.” Often, the daughter is an isotope of a different element.
- Radioactive decay releases heat and radioactive particles, including alpha particles, beta particles, and gamma rays.
- Each radioisotope decays at a certain fixed rate, expressed by the isotope’s half-life.
- One half-life is the time it takes for exactly half of the atoms of a particular radioisotope in a sample to decay.
- The age of a rock, archaeological artifact, or skeletal or undecomposed remains of an organism can often be determined by measuring the concentrations of both the parent and daughter isotopes in the sample.
- For such age dating to be accurate, no daughter isotope must have been present when the sample was formed. In addition, no parent or daughter isotope atoms must have been gained or lost by the sample after it was formed. These conditions are often not met.
- Different parent-daughter isotope pairs must be used to date samples of different ages. Parent isotopes with long half-lives are useful only for dating older samples.

- Radioisotope dating is often used to calibrate other, less expensive dating techniques, such as fossil and stratigraphic dating, that reveal only the relative dates of samples from within a group of samples under study.

UNDERSTANDING THE CONCEPT

Atoms of any element may occur in several different forms called isotopes. Isotopes differ from each other in atomic weight, but they are virtually identical in their chemical properties. Some isotopes are stable, and others are radioactive. **Radioisotopes** (but not stable isotopes) called “parent isotopes” are converted, or decay, by losing part of the atom to become a different isotope (often of a different element), called a “daughter isotope.”

During **radioactive** decay, the radioisotope releases heat and one or more types of radioactive particles, including alpha particles, beta particles, and gamma rays. Radioactive decay takes place at a fixed rate, different for each radioisotope. This rate is expressed by the radioisotope **half-life**. One-half of the atoms of a specific radioisotope in a sample will decay during one half-life for that isotope, regardless of how many atoms were present initially. Thus, after one half-life, one-half of the atoms originally present have decayed. After two half-lives, one-half of these remaining atoms have also decayed, and one-quarter of the original number of atoms of the isotope remain. After three half-lives, one-eighth of the original atoms of the isotope remain; after four half-lives, one-sixteenth of them remain; and so on.

The age of various materials, including rocks, archaeological artifacts, and the remains of living organisms, can be determined by radioisotope dating. To use this technique, the fraction of the atoms of certain radioisotopes that remain undecayed from the time the rock or other sample was formed must be determined. We can do this by measuring the concentrations of both the parent and daughter isotopes in the sample. If the sample is one half-life old, the parent and daughter will be present in equal concentrations. After two half-lives, there will be three times as many daughter atoms as parent atoms. After three half-lives, there will be seven times as many daughter atoms as parent atoms, and so on for other ages.

Radioisotope dating depends on several critical assumptions. We must usually assume that no atoms of the daughter isotope

TABLE CC7-1. Some radioisotopes Used for Age Dating and Their Half-Lives

Parent Isotope	Stable Daughter Isotope	Half-Life ^a (years)
Carbon-14	Nitrogen-14	5560
Uranium-235	Lead-207	700 million
Potassium-40	Argon-40	1.3 billion
Thorium-232	Lead-208	1.4 billion
Uranium-238	Lead-206	4.5 billion

^aDating methods are most accurate when used on samples with ages between 0.5 and 3.5 times the half-life

were present when the sample was formed and that neither parent nor daughter has been added or removed, except by radioactive decay, since then. These assumptions are often incorrect and are always difficult to test because samples undergo physical and chemical changes over time that can add or remove elements. Therefore, to be certain of the measured age, we must often date a sample by two or more different methods.

A particular radioisotope is not useful for age dating if only a very small fraction of the original parent atoms have decayed or, conversely, if almost all the original parent atoms have decayed. Hence, the ideal radioisotope for dating depends on the sample’s age. Several of the most frequently used radioisotopes and their daughters and half-lives are listed in **Table CC7-1**. We can see that carbon-14 is the most useful radioisotope for dating geologically recent samples (if they contain carbon), whereas uranium-235 is more suitable for older materials. The other radioisotopes listed are suitable only for dating ancient samples.

The very sensitive, accurate measurements required for radioisotope dating, along with the equipment, can be expensive. In addition, the determined dates must be verified because the necessary assumptions for this technique are not always valid. For these reasons, radioisotope dating is often supplemented by relative dating techniques based on magnetic properties or **fossils** (**Chap. 6**). In some studies, radioisotope dating is performed on only a few representative samples, and the dates of these samples are used to calculate absolute dates for other samples dated by relative dating methods.

CRITICAL CONCEPT 8

Residence Time

ESSENTIAL TO KNOW

- Substances, including water, move through different parts of the oceans, atmosphere, and biosphere at different rates.
- The concentration of a substance in a part of the ocean, atmosphere, or biosphere is determined by how fast, on average, a molecule of the substance moves from input to output within that particular segment of the environment. This rate is measured as a residence time.
- Residence time is calculated by dividing the total quantity of a substance within an environmental segment (or “box”) by the rate of either its input or its removal from the segment.

- Residence time can be used to estimate the concentration increases that will result from increased discharges of a substance into a segment of the ocean. It is particularly useful for pollution assessments in estuaries and coastal waters. Long residence time and small water volume lead to high concentrations of a contaminating substance.
- In certain situations, we can use residence times to estimate the relative magnitudes of inputs to, or outputs from, a particular ocean segment without measuring them directly.

UNDERSTANDING THE CONCEPT

Marine and atmospheric sciences involve the study of the

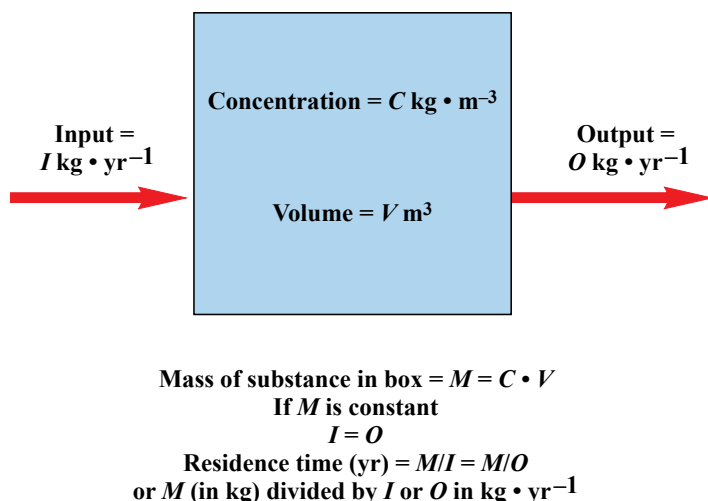


FIGURE CC8-1. This simple model shows how we calculate residence time by dividing the mass of the substance in the environmental segment (“box” or “compartment”) by either the rate of input or the rate of removal (output) of the substance from the box.

movement of elements, water, and other substances within the oceans and atmosphere, and between the oceans, atmosphere, biosphere, and Earth’s mantle and crust. These movements are components of complex **biogeochemical cycles** (Chap. 5). The complexity of chemical movements within these cycles makes it very difficult, if not impossible, to study a cycle as a whole. Accordingly, the **environment** must be divided into segments for most studies.

A segment of the environment is chosen by a characteristic that distinguishes it from adjacent segments in some way. A segment, which, for convenience, is usually called a “box” or “compartment,” can be defined as almost any part of the Earth, ocean, or atmosphere. For example, a box could be defined as the entire world ocean, a single basin such as the Mediterranean Sea, or a simple square section of open ocean that is identified only by boundaries defined by an oceanographer. A box could also be the entire ocean **biomass** or the members of a particular fish species within a specified geographic area. Boxes are often defined to investigate the **residence times** of substances.

The substance of interest enters a box through certain identifiable pathways (inputs) and leaves through other identifiable pathways (outputs). Substances passing through the box can move quickly from input to output, or remain within the box for a long period before exiting. The average time that an atom or molecule of a relevant substance remains in the box is an important characteristic of the system and is expressed as its residence time.

To calculate the residence time in a defined box, one must measure the total quantity of the substance in the box, and either its rate of input to or its rate of output from the box. For example, if we want to determine the residence time of sodium in the Mediterranean Sea, we estimate the total quantity of sodium by measuring the sodium ion concentration at different places, averaging the result, and multiplying by the Mediterranean Sea’s volume. The rate at which sodium is entering or leaving the Mediterranean must also be measured. All of the sea’s inputs or outputs must be considered, including **river runoff**, atmospheric fallout and rain, and exchange of water between the Mediterranean and the Atlantic Ocean through the Straits of Gibraltar, between the Mediterranean and the Black Sea through the Bosphorus, and between the Mediterranean and the Red Sea through the

Suez Canal. We could determine the residence time of the water itself by measuring the total quantity of water and its inputs and outputs from the Mediterranean in the same way, but in this case, we would have to consider an additional output: evaporation.

Once we know the total amount of the substance of interest in the box and the rate of its input or output, we can calculate a residence time if we assume that the system we are studying is in a **steady state** (in other words, if we assume that the amount of substance in the box remains stable over time). Because the Earth’s biogeochemical cycles have had billions of years to reach equilibrium, we can usually accept that assumption when we are studying large boxes, such as an entire ocean basin. However, the smaller the box, the less likely the assumption is to be true. Although the Earth’s biogeochemical cycles are in approximate steady state when averaged globally, these processes can vary substantially at any one place and time because of such changes as variations in the year-to-year rate of river flow.

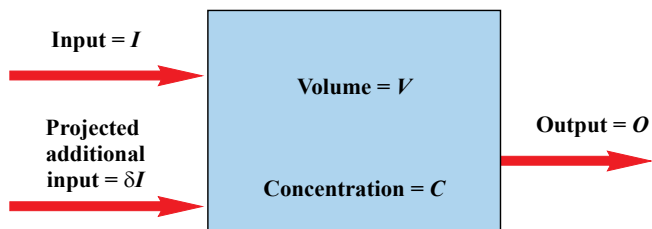
If the total amount of substance in the box remains stable over time, the rate of input must equal the rate of output (Fig. CC8-1). The residence time is calculated as the total quantity of the substance in the box divided by either the output or input rate. Measured in years or another time unit, residence time is the time it would take for the outputs to remove all the substance from the box if all inputs were stopped, and vice versa.

The concept of residence time has many uses. First, it provides a method of estimating the rates of some processes that are difficult to measure directly. For example, if we can measure the concentration of a particular element in Mediterranean seawater and the rate of input of this element from rivers and adjacent seas, we can calculate the rate at which this element is being removed to the **sediment**. We did not include **sedimentation** (or evaporation and salt precipitation) in our preceding examination of sodium cycling in the Mediterranean because, for sodium (as opposed to other elements), the rate of removal to the sediment is extremely slow at present, so it is negligible in relation to other inputs and outputs.

A second use of residence time is to determine the probable fate of **contaminants** released to the oceans. If we release contaminants with long residence times, they will remain in the oceans for a long time and may cause **pollution**. If contaminants have short residence times, human additions to the oceans will be quickly removed (usually to the ocean sediment). Residence times of the elements in the world oceans are discussed in **Chapter 5**.

One important characteristic of biogeochemical cycles is that mechanisms for the removal of a substance from a box are generally related to the concentration of the substance in the box. If the concentration increases, the output rate will also increase. For example, if the concentration of a substance in the Mediterranean is increased, the rate of loss of the substance from the Mediterranean will increase as seawater flows out of this sea to an adjacent sea. Similarly, higher concentrations lead to faster uptake by **biota** and faster removal to sediment, although there is often not a simple linear relationship in these instances.

An increase in the input rate to a box generally raises the concentration in the box until the output rate matches the elevated input rate. If the expected increase in input is known, the resulting increase in concentration can sometimes be calculated and predicted. This method is extremely important in determining the potential of rivers, bays, or segments of the coastal ocean to be



Residence time in the “box” $T = \frac{V \cdot C}{I}$

or $C = \frac{T \cdot I}{V}$ (Equation 1)

Residence time remains the same when input is increased.

Residence time $T = \frac{V \cdot C}{I} = \frac{V (C + \delta C)}{(I + \delta I)}$

Where δC is the increase in concentration that occurs in the box when the input is increased by δI .

Thus, $\frac{T}{V} \cdot (I + \delta I) = C + \delta C$

or $\frac{T \cdot I}{V} + \frac{T \cdot \delta I}{V} = C + \delta C$

substituting from equation 1

$C + \frac{T \cdot \delta I}{V} = C + \delta C$

or $\frac{T \cdot \delta I}{V} = \delta C$

or the incremental increase in concentration in the “box” is equal to the residence time multiplied by the incremental increase in input rate, divided by the volume of the “box”.

FIGURE CC8-2. A calculation, which is based on the residence time of a substance in an estuary or similar water mass, can be used to predict the effects of a new source of contamination that increases the input rate of the substance to a defined segment of the environment, or “box.”

affected by contaminant inputs. A hypothetical example illustrates how this assessment can be made. **Figure CC8-2** shows the equations we could use for an estuary into which is discharged river water that carries a natural concentration (C) of a substance. An industrial complex is proposed that will discharge additional

amounts of this substance to the estuary. For simplicity, we will assume that the substance of interest is not removed to sediment within the estuary and that it can leave the estuary only dissolved in seawater. From the calculations in **Figure CC8-2**, we see that the change in concentration (δC) that can be expected to occur in the estuary after the additional input (δI) begins is easily estimated if the residence time, volume of the estuary, and magnitude of δI are known. The important observation to be made about this result is that the concentration increase caused by an increased input of a contaminant substance is greater if the residence time (T) in the discharge water body is long in relation to its volume (V).

This finding is important because some water bodies, particularly estuaries, may have long residence times and small volumes. Such locations are not good choices for waste discharge. San Francisco Bay is a particularly good example. It has two segments: South San Francisco Bay, which has a water residence time of weeks to months; and North Bay, which has a similar volume but a water residence time of only a day or two. Inputs of treated municipal wastewater to South San Francisco Bay have caused serious water quality problems, whereas much larger inputs of treated municipal wastewater and industrial wastewater to North Bay have caused comparatively minor problems. The difference is the longer residence time in South San Francisco Bay.

Residence time calculations can be more complicated when there are multiple, separated inputs or outputs in the area of interest. For example, the calculation in **Figure CC8-2** is more complicated if a fraction of the contaminant is removed to sediments within the estuary. Nevertheless, residence time observations and calculations can be useful tools in such situations. If we can measure only the inputs and the contaminant concentrations in the estuary shown in **Figure CC8-2**, the calculated residence time will not tell us how much of the substance is flushed out to the ocean or how much is removed to the estuary sediment. We need not perform the relevant calculations here, but if we measure the inputs of the substance, its concentrations, and the residence time of water (not the residence time of the substance) in the estuary, we can calculate how much of the substance is removed with the water to the ocean and how much is retained in the estuary sediments and/or biota.

CRITICAL CONCEPT 9

The Global Greenhouse Effect

ESSENTIAL TO KNOW

- The average temperature of the Earth’s atmosphere is determined by the balance between solar energy that penetrates the Earth’s atmosphere and is absorbed by gases, liquids, and solids, and energy that is lost to space by the Earth’s radiation.
- The sun’s radiated energy is concentrated in the visible portion of the electromagnetic spectrum, whereas the Earth’s radiated energy is concentrated at longer wavelengths in the infrared portion of the spectrum.
- The Earth’s atmosphere absorbs energy in the infrared and ultraviolet portions of the spectrum more effectively than it does in the visible portion of the spectrum.
- Because the Earth’s atmosphere absorbs incoming solar radiation less effectively than it absorbs energy reradiated to space by the Earth, the Earth’s climate is warmer than it would otherwise be. This is the greenhouse effect.
- The efficiency of absorption of infrared energy by the atmosphere is increased by higher concentrations of certain gases, including carbon dioxide, methane, chlorofluorocarbons, nitrogen oxides, and ozone, which are called “greenhouse gases.”
- Concentrations of greenhouse gases in the atmosphere have increased and continue to increase as a result of human activities. Carbon dioxide concentrations have risen more than 50% since 1800.
- Radiation of heat from the Earth to space has probably been reduced and will continue to be reduced as a result of the absorption of radiated infrared energy by the

increased greenhouse gas concentrations. In contrast, the amount of solar radiation passing through the Earth's atmosphere has not been reduced by such absorption. If this were the only change occurring, the average temperature of the Earth's atmosphere would increase. This is the global climate change that occurs as a result of the enhancement of the greenhouse effect by greenhouse gas releases.

- The Earth's average atmospheric temperature has increased by between 1.1 and 1.3°C since 1880 (the industrial era).
- The increase in greenhouse gas concentrations in the atmosphere and the consequent reduction in the Earth's radiative energy loss to space have many complex effects on the atmosphere, oceans, and biological systems. Many of these changes also have consequences that, on their own, would lead to warming or cooling of the Earth's atmosphere. These positive and negative feedbacks would, respectively, enhance or offset predicted global warming.
- Most analyses and models of this system predict that the net effect of greenhouse gas emissions will be a global warming of an additional 2°C to 6°C in the next century, but the magnitude of such warming is not certain. Indeed, the ultimate net climate change may be global or regional cooling. The only certainty is that the increases in greenhouse gases will cause global climate changes of currently unknown magnitude.
- The consequences of global warming would be damaging to human civilization. Possible changes include flooding of coastal cities due to rising sea level and drastic changes in rainfall patterns that would lead to the loss of agricultural land.

UNDERSTANDING THE CONCEPT

We experience temporal and spatial fluctuations of temperature in the Earth's atmosphere as weather and seasonal change. However, the average temperature of the Earth's atmosphere remains almost constant from year to year. Changes in average temperature of several degrees have occurred in the geological past (Fig. CC9-1). These changes, although generally small, profoundly affected the Earth's **climate**. For example, they caused **ice ages** to begin and end.

Changes in the chemistry of the Earth's atmosphere due to human activity have already caused about a 1°C increase in the Earth's average temperature and may cause it to increase by as much as 1°C to 3°C more within the next decade or so. This predicted change is usually called the greenhouse effect or “global warming.” The fear is that the Earth's average temperature could change faster than it has ever changed during human history, and that the extent of the predicted change could be great enough to have major effects on the Earth's climate. Rapid climate changes would have numerous damaging effects on nature and on humans. Consequently, there is considerable interest in determining how much of the post-industrial atmospheric temperature increase is caused by anthropogenic greenhouse gas emissions, if the increase is likely to continue and how fast, and whether the problem can be reduced by altering human activities.

The only significant source of heat energy to the atmosphere is solar radiation. Several other sources, including heat **conducted** from the Earth's interior, tidal friction, and heat released by burning **fossil fuels**, are all negligibly small in comparison

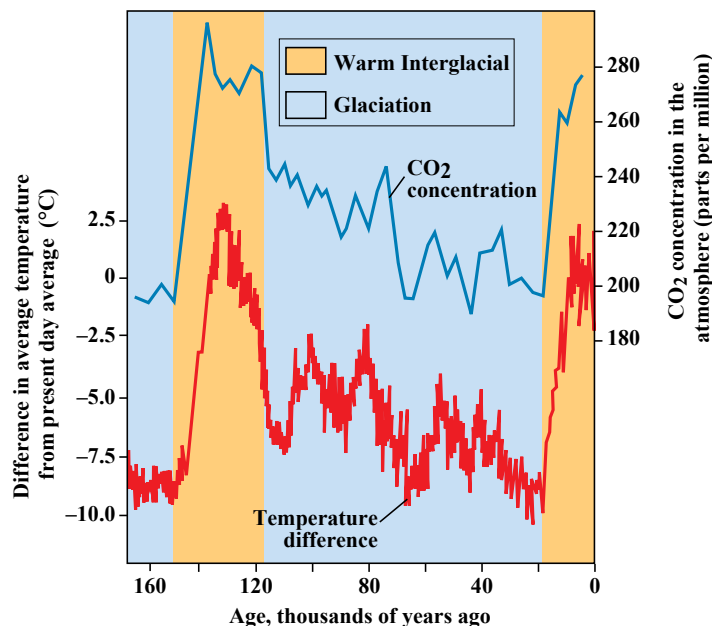


FIGURE CC9-1. Diagram showing the close relationship between variations in average air temperature and atmospheric carbon dioxide concentration during the past 160,000 years. These data were obtained from an ice core drilled into the Antarctic ice sheet. Notice that the Earth's average temperature at present (in geological time), is approximately at its highest level since more than 120,000 years ago. The diagram would show the post-industrial temperature change of about 1°C as only a thin line at the edge of the plot, and if the 50% CO₂ concentration increase were shown, it would plot above this figure—off the top of this page.

with solar input. Part of the solar radiation reaching the Earth is reflected directly back to space (primarily by clouds or snow). The rest is absorbed by the land, oceans, or atmosphere. Heat is transferred between oceans and atmosphere and between land and atmosphere, and it is radiated back to space by land, ocean, and atmosphere (Fig. 7-6). This system of energy transfer is discussed in **Chapter 7**. One important feature is that the amount of solar energy that reaches the Earth must equal the amount of energy reflected or radiated back to space. If the amount of incoming energy does not equal the amount of outgoing energy, the Earth's atmosphere will either gain or lose heat continuously, and the atmospheric temperature will increase or decrease.

To understand how the chemistry of the atmosphere can affect the balance between incoming and outgoing radiation, we must understand the nature of **electromagnetic radiation** (Fig. 5-17). All objects radiate electromagnetic energy, but the **wavelength** range emitted by an object depends on its temperature. The hotter the object, the shorter the wavelengths of electromagnetic energy that it radiates. The sun, which is very hot, radiates much of its energy in the visible and ultraviolet wavelengths, whereas the Earth and oceans, which are much cooler, radiate energy in the much longer wavelengths of the infrared region of the spectrum.

We can understand this relationship of radiation wavelength to temperature from our everyday experience. If you turn the burner of an electric stove on at full power and hold your hand close to it, you quickly feel the heat that it radiates. You see no immediate change in the metal of the burner, because the burner is radiating all of its energy in long wavelengths in the (infrared) part of the electromagnetic spectrum, not visible to humans. As the burner heats further, it begins to glow red. At this higher temperature, it is radiating energy at wavelengths in the visible

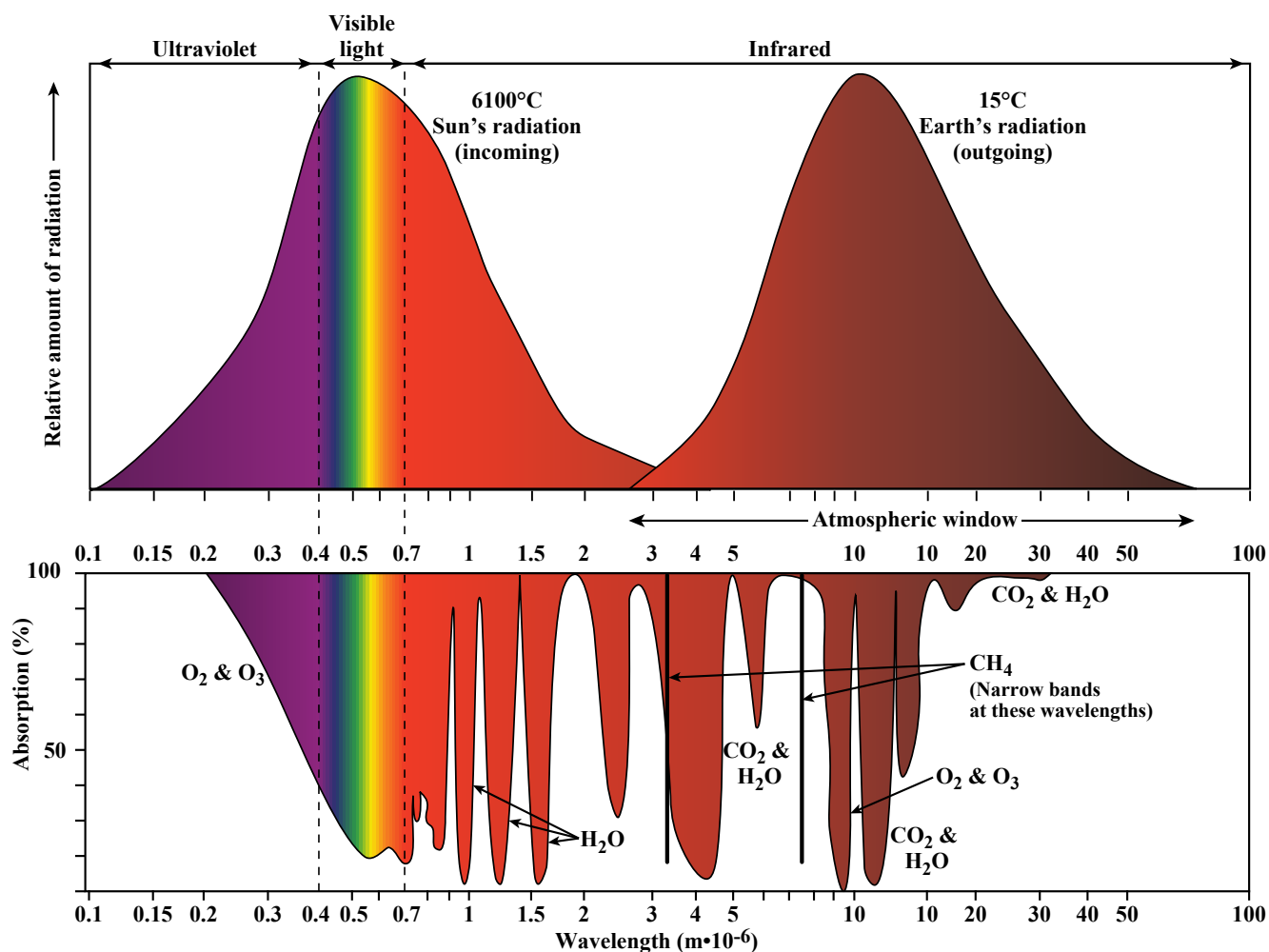


FIGURE CC9-2. The upper part of this figure shows the emission spectra of the Sun and the Earth. The lower part shows the absorption spectrum of the Earth's atmosphere. The sun emits most of its energy in the visible, ultraviolet, and near-infrared parts of the spectrum, whereas the much cooler Earth emits radiation at much longer, infrared wavelengths. The amount of energy that enters the atmosphere from the sun equals the amount lost from the Earth to space by radiation. The sun's radiation is partially absorbed by the Earth's atmosphere, particularly in the ultraviolet region. The Earth's radiation is also partially absorbed by the atmosphere, but at the much longer infrared wavelengths. Carbon dioxide and water vapor strongly absorb radiation at these longer wavelengths, but they do not absorb strongly at most wavelengths of the incoming solar radiation. Other greenhouse gases also absorb radiation primarily in the far-infrared wavelengths.

red and near-infrared part of the spectrum. If you could heat the burner even further, it would radiate at progressively shorter wavelengths until it appeared white, like the sun or the filament of an incandescent lightbulb.

Electromagnetic radiation from the sun must pass through the atmosphere before reaching the Earth's surface. Similarly, longer-wavelength radiation emitted by the Earth, ocean, and clouds must pass through the atmosphere as it is radiated to space. You can feel the radiated heat by placing your hand near a rock or paved surface, especially immediately after dark in summer. As the longer wavelength radiation from the Earth passes through the atmosphere, it is partially absorbed by atmospheric gases, but these gases do not absorb all wavelengths of electromagnetic radiation equally. Each gas absorbs some wavelengths more effectively than others.

The **absorption spectrum** of the atmosphere and the gases that contribute the strongest absorption in parts of that spectrum are shown in **Figure CC9-2**. Compare this absorption spectrum with the **emission spectra** of the sun and the Earth (**Fig. CC9-2**). We see that the sun's radiated energy is strongly absorbed by the atmosphere at ultraviolet wavelengths, but there is little absorption in the visible region of the spectrum where the sun's radiant

energy is concentrated. In contrast, there are strong absorption bands in the infrared portion of the spectrum in which most of the Earth's radiation is emitted.

The atmosphere does not absorb all radiation that passes through it, even at wavelengths where the atmosphere is shown to have an absorption band (**Fig. CC9-2**). Only a fraction of the energy at such a wavelength is absorbed as it passes through the atmosphere. The rest is transmitted either to the Earth's surface or to space. The fraction absorbed is a function of the number of absorbing molecules of the gas that lie in the radiation's path as it passes through the atmosphere. If the atmospheric concentration of a gas or gases that absorb at a particular wavelength is increased, the fraction of the energy absorbed at this wavelength will increase. Consequently, the fraction transmitted to the Earth's surface or to space is decreased.

Carbon dioxide does not absorb strongly in most of the visible wavelengths that correspond to the sun's energy spectrum (**Fig. CC9-2**). However, it does absorb strongly throughout almost the entire wavelength range of the Earth's radiation. If the carbon dioxide concentration in the atmosphere is increased, there will be little effect on the quantity of solar energy passing through the atmosphere to reach the Earth's surface. However, the quantity of

energy that is radiated by the Earth and atmosphere and transmitted to space will decrease. Because less energy will now escape to space, but the same amount will be received from the sun, excess energy will be retained in the atmosphere, and the global climate will warm. This process is known as the “greenhouse effect.” The name greenhouse effect refers to the similarity of this phenomenon to the effects of the glass of a greenhouse window. Glass transmits visible light but absorbs or reflects infrared energy radiated by objects within the greenhouse, keeping the greenhouse warmer than the surrounding air.

What we normally mean when we say the greenhouse effect is actually incorrect. Anthropogenic addition of greenhouse gases should be called the “enhanced greenhouse effect” because carbon dioxide and other gases that occur naturally in the atmosphere already cause a natural greenhouse effect. The temperature at the Earth’s surface would be much lower than it is now if there were no natural greenhouse effect.

We focus on carbon dioxide as a greenhouse gas because its concentration in the atmosphere has increased more than 50% since 1800, and is now increasing at a faster rate each year (Fig. 1-2). This increase is the result of burning fossil fuels and deforestation (live vegetation absorbs atmospheric carbon dioxide and releases oxygen). Several other greenhouse gases may also be important, including methane, **chlorofluorocarbons** (freon and similar compounds, often called “CFCs”), nitrogen oxides, and **ozone**. Each of these gases preferentially absorbs the Earth’s long-wavelength radiation more than the sun’s shorter-wavelength radiation. In addition, each of these gases is released to the atmosphere in substantial quantities as a result of various human activities, including fossil fuel burning, industry, and farming. Although the quantities of carbon dioxide released to the atmosphere are much larger than the quantities of these other gases released, the other gases absorb long-wavelength radiation more effectively. Therefore, global climate change studies must also take into account these other gases.

We know that the carbon dioxide concentration in the atmosphere has increased in the past two centuries. The concentrations of other greenhouse gases have undoubtedly also increased, although we have only limited information about how much. Because more of the Earth’s radiated heat energy is trapped and prevented from escaping to space while the amount of solar radiation reaching the Earth’s surface has not been substantially changed, the Earth’s atmosphere should be heating up. This is why the expected climate change has been called the “global warming problem.” However, the atmospheric heat balance is more complicated than the simple input and output of radiated energy. Many other processes affect the balance between heat absorption and heat loss by the atmosphere, and most of these processes are, in turn, affected by changes in atmospheric temperature or chemistry.

These secondary effects can either add to the warming that would be expected from the greenhouse effect or act in the opposite direction and cause a tendency for the Earth’s atmosphere to cool. Hence, they provide what are known as positive or negative **feedbacks** (often called “feedback loops”) to the primary greenhouse effect. Positive feedbacks would cause faster atmospheric warming than would occur if just the simple enhanced greenhouse effect were operating. Negative feedbacks would cause slower atmospheric warming. If negative feedbacks were great enough, the net effect of the release of greenhouse gases could

actually be a colder climate.

Most models of the complicated atmospheric heat balance predict that negative feedbacks will not dominate, and the net result of the enhanced greenhouse effect will be global warming. This conclusion is supported by the very good correlation in the Earth’s past between high atmospheric carbon dioxide concentrations and high average climatic temperatures (Fig. CC9-1). However, there is considerable scientific uncertainty about the many feedback mechanisms, their magnitude, and how quickly they will respond to the extremely rapid (in geological time) increase of carbon dioxide concentrations in the atmosphere. Although most scientists believe that the release of greenhouse gases will cause global warming (estimated by models to be an increase of an additional 1°C to 6°C during the next century), this outcome is not certain. What is certain is that the release of greenhouse gases has altered the atmospheric heat balance and that the alterations will alter the Earth’s climate. For this reason, scientists now refer to the issue not as the “greenhouse effect” or “global warming,” but instead as “global climate change.”

To see the complexity of the feedback mechanisms involving ocean-atmosphere interactions, we can briefly examine two such mechanisms, neither of which is currently well understood. First, consider a positive feedback. The oceans contribute the greenhouse gas methane to the atmosphere (Chap. 5). The methane is primarily a by-product of the biological activity in the oceans. The rate of methane production and its release to the atmosphere could increase if biological productivity in the oceans increased. About one-quarter of all the carbon dioxide released by human fossil fuel burning has now become dissolved in the oceans. Higher dissolved carbon dioxide concentrations may result in higher rates of **photosynthetic** growth and production in the oceans, which may lead to an increase in methane release from the oceans. Because methane is a greenhouse gas, it would add to the greenhouse effect, trapping more heat.

Second, consider a negative feedback. Clouds reflect a fraction of the sunlight that hits their upper surfaces back to space. The extent of the Earth’s cloud cover is determined largely by the rate of evaporation of water from the ocean surface. The primary effect of greenhouse gases is to increase atmospheric and thus ocean surface temperatures. Higher water temperatures would lead to increased evaporation. Consequently, increasing global greenhouse gases may cause the extent of cloud cover to increase, and more solar radiation may be reflected to space to offset the additional heat trapped in the atmosphere by the greenhouse gases. This negative feedback is a particularly good example of the complexity of the climate change question, because additional clouds will not necessarily result in a greater reflection of solar heat to space. Certain types of clouds at different levels in the atmosphere are better reflectors than others. In addition, clouds not only reflect but also absorb incoming solar energy, and they both absorb and reflect the radiation outgoing from the Earth. Each of these factors may vary continuously as the rate and location of evaporation at the ocean surface and the ocean and atmospheric circulation change continuously in response to each other.

Of the many positive and negative feedbacks, some may respond almost immediately to greenhouse gas concentration changes in the atmosphere, and others may take centuries or millennia to respond to such changes. A number of the more important feedbacks involve interactions between the atmosphere and the oceans, and they are subject to intensive research because

they must be understood if we are to successfully model and predict global climate change.

None of the mathematical models that were used up to about 1991 were able to include any of the relevant interactions between the atmosphere and oceans. The reasons were that too little was known about these interactions to include them in a meaningful way, and even the supercomputers that were available would have been overwhelmed by the additional computations needed to include them. As computational power and knowledge of ocean-atmosphere interactions continue to improve, and as more information about these feedbacks is included in the global climate models, the models may dramatically change predictions of the climate changes that we can expect as a result of the enhanced greenhouse effect. Nevertheless, it will be many years, or even decades, before these complex processes will be understood well enough that we can expect there to be a high degree of probability in model-based predictions of the future of Earth's climate.

Fortunately, evidence is strong that the presence of many feedbacks within a complex system tends to produce stable situations that are difficult to change. The feedbacks generally tend to return the system toward a stable configuration. However, most of the feedbacks are nonlinear responses (in other words, the response does not vary in exact proportion to the change in the stimulant). Hence, the ocean-atmosphere is a **chaotic** system (CC11).

One important characteristic of chaotic systems is that they

tend to remain relatively stable when disturbed, until the disturbance passes a critical point at which the system suddenly shifts to a completely different, but also relatively stable, state. Possibly the beginnings and endings of ice ages are driven by such chaotic dynamics. Could it be that the enhancement of the greenhouse effect will not significantly affect the Earth's climate until a critical point is reached at which the climate will change abruptly to something as different as an ice age? Only decades of research or time can provide an answer.

The consequences of global climate change caused by the release of greenhouse gases are mostly unknown and difficult to predict. However, they are likely to be severe. For example, one effect will be the thermal expansion of ocean water as its temperature increases. The resultant sea-level rise will likely inundate many coastal cities in low-lying areas, such as Florida and some island nations. Partial melting of the polar ice sheets is likely to also contribute to sea-level rise. In addition, rainfall patterns might be changed. The bands of rainfall that now sustain the breadbaskets that feed much of the global population (**Chap. 7**) are likely to move northward, causing these regions to become deserts like those of North Africa. Because **hurricanes** are fueled by the heat energy of ocean water, another effect of a warmer atmosphere and ocean might be an increase in the intensity of hurricanes. Hurricanes might also be able to sustain damaging winds while traveling farther away from the equator than they do at present.

CRITICAL CONCEPT 10

Modeling

ESSENTIAL TO KNOW

- To understand environmental processes, we generally must create a conceptual or mathematical model of the system (or parts of the system) that describes the components of the system and their interactions. The parts are often called “cells,” “boxes,” or “compartments.”
- There are many types of interaction. Examples include heat transfer between the ocean and the atmosphere, and the effect of changes in food supply on the population of an animal species.
- Conceptualized models must be parameterized and tested. Parameterization includes applying data to each model cell and developing equations to describe the flow of the modeled parameter between each adjacent pair of cells. These equations must be based on the physical laws that apply to each interaction. Testing consists of running the model to see how it fits the measured data. Parameters are adjusted between tests until the model fits the data set.
- Models must be validated by being applied to a data set that was not used in testing the model. This is often difficult to do for models for which the environmental data sets are sparse or limited to short time scales.
- Sensitivity analysis is performed by changing model inputs or assumptions and observing the changes in output. This type of analysis can help identify the parameters or processes that most affect the system's behavior and, thus, are important for field studies or monitoring.

- To facilitate model creation, interactions are often assumed to be linear. Interactions are linear if one component of a system varies in direct proportion to a change in the other. However, most, if not all, natural systems include many nonlinear interactions.
- The nonlinear nature of natural systems limits the degree to which models can be used to make accurate predictions of future changes in the modeled system. Models of systems that are nonlinear in nature can be used only to predict a range of likely future conditions, and these predictions become more uncertain with increasing time into the future.

UNDERSTANDING THE CONCEPT

The Earth, oceans, atmosphere, and the **biota** that inhabit these **environments** are all intimately linked in a system governed by physical laws. Each atom interacts with many others according to these laws, and it is the sum of these interactions and the atoms themselves that constitute the system. No matter how much information we obtain or how many computers we construct and program, describing this system accurately in all its details is impossible. Therefore, we must create a highly simplified system, or model, that replicates the real system in as much detail and as accurately as possible. We can create this model system either conceptually in our minds or mathematically by listing the components of the system and the mathematical relationships between each connected pair of these components.

Models that represent the entire Earth can be constructed at only a very gross level of detail. Hence, we generally construct models of parts of the system (subsystems) that we believe to be somewhat distinct from the larger system. We then represent the rest of the system by inputs and outputs to the subsystem. The descriptions of **plate tectonics**, atmospheric and ocean circulation, and marine biology, as well as other discussions in this textbook, are conceptual models of the real world.

Many different types of interactions must be described to construct models of the environment. Each model must take into account different interactions depending on which subsystem is being modeled. For example, models of atmospheric circulation must include a representation of the processes of heat transfer between ocean and atmosphere, land surface and atmosphere, and sea ice and atmosphere, and a representation of the internal **convective** processes (**CC3**) due to differential heating of air masses at different locations. Similarly, models of the changes of fish populations over time must represent the effects of varying food supply and varying abundance of predatory **species** on the growth, survival, and reproduction rates of the species involved. Models, including the two just mentioned, can be simple or extremely complicated, especially if they represent many different processes within a subsystem.

One example of a highly simplified model, and of how it can be useful, is the model describing the possible effects of changing the rate of input of a **contaminant** to an **estuary** or other body of water that is illustrated in **Figures CC8-1 and CC8-2**. At the other extreme, in coupled ocean-atmosphere global **climate** models, which are used to investigate, among other things, the likely consequences of the greenhouse effect, the ocean is typically segmented into “cells” that are formed by division of the ocean depths into 20 to 60 horizontal layers, or slices, and subdivision of those slices into segments of 2° to 4° of **latitude** and **longitude**. The atmosphere is typically represented by up to 90 horizontal levels subdivided at 1° to 5° intervals of latitude and longitude. The result is models that can have several million to tens of millions of adjoining cells. For each interface between two cubic cells (6 per cell), the model must have equations, based either on observed data or on physical laws, that reflect the rate of transport of water, heat, water vapor, and other parameters between the two cells as the properties of the water and air in each cell change.

These massive global climate models can be run on only the fastest computers, which operate at more than a exaflop (a exaflop is a measure of computer speed and is equivalent to a billion billion or 10^{18} floating point operations per second). Even with these fast computers, running a model through as few as 100 years of climate can take days to months of computer time, and the output data from such a model run can consist of tens to hundreds of terabytes. A terabyte is 1024 gigabytes, or 1,099,511,627,776 (more than a million million) bytes.

The development and application of models involves three fundamental steps, each of which must be revisited and modified as needed when new information is available. The first step is to conceptualize the system under study as a series of separate but interacting cells (often called “compartments” or “boxes”). Most often, these cells are **spatial** areas, but they can be, for example, individual species within an **ecosystem**. The diagrams of the phosphorus and nitrogen cycles in **Figures 12-6 and 12-7** are models in which the cells are both spatial (positioned within the

atmosphere, mixed layer, or deep layer) and representative of the different chemical forms (e.g., dissolved ammonia, nitrate, nitrite, dissolved organic nitrogen, nitrogen in animal tissues, nitrogen in **bacteria** and **detritus**). **Figures 12-10, 12-11, and 12-12** are models in which the cells are individual species in a food chain or food web.

In developing a conceptual model, it is important to know exactly what question is being asked, so that the model can be as simple as possible. For example, if we were investigating the fate of **anthropogenic** nitrogen compounds released to the ocean, the model in **Figure 12-7** would be adequate if we were interested only in how much of the nitrogen was transported to the deep layer and how fast. However, a much more complicated (and computer-intensive) model would be needed if we were interested in how the nitrogen released from North America was distributed to each ocean basin.

The second step in the modeling process is parameterization. In this process, values of the modeled parameter (e.g., nitrogen concentration, **biomass**, or heat content) must be assigned to each cell, and equations must be written to express the flows of the parameter between each pair of adjacent cells. These flow equations can be complex if they are based on physical principles. For example, heat transfer between two ocean water cells is a combination of heat **conduction**, **diffusion**, and mixing across the cell boundary, in addition to water transport between the cells. Usually, a single equation is written for such flows on the basis of observed data, as the processes involved are too complex to express as physical laws. Often, the observed data are very limited or available only for a small number of cells, particularly in complex models such as global climate models. After parameterization, the model is tested by being run with the parameters identified to see if it reproduces the data. For example, the global climate model would be examined for how closely it reproduces changes in temperature data at locations for which real-world data are available. Next, the parameters of the model are adjusted to improve the match to the real-world data. The model may be tested and parameters adjusted in this way many times.

The third step in model development is validation to determine whether the model actually does represent the way the natural system functions. In this step, the model must be tested with a data set that is completely different from the data used for the parameterization step. This is important because the process of parameterization forces the model to fit the input data. Thus, the model parameters may represent only the input set of conditions rather than the underlying processes involved.

Validating some models is relatively easy. For example, global climate models can be parameterized with decades of climate data and then validated by the use of data from a completely different time period. This is possible because relatively detailed historical atmospheric and ocean climate data are available for a number of decades in the past. Validating other models, especially complex ecological models, is much more difficult because the amount and historical coverage of the existing data are often very limited, and parameterization requires using most or all of the available data. Models that cannot be properly validated may still be useful for many research studies, especially sensitivity analysis, which is described in the next paragraph, but they are inherently unreliable if used to extrapolate beyond the existing data set, and they cannot be relied upon for prediction.

The most familiar use of models is to predict future condi-

tions, particularly **weather** and climate. However, the most important use of models is to determine the parameters or processes that most affect a system's behavior and, thus, are the most important for field studies or monitoring. To do this, we subject the models to sensitivity analysis. Sensitivity analysis is a simple, albeit sometimes tedious, process of repeatedly running the model, each time making small changes in a different input parameter or combination of input parameters (the initial value in a cell, or the coefficients of the equations relating flow between cells). This analysis can reveal which input parameters produce substantial changes in the model outputs and which produce only small changes. The former are sensitive parameters that must be accurately determined (or formulated for a flow equation) because small errors will greatly affect the accuracy of the model's representation of the real world. They are also probably the most important parameters in controlling change in the real-world system that is being modeled. As a result, the most sensitive parameters of a model are often the most important parameters to focus on in field research and monitoring programs.

Models that have been adequately tested and validated can be used as predictive tools. The best-known examples of such uses of models are the weather forecasts seen nightly on television, the predictions of the future paths of **hurricanes**, long-term (months) climate predictions that are of particular importance to farmers, and predictions of future climate change associated with the past and future releases of greenhouse gases. These predictions are all based on weather and climate models that have been parameterized, tested, and validated with extensive historical data. However, as we all know, the predictions are always somewhat general in nature and are often inaccurate.

For example, as hurricanes travel across the Caribbean, several days away from the U.S., the model predictions of the storm's path are shown as a wedge, with the future position more and more uncertain the further out in time the prediction is. The pre-

dicted range of possible landfall times and locations of the storm is narrowed as the storm approaches, but the predicted impact point often shifts and is no longer even within the range of the impact areas predicted several days earlier. In addition, predictions of strengthening or weakening of the storm as it approaches over the ocean are often incorrect, and sometimes the storm intensity suddenly and unexpectedly increases or decreases. Similar observations could be made about other model predictions.

Why are models not better predictive tools? There are several reasons. First, the models, although very sophisticated, are not the real world, and they do not account for variability within the cells of the model. Model cells often represent a large segment of the oceans or atmosphere, within which there is heterogeneity and variability that is not taken into account in the model. Second, the equations describing flow between cells are "fitted" to the testing and validation data and do not precisely describe all the factors that can affect these flows. Third, and this is the most important reason, natural systems include many nonlinear processes, and thus, these systems are **chaotic**, as discussed in more detail in **CC11**.

Even the most sophisticated models cannot make detailed and accurate long-term predictions for chaotic systems. In these systems, models will at best be able to accurately predict a range of likely future conditions. This is indeed exactly what hurricane path predictions do. These predictions are generally quite accurate for approximately a 24-h period ahead of the prediction, but they become progressively less accurate as the predictions extend out in time. Similarly, weather predictions are often quite accurate up to about 24 h ahead, but they become much less accurate farther out in time. In fact, there is a theoretical limit to future predictions in chaotic systems. Because this limit is only several days for weather predictions, these predictions will never be accurate for more than a few days into the future.

CRITICAL CONCEPT 11

Chaos

ESSENTIAL TO KNOW

- Some interactions in natural systems are linear (or almost linear over the range of values present in the environment), but many are nonlinear because the components do not vary in direct proportion to each other. For example, doubling the food supply may cause the population of an animal species to double if food is scarce, but it may result in little change in population if food is already abundant.
- Linear systems are predictable. If we increase one parameter (e.g., the available food supply), the system will adjust to equilibrium (the population will rise to a new stable value that is related in direct proportion to the new food supply).
- Nonlinear systems behave unpredictably. Changing one parameter may lead to a new equilibrium (which may be drastically different from the original system) or may cause the system to oscillate in a chaotic way.
- Complex nonlinear systems with many interactions are

chaotic and unpredictable, but they tend to vary irregularly within an identifiable range of conditions. For example, weather is inherently unpredictable, but climate (range and average conditions) does not generally change from year to year.

- Complex nonlinear systems may respond to unusual changes in one or more parameters by "jumping" to oscillate around a different set of average conditions.
- Natural systems include many interactions, at least some of which are nonlinear. All Earth environmental processes including, atmosphere, and ocean processes are therefore chaotic.
- Because natural systems are chaotic, their exact state cannot ever be forecasted (predicted) accurately. These systems are inherently variable, and they may undergo sudden changes to new average conditions.

UNDERSTANDING THE CONCEPT

In most cases, we conceptualize the world around us as be-

having according to simple rules. For example, it seems reasonable that if we doubled the temperature difference between ocean water and the overlying air, we would find that the rate of heat transfer between the two also doubled; or if we doubled the rate of food production, we would find that the animal species eating this food doubled in population. These are linear relationships in which one parameter changes in direct proportion to another. To be linear, relationships need not be one-to-one (as just described).

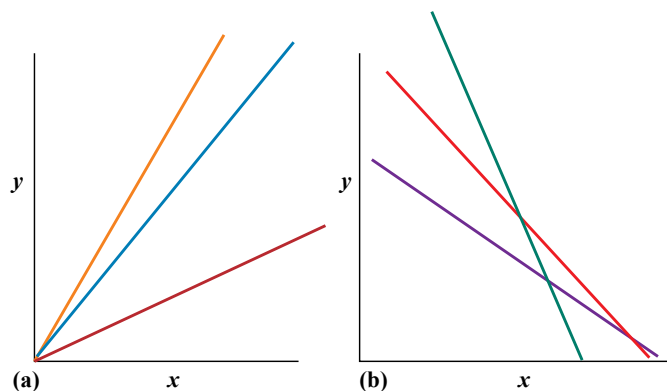


FIGURE CC11-1. Linear relationships can be depicted by the straight line on a graph that plots one variable (x) against another (y). The value of one variable may (a) increase with an increase in the other variable, or (b) decrease with an increase in the other variable.

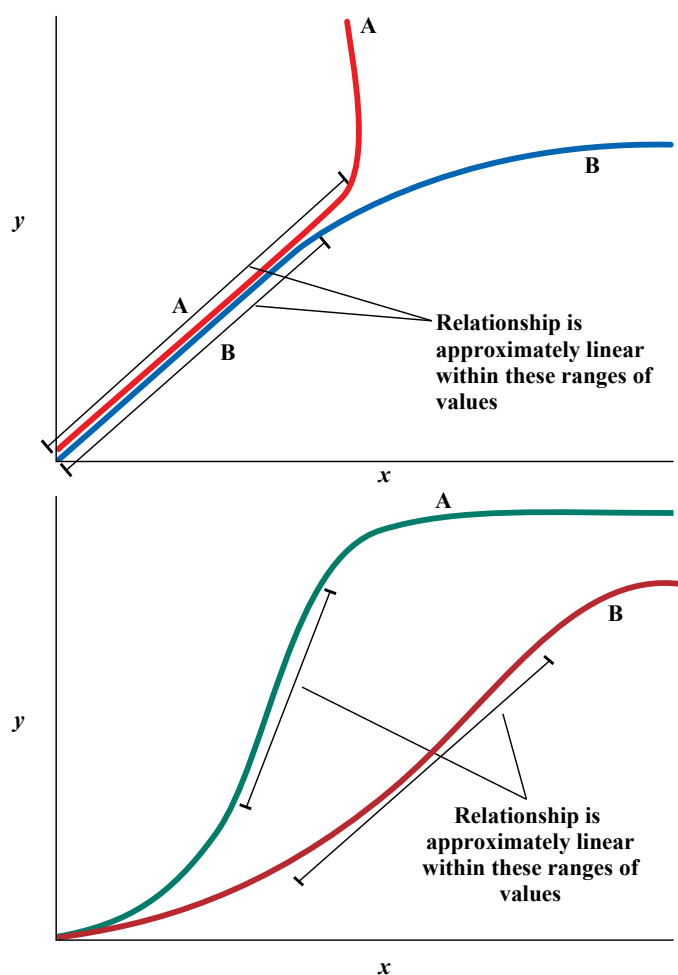


FIGURE CC11-2. Nonlinear relationships can be identified by the curvature of a graph that plots one variable (x) against another (y). Although the curvature can be complex, there is often a range of two values within which the relationship is approximately linear.

The only requirement is that a certain amount of change in one parameter will always result in the change of another parameter by the same amount, even if one parameter increases and the other decreases (**Fig. CC11-1**).

Linear relationships are simple to conceptualize and to express mathematically. They are also intuitively attractive. For example, the hypothesized linear effect of changing food production rate on a consumer population seems perfectly reasonable at first. However, it does not represent the true behavior of the systems. To understand why, think about what would happen if we were to drastically increase or decrease the food supply for an animal species that has no predators. If we drastically increased the food supply, the population would increase, but at some level, the species would become so crowded that some other factor (e.g., breathing space) would prevent it from increasing further. If we drastically decreased the food supply, the population would decline. Eventually, it would reach such a low level that the species would not be able to reproduce successfully because, for example, the survivors would have to spread so far apart to obtain sufficient food that there would be too few available mates. Because the hypothesized relationships do not maintain proportionality if the value of one parameter is changed too much, they are nonlinear.

Examples of nonlinear relationships are depicted in **Figure CC11-2**. Many, if not all, of the relationships that control the real world are nonlinear, which critically affects our ability to model the real world. Fortunately, many relationships, such as the one just discussed, may be linear or nearly so over a wide range of values of the two parameters (**Fig. CC11-2**). For instance, the food–population relationship discussed earlier may be nearly linear within the range of food production rates that are observed in the real world. In these circumstances, we can use linear relationships in our models and the models will be similar enough to the real world to be useful.

Not until the 1970s did scientists begin to examine the behavior of mathematical models that incorporate nonlinear relationships. What they found was startling. Models that use only linear relationships, no matter how complicated, always reach a new equilibrium if one of the parameters is changed (unless parameters are constantly changing). However, even the simplest of nonlinear models involving only two parameters may never reach an equilibrium. The system may oscillate in a seemingly random, or **chaotic**, way, never exactly repeating itself.

The mathematics of nonlinear systems can be complicated. Fortunately, we can understand chaotic behavior by considering a simple system called the “Lorenzian waterwheel,” named for Edward Lorenz, who first investigated the chaotic behavior of this system (**Fig. CC11-3**). In this waterwheel, the buckets are pivot-mounted so that they always stay upright (like the chairs in a Ferris wheel). Each bucket has a small hole through which water can escape (but not drip into a bucket below). This system can be modeled by three simple but nonlinear equations, and it has also been constructed and tested as an actual physical model.

If we pour water into the top bucket, the waterwheel’s behavior depends on how fast the water is poured, how fast each bucket empties, and how quickly the wheel turns (which determines how long each bucket is in the right part of the rotation to be filled).

What do you think happens if we pour water into the Lorenzian waterwheel at different rates but leave everything else unchanged? If we pour the water in at a very slow rate, we might

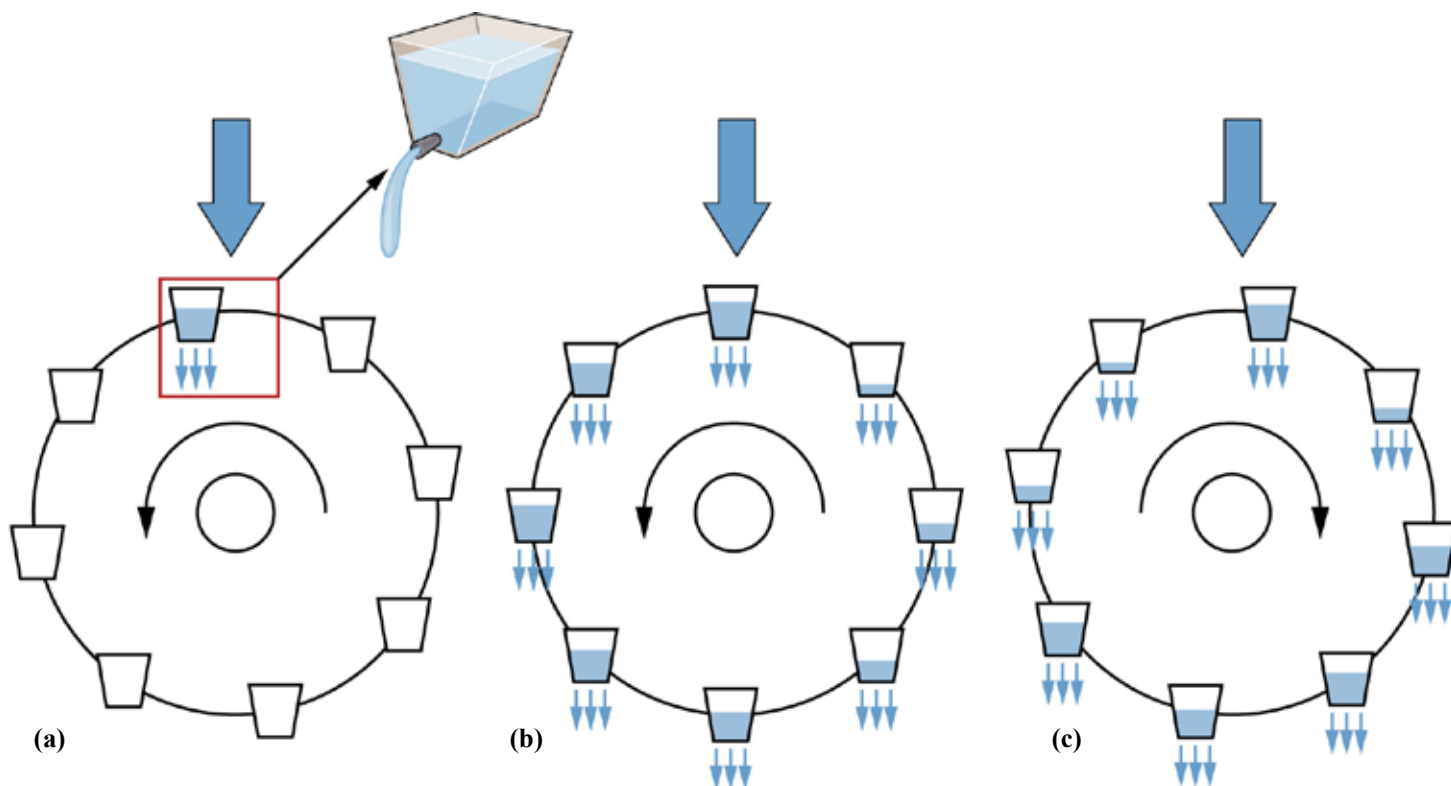


FIGURE CC11-3. The Lorenzian waterwheel. (a) If water is poured into the top bucket faster than it can escape from the bucket, the wheel will be set in motion. (b) If the water flow to the buckets is maintained, the wheel may settle into an equilibrium in which it rotates at constant speed. This can happen when, for example, the buckets are just empty by the time they reach the top and are refilled. (c) The acceleration and speed of the wheel depend on the amount of water in each bucket, which depends on the length of time that each bucket spends under the fill point, which depends on the speed of rotation of the wheel. These are nonlinear relationships, and the wheel may never settle into a constant rotation rate, even if the water input rate remains constant. Furthermore, the rotation rate may vary constantly, and the direction of rotation may even be reversed.

expect it to escape from the top bucket fast enough that the bucket does not fill and the waterwheel will not turn. This is indeed what happens. If we pour the water faster, the top bucket will partially fill and set the wheel in motion, bringing the next bucket under the pouring point to fill in turn and continue the rotation (**Fig. CC11-3b**). As each bucket moves through the rotation, it will progressively empty. We might expect that, if we maintain a stable rate of filling, the rotation will be smooth and the rate of rotation will increase as the fill rate increases. This is what happens when the fill rate is relatively slow. However, at faster fill rates something interesting happens.

Consider the wheel rotating relatively slowly. One or more successive buckets fill almost to the top as the wheel rotates. Because these buckets are heavy, they accelerate the rotation as they move toward the bottom point of the wheel, causing buckets behind them to fill for a shorter time. The full buckets can pass through the low point in the wheel before they have emptied, whereas the following buckets are filled less and empty completely. Once past the low point in the rotation, a once-full bucket (now partially full) may be heavier than the once partially filled buckets (now empty) behind it, so it now tends to slow the rotation of the wheel (**Fig. CC11-3c**).

We can see that the rotation rate of the wheel will vary constantly and the direction of rotation may even reverse. However, what is most surprising about the motion is that, even though we do not vary the fill rate, the wheel's rotation does not, as we might intuitively expect, settle into a regular back-and-forth oscillation. Instead, the rotation varies continuously back and forth in a pattern that is never repeated, a "chaotic" situation. This

happens because of the nonlinearity in relationships, such as the relationship between the rate of rotation and the amount of filling of each bucket.

Because natural systems involve many nonlinear relationships, they can act like the Lorenzian waterwheel: sometimes they can appear stable or change smoothly to a new equilibrium if one of their important parameters changes, or they can oscillate chaotically. For example, a **convection** cell is similar to the waterwheel, although inverted. If heat is supplied very slowly to the lower part of a fluid and lost slowly from the top, the heat can be transferred by **conduction** within the fluid and there is no movement. At somewhat higher heating or cooling rates, smoothly flowing convection cells may form (**CC3**), and at even higher rates the convection cell motion becomes **turbulent** and chaotic.

Even seemingly simple biological systems can exhibit this type of behavior. For example, as discussed previously, populations respond in a nonlinear way to changes in food availability. If food supply increases, then population increases accordingly, but as population continues to increase, other factors eventually limit and reduce the population. Depending on the nature of these relationships, an increase in food supply can lead to a new, higher population of the consumer species, chaotic variations of this population, or even collapse and extinction if, for example, a small increase in population drastically reduces reproduction rates because of overcrowding.

Natural systems generally depend on many interactions, some of which are nonlinear. As a result, most natural systems are chaotic, and this fact has many implications for science. Since the 1970s, a new scientific subdiscipline has developed to study

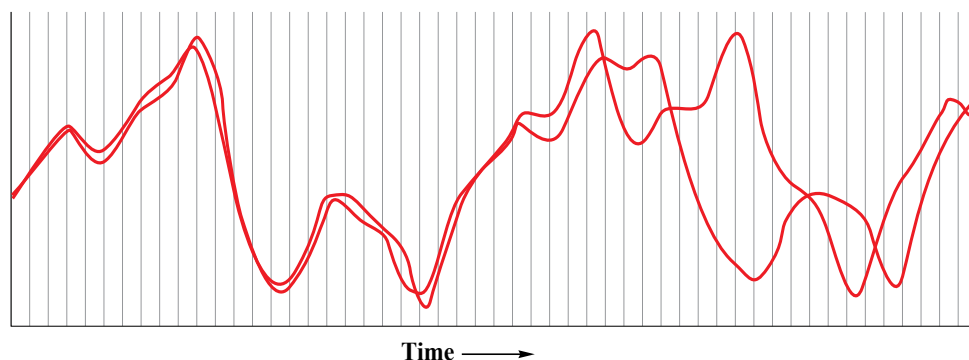


FIGURE CC11-4. These two plot lines depict the results obtained from the same nonlinear mathematical model that simulates weather patterns. The only difference between the two model runs is that the starting-point data were rounded to a smaller number of decimal places for one model run than for the other. The patterns grow farther and farther apart until they lose any resemblance to each other. Many natural systems show this sensitive dependence on initial conditions. As a result, it is impossible to measure and model such systems well enough to predict their precise future behaviors.

chaos. The term chaos has now generally been replaced by complexity. Complexity is often the preferred term because certain characteristics of complex nonlinear systems can be deduced that enable us to understand their behavior and how their component parameters might vary under certain circumstances, even though the precise future status of the system cannot be predicted. We do not need to fully understand the behavior of nonlinear systems unless we become research scientists, but several of their characteristics are important to all of us.

Nonlinear system models are extremely sensitive to very (infinitely) small differences in initial conditions. This characteristic was discovered by Edward Lorenz when he was using a computer model of **weather** patterns, and these observations are what started the systematic study of chaos. Lorenz found that if he started his mathematical model several times from the same set of input parameters (e.g., temperature and **humidity** distribution), the predicted weather exhibited variations that, although complicated, were reproduced exactly each time the model was run.

Lorenz then decided to start his model at a time interval later than the starting point of his previous model runs. He started this new set of calculations with the values of input parameters that his previous model runs had predicted would occur at this time interval. Unexpectedly, he found that his model quickly began to predict weather patterns that differed from those that the model had predicted for this same time period when the model had been started at the earlier time (**Fig. CC11-4**). The reason for this difference, Lorenz found, was that the nonlinear equations in his model generated completely different weather patterns when the later-start model run used input data that were rounded from the values produced in the earlier-start model runs. When he ran his model from the beginning, the computer maintained each parameter's value at each time interval to six decimal places (e.g., 0.267902). When he started his model in the middle, he rounded the input data to three decimal places (e.g., 0.268).

This sensitive dependence on initial conditions has often been called the “butterfly effect.” This term originated from the example that a butterfly stirring the air one day in Beijing could alter the storm systems in New York the next month. As a consequence of the butterfly effect, models of nonlinear systems can never predict future behavior exactly. In addition, the further into the future a model predicts, the less likely it is to be accurate. For example, weather forecasts will never be much better than they are today. We will never be able to accurately predict the exact time or place a thunderstorm will occur on the next day, nor predict with certainty whether it will rain a few days in the future. Similarly, we will never be able to forecast or predict the exact future state of any complex nonlinear system, including **plate**

tectonic movements, **ecosystem** dynamics, or fish and wildlife population variations.

Fortunately, although many complex nonlinear systems never reach equilibrium, they do tend to oscillate chaotically within a range of conditions (values of the component parameters) that models can predict. For example, weather is chaotic and unpredictable, but **climate** (the average and range of temperature, rainfall, etc., at a specific location and time of year) can be predicted with some confidence because it changes little from year to year. Hence, models of complex systems (**CC10**), if they adequately resemble the real world, can be used with reasonable success to predict the average conditions and range of variations that will occur in the future.

Unfortunately, although complex nonlinear systems do tend to oscillate chaotically within a predictable range of conditions, this dynamic “equilibrium” can be disturbed by small changes in component characteristics. In some instances, if a critical value of one or more components is changed, the system can “jump” from one set of average conditions to a completely different set of average conditions around which it oscillates chaotically. Thus, a system may appear to suddenly change drastically, even though none of the components of the system changed substantially before the jump occurred. For example, a very small increase in a parameter, such as the average temperature of the oceans and atmosphere, could have little effect on climate until a critical point is reached. At that point, the Earth's climate could suddenly become much warmer or colder. Recognition of this concept has led to the now well-known term “tipping point”.

We now know that this type of sudden climate change has occurred in the past (**Chaps. 9, 10, Fig. 7-21**). Similarly, sudden changes in ecosystems, such as drastic declines or blooms in certain species, may be a natural consequence of nature's nonlinearity. Sudden changes in other complex systems, such as ocean circulation and the motions of tectonic plates, may have occurred in the past for this same reason. It is important to realize that sudden changes in natural systems have occurred in the past and will occur again in the future. Furthermore, no matter how well we model such systems, it will be very difficult to reproduce these changes faithfully, and virtually impossible to develop an accurate predictive capability that will alert us to them before they occur.

CRITICAL CONCEPT 12

The Coriolis Effect

ESSENTIAL TO KNOW

- When set in motion, freely moving objects, including air and water masses, move in straight paths while the Earth continues to rotate independently.
- Because freely moving objects are not carried with the Earth as it rotates, they are subject to an apparent deflection called the “Coriolis effect.” To an observer rotating with the Earth, freely moving objects that travel in a straight line appear to travel in a curved path on the Earth.
- The Coriolis effect causes an apparent deflection of freely moving objects to the right in the Northern Hemisphere and to the left in the Southern Hemisphere. The deflection is said to be *cum sole*, or “with the sun.”
- The Coriolis deflection is greatest at the poles and decreases at lower latitudes. There is no Coriolis effect for objects that move directly east to west or west to east at the equator.
- Regardless of their speed, freely moving objects at the same latitude appear to complete a circle and return to their original location in the same period of time. This period, called the “inertial period,” is 12 h at the poles and increases progressively at lower latitudes. The inertial period is 24 h at 30°N or 30°S, and it approaches infinity near the equator.
- For objects moving within 5° on either side of the equator, the Coriolis effect can often be ignored because the deflection is small and the inertial period is very long.
- Freely moving objects moving at the same speed appear to follow circular paths with smaller radii at higher latitudes. Freely moving objects at the same latitude appear to follow paths with larger radii if they are moving at higher speeds.
- The magnitude of the Coriolis deflection (the rate of increase of distance from a straight-line path) increases with increasing speed.

UNDERSTANDING THE CONCEPT

From common experience, we know that freely moving objects, such as bullets, move in a straight line unless acted on by an external force. However, all is not what it seems. If we very carefully examine the bullet's flight, we find that its path is deflected very slightly to one side (in addition to the downward deflection due to **gravity**). Curiously, this deflection is always to the right in the Northern Hemisphere and to the left in the Southern Hemisphere. This apparent deflection, called the **Coriolis effect**, occurs because we live on a rotating Earth, and it occurs for all objects that move freely without connection to the solid Earth surface, including bullets, thrown footballs, and moving air and **water masses**.

Some simplified examples help to explain how the Coriolis deflection works.

Merry-Go-Round without Gravity

Consider a merry-go-round rotating counterclockwise when viewed from above (**Fig. CC12-1a**). This is equivalent to the Earth seen from above the North Pole (**Fig. CC12-1c**). Imag-

ine that you are an observer sitting on the merry-go-round and that someone at the center of the merry-go-round throws a ball outward. As the ball travels through the air in a straight line, the merry-go-round continues to turn, carrying you around with it (**Fig. CC12-1a**). You do not see the ball travel in a straight line. Instead, the ball appears to travel in a curve deflected to the right (**Fig. CC12-1b**). If the merry-go-round spins clockwise (equivalent to viewing the Southern Hemisphere from above the South Pole; **Fig. CC12-1d**), the apparent deflection of the ball is to the left (**Fig. CC12-1e**).

If you sit on the merry-go-round away from its center of rotation and throw a ball from there at a target, the ball again appears to be deflected to the right. However, it misses the target, even though it flies in a straight line. The reason is that the ball was moving with the merry-go-round when it was thrown. While in the observer's hand, the ball is constrained to move in a circle with the merry-go-round, but when released, it is no longer constrained. In fact, if the ball were simply released on the merry-go-round instead of thrown, it would roll off in the direction it was moving at the instant it was released. The same thing occurs when we twirl an object on the end of a piece of string and then let go. The object flies off in a line whose direction is determined by where it was in its circle of rotation when it was released (**Fig. CC12-2a**). If a ball is thrown from a rotating merry-go-round, it is given straight-line motion by the throw. However, because of the rotation of the merry-go-round, it is already in motion in a direction tangential to the ball's circle of rotation at the point it was released. These two straight-line motions (components) are combined, so the ball's actual direction is a straight line between the directions of the two components (**Fig. CC12-2b**). The ball's direction and speed are easily calculated from the two components, which are called “vectors” (**Fig. CC12-2c**).

Centripetal Force

Spinning objects, like the object on the end of a twirling string, fly off if released. This tendency is often called “centrifugal” force. However, centrifugal force does not exist. Circling objects tend to fly off when released because the force restraining them in their circular path is no longer present. A force that acts toward the center of rotation, called **centripetal force**, must be applied to maintain any orbiting object in its rotating orbital path (which can be circular or elliptical). This force may be exerted in several ways. For example, **friction** between the observer and the merry-go-round keeps the observer from flying off, whereas tension in the string prevents a twirling object from leaving its circular path. Earth's gravity is the centripetal force that prevents us from flying off the Earth.

Centripetal force must increase as the rate of rotation increases. For example, the amount of friction needed to keep an observer on the merry-go-round becomes greater as the merry-go-round speeds up. The rate of rotation is measured by the angular velocity (the rate of rotation measured in degrees of angle per unit time; **Fig. CC12-3a**). For a constant angular velocity, orbital velocity increases at points more distant from the center of rotation (**Fig. CC12-3a**). At a fixed distance from the center of rota-

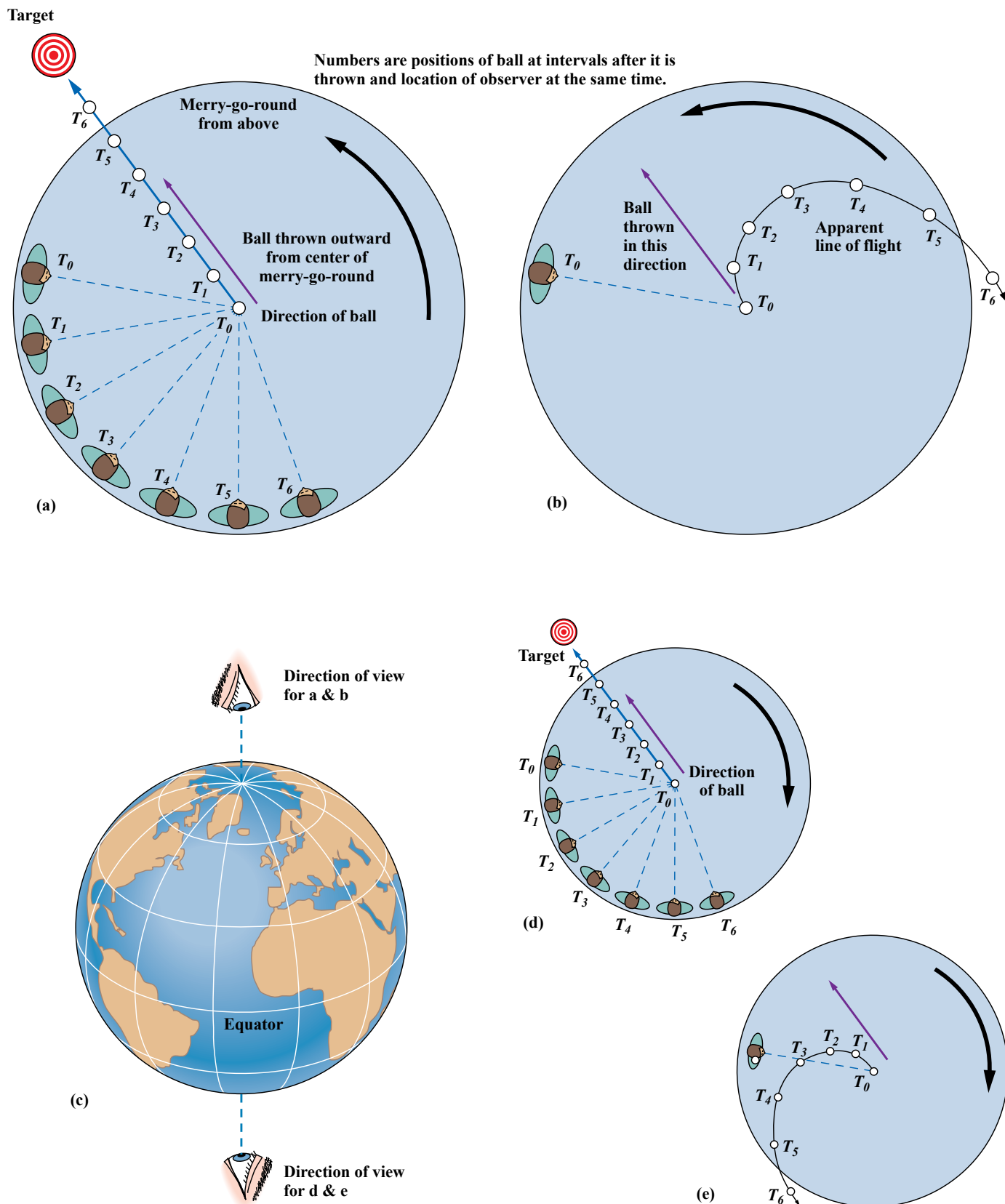


FIGURE CC12-1. (a) A ball thrown outward toward a target from the center of rotation of a merry-go-round has zero orbital velocity due to the merry-go-round's rotation, and will travel outward in a straight line (compared to the ground) to hit the target. However, an observer sitting away from the axis of rotation moves around with the merry-go-round such that the observer's angle of vision to the ball and target changes. (b) If the observer is not aware of the merry-go-round rotation (just as we are not aware of the Earth's rotation), the observer will see the ball appear to move in a circular path, curving to the right. To observe this phenomenon, rotate part (a) to place the observer at the observing location in part (b) for each of the time intervals T_1 , T_2 , and so on. (c) Looking down on the merry-go-round depicted in parts (a) and (b) is equivalent to looking down on the Earth from above the North Pole. If we were to look down from above the South Pole, the merry-go-round would appear to rotate clockwise instead of counterclockwise. (d) If the merry-go-round rotates clockwise (the equivalent of the Southern Hemisphere), the ball still moves in a straight line. However, (e) the apparent deflection is in the opposite direction, and the observer sees the ball curve to the left.

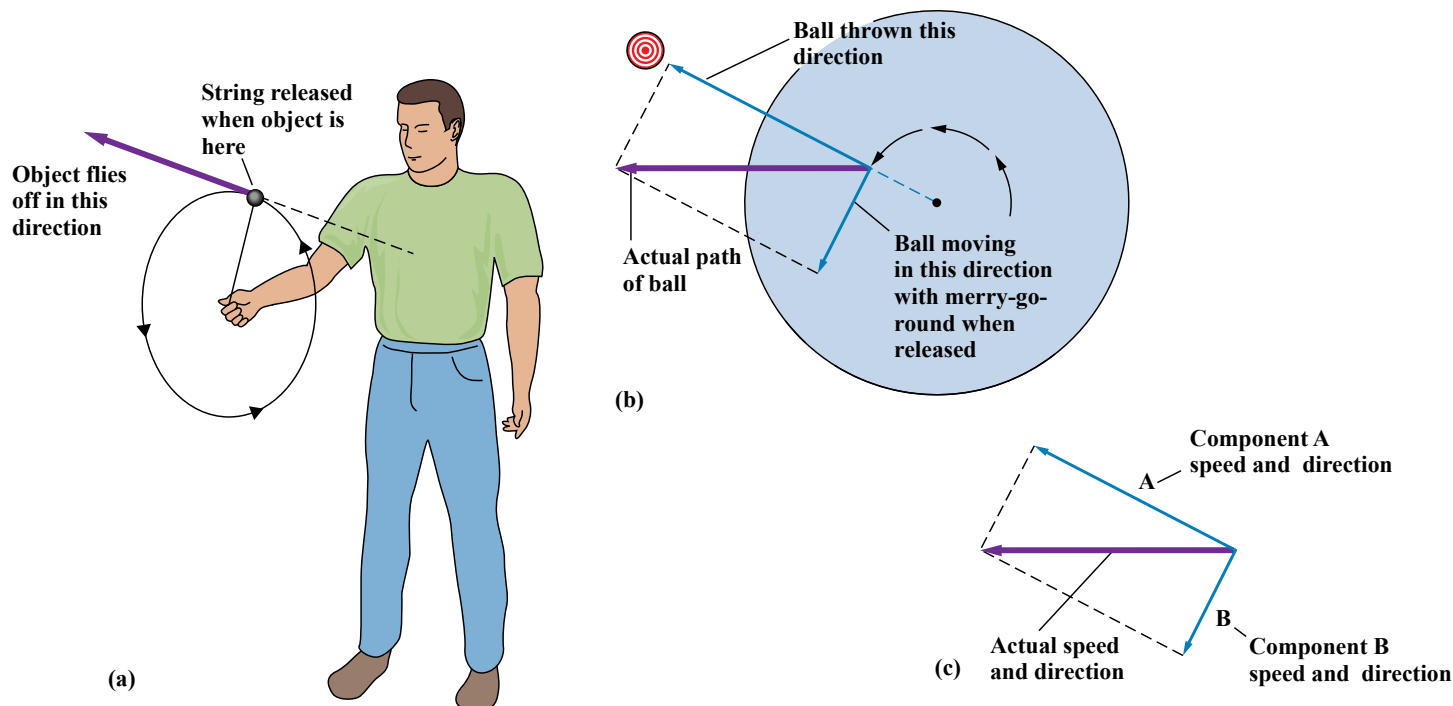
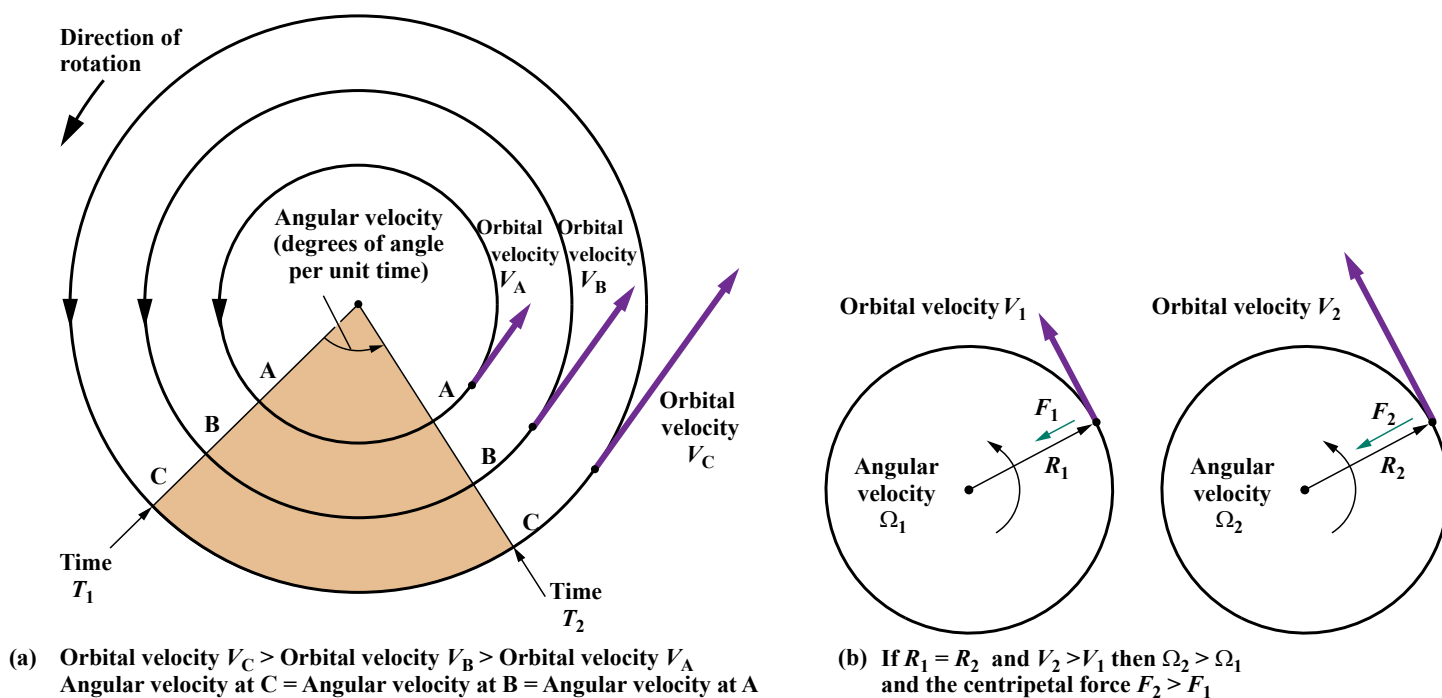


FIGURE CC12-2. (a) A centripetal force is needed to keep an orbiting object in its orbit. When an object is twirled on the end of a string, we supply this force with our muscles through tension in the string. If we release this tension by letting go of the string, the object will fly outward. (b) When a ball is thrown outward from a rotating merry-go-round, it flies off in a straight line (just as the object twirled on the end of a string does), but it also has an additional velocity (speed and direction) imparted to it by the throw. The actual velocity with which the ball moves is a combination of the two velocities. (c) We can determine the ball's actual velocity by geometrically combining its two component velocities. The lines A and B are drawn in the respective directions of the two velocities. The length of each line is proportional to the speed imparted in that direction. The actual direction of the ball's motion can be determined by drawing lines (the dotted lines) parallel and equal to each arrow from the head of the other to form a parallelogram (in this case, it is a rectangle). The diagonal of this parallelogram shows the actual direction of motion, and its length shows the speed.



(a) Orbital velocity $V_C > \text{Orbital velocity } V_B > \text{Orbital velocity } V_A$
Angular velocity at C = Angular velocity at B = Angular velocity at A

(b) If $R_1 = R_2$ and $V_2 > V_1$ then $\Omega_2 > \Omega_1$
and the centripetal force $F_2 > F_1$

FIGURE CC12-3. The velocity of an orbiting object can be expressed as either angular velocity or orbital velocity. (a) The angular velocity is the number of degrees of angle moved through in unit time. The orbital velocity is the distance the object moves along its orbital path per unit time. On a rotating solid object (like a merry-go-round or the Earth), the orbital velocity of a point increases with distance from the center of rotation, and the angular velocity is the same for all points regardless of location. (b) Angular velocity, orbital velocity and centripetal force are related. For two objects at the same distance from the center of rotation, if the angular velocity (rate of rotation) is increased, the orbital velocity is also increased and a larger centripetal force is needed to maintain the object in orbit. For example, the tension in a string becomes greater as we twirl it and its attached object faster.

tion, angular velocity increases if orbital velocity increases, and decreases if orbital velocity decreases (Fig. CC12-3b).

Merry-Go-Round with Gravity

Now imagine a merry-go-round with gravity that acts toward a point located below its center of rotation (Fig. CC12-4a). This merry-go-round is equivalent to a small circular area of the Earth's surface centered at the North Pole with gravity acting toward the Earth's center (Fig. CC12-4b).

If a ball is placed at any point on the merry-go-round other than the center, a small component of the gravitational force pulls it toward the center of rotation (Fig. CC12-4c). If the merry-go-round and ball are rotating and the ball is released to roll freely, the ball will move outward, away from the center of rotation, just as the object on the end of a twirling string flies outward if released. However, except at the center, the component of gravity that acts toward the center of rotation (Fig. CC12-4c) pulls the ball toward this center. If the ball is at the correct distance from the center of rotation, the component of gravitational force parallel to the merry-go-round surface provides exactly the centripetal force needed to balance its rotation with the merry-go-round. Thus, the ball remains in this position and rotates with the merry-go-round (Fig. CC12-4d).

If we set the ball in motion with respect to the merry-go-

round, this balance will be disturbed. For example, if the ball is moved toward the center of rotation (on the Earth, equivalent to moving it toward the pole), the component of gravity parallel to the merry-go-round surface is reduced (Fig. CC12-4c). Because this component of gravity is now less than the centripetal force needed to maintain the ball in its smaller orbit, the ball tends to move back away from the center of rotation. If the ball is moved away from the center of rotation, a small additional gravitational attraction tends to pull it back toward the center.

Next, imagine that the ball is moved in the direction of rotation (equivalent to moving from west to east on Earth). This motion is in the same direction as the ball's path as it rotates with the merry-go-round, so the ball's orbital and angular velocity are increased, and a larger centripetal force is needed to keep the ball in this faster orbit. However, the gravitational force component remains unchanged. Therefore, the ball tends to move outward. Similarly, if the ball is moved in a direction exactly opposite that of the merry-go-round (east to west), the angular velocity is decreased, and the ball tends to move inward. Thus, motions of the ball directed with or against the direction of rotation (east to west or west to east) are turned away from or toward the center of rotation, respectively.

No matter which direction the ball is started in motion, it

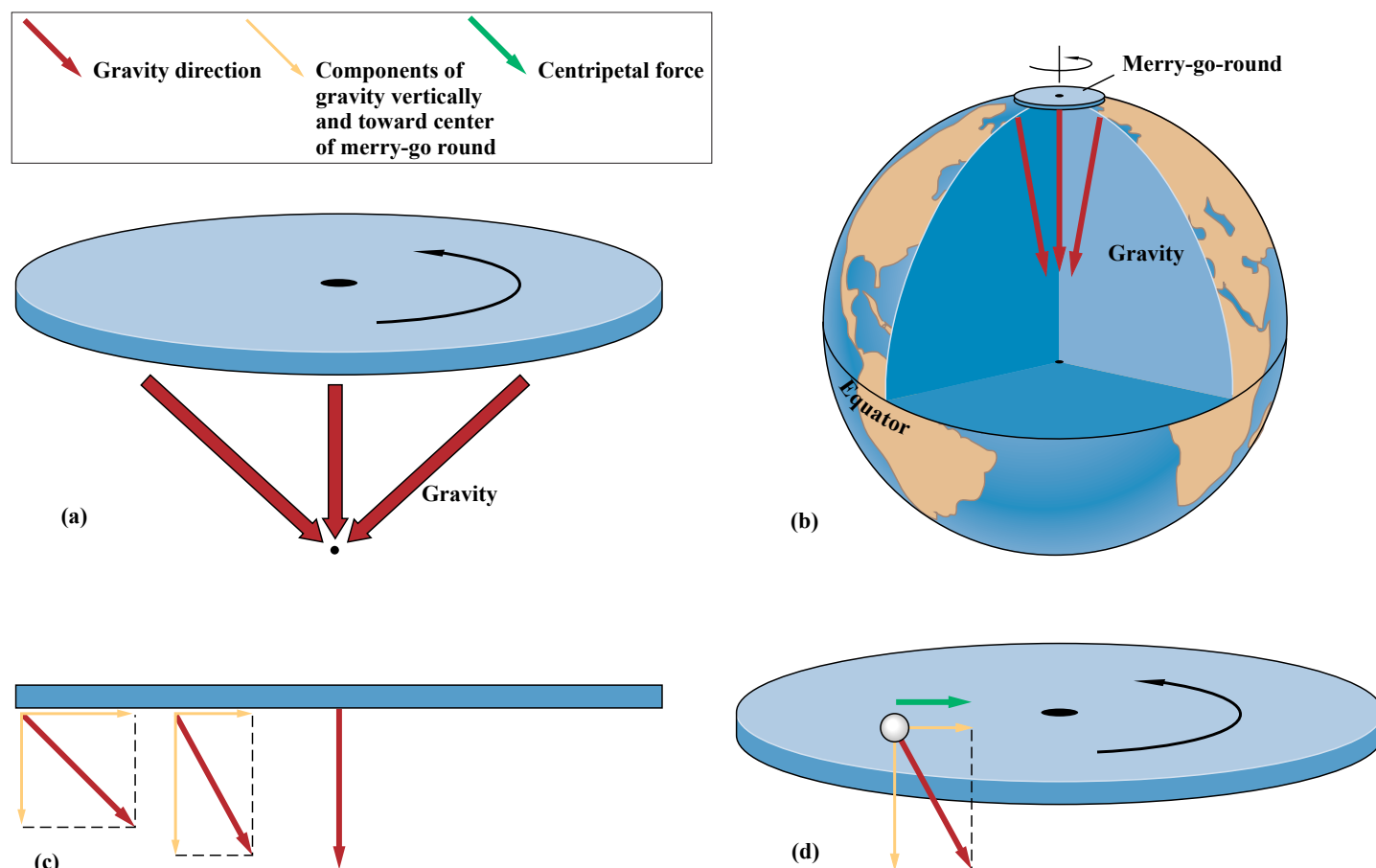
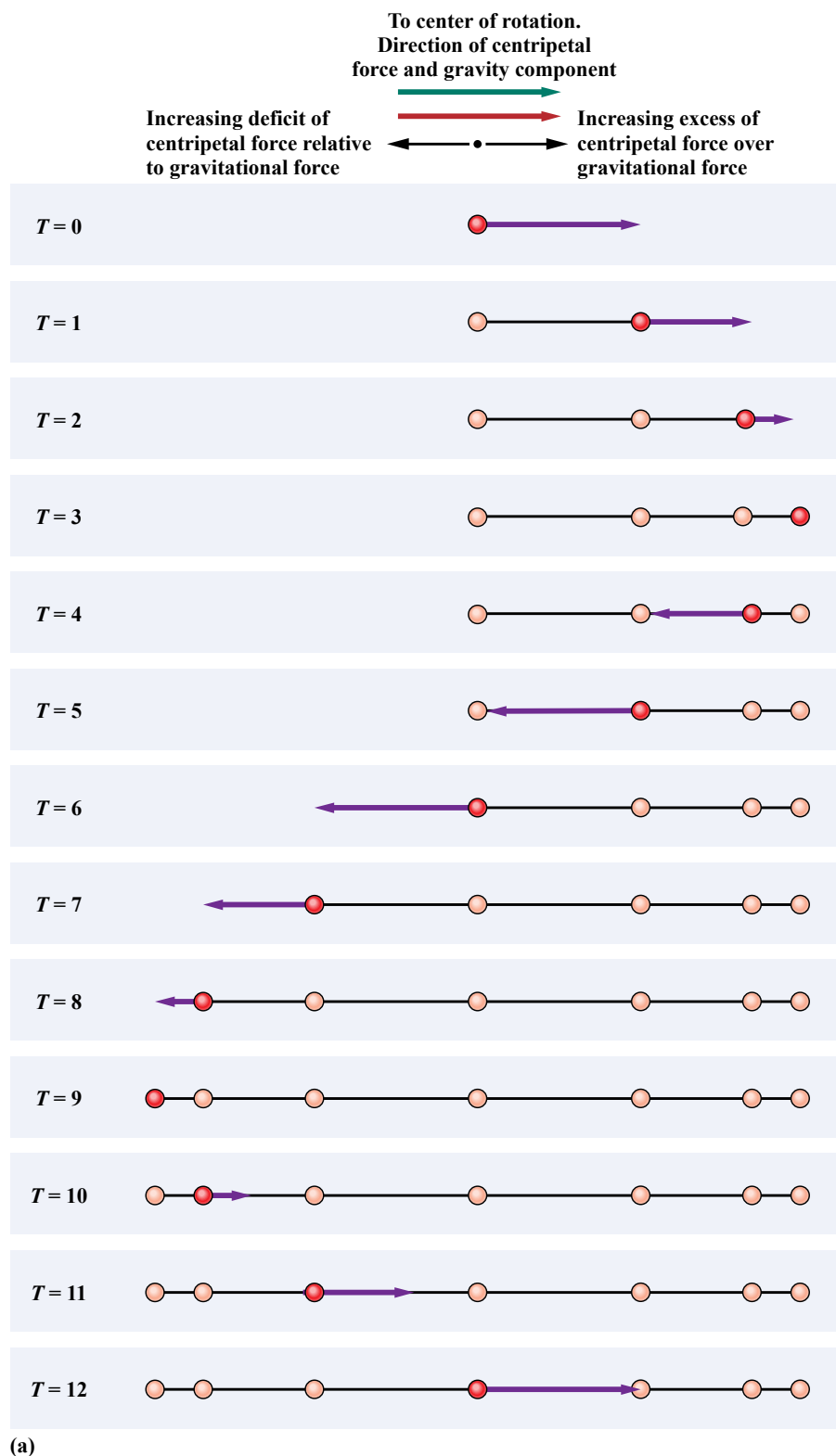


FIGURE CC12-4/ (a) A hypothetical merry-go-round with a gravitational force acting toward a point beneath its center. This hypothetical merry-go-round is like a section of the Earth near the pole (b). However, in this example, the distance to the center of gravity is greatly reduced to help us understand the effects of changes in the directions of gravity and centripetal force with latitude on the Earth's surface. (c) The gravitational force is the same strength (or very nearly so) at all points on the merry-go-round, but it is directed vertically down only at the center of rotation. At all other locations, it can be resolved into two components (the yellow lines), one parallel to the merry-go-round surface and another vertically downward. The component parallel to the surface increases in strength with increasing distance from the center of rotation. (d) If we place a ball on the surface of the merry-go-round, hold it in place so that it rotates with the merry-go-round, and then release it, it will tend to fly outward like an object twirled on a string. However, there is a component of gravity that acts toward the merry-go-round's center. This component provides a centripetal force that, if exactly balanced with the rotation rate of the ball on the merry-go-round (its orbital velocity), will prevent the ball from flying outward.

will move either toward or away from the center of rotation, and this movement will be counteracted by an imbalance between gravity and centripetal force. The counteracting force will grow as the ball moves farther inward or outward. This force will slow the ball's inward or outward motion and eventually reverse its direction. The ball will oscillate toward and away from the center of rotation (Fig. CC12-5a) in a motion similar to that of a

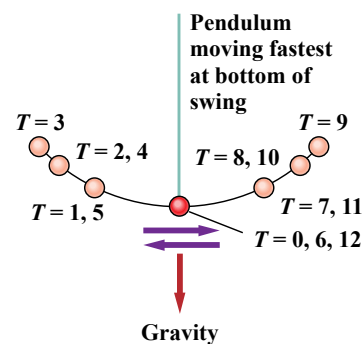
pendulum (Fig. CC12-5b).

As the ball oscillates back and forth, it is still moving in its circular orbit with the merry-go-round. We can show mathematically that the ball would always go through one complete oscillation from its extreme innermost point outward and back to this innermost point in exactly the time it takes for the merry-go-round to complete one revolution. You need not be con-



T = Time interval after motion initiated
 Velocity of ball. Length proportional to speed
 Location of ball
 Prior location of ball

FIGURE CC12-5. (a) If a ball is set in motion toward the center of a hypothetical merry-go-round with a gravitational force that acts toward a point beneath its center (see Figure CC12-4), the ball will oscillate back and forth. As it moves toward the center, the component of gravity in this direction decreases while the required centripetal force increases (because the ball retains its orbital velocity but is now in a smaller-radius orbit). The imbalance between centripetal force and gravitational attraction slows and then reverses the ball's direction of motion. The ball is accelerated until it passes through its original location relative to the center of rotation, but it has enough momentum to continue outward. As the ball moves farther away from the center, the component of gravity increases while the required centripetal force decreases. As a result, there is an excess of gravitational force over centripetal force, and the ball is slowed, and its motion eventually reverses back toward the center of rotation. (b) These oscillating motions are very similar to the movements of a pendulum.



(b) Pendulum slows as it swings upwards.

cerned with this calculation, but the result is important because we can now see what the path of the ball set in motion would be (Fig. CC12-6). To an observer anywhere on the rotating merry-go-round, the ball would appear to travel in clockwise circles. It would complete a circle twice for each time the merry-go-round

completed one revolution (the equivalent of two circles every 24 h on the Earth).

The Earth, Centripetal Force, and Gravity

The Earth is more complicated than a merry-go-round because it is a sphere, although not quite an exact sphere. Gravity

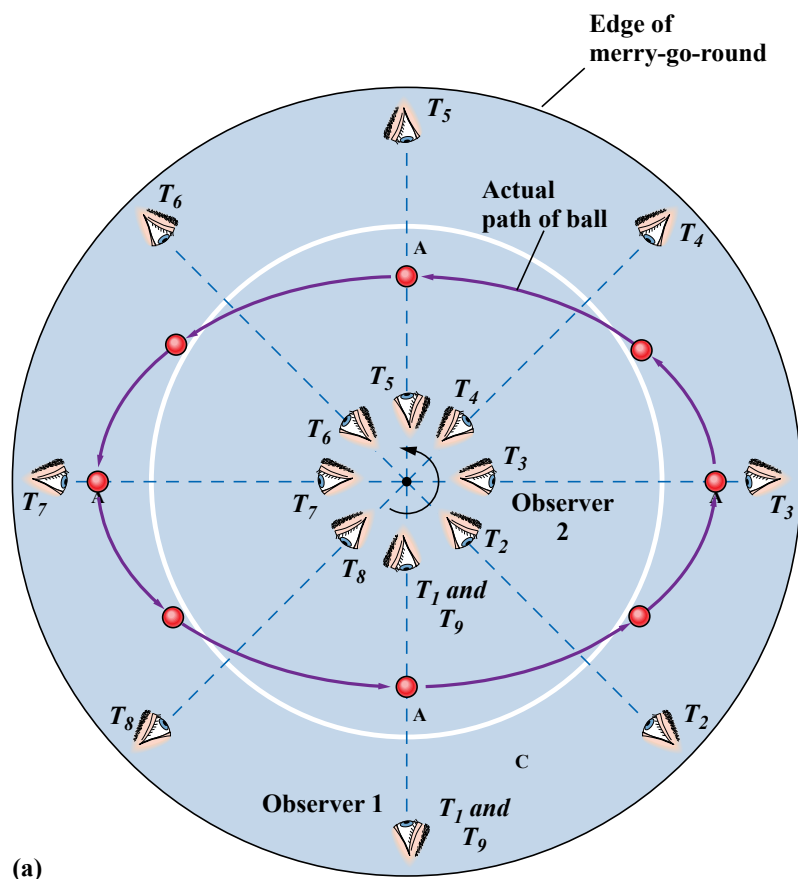
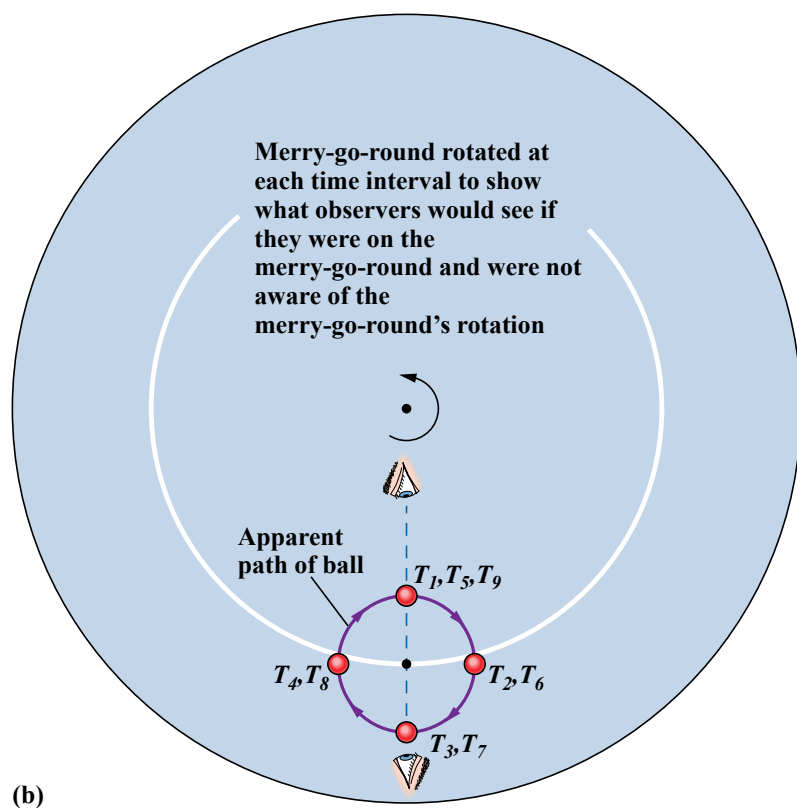


FIGURE CC12-6. (a) If a ball is set in motion toward the center of a hypothetical merry-go-round with a gravitational force that acts toward a point beneath its center (see Figure CC12-4), this motion is added to the orbital motion that the ball already has due to the rotation of the merry-go-round. The added motion and the interaction of centripetal force and gravity cause the ball to oscillate back and forth toward and away from the center, as shown in Figure CC12-5. When the rotational motion is added, the actual path of the ball on the merry-go-round is an ellipse, and it returns to its original position after exactly one revolution. (b) To an observer at any location on the rotating merry-go-round, the ball appears to move in a circle. Follow the ball's direction and distance from each observer's eye in part (a) of this figure and see how they plot in a circle in part (b).



acts toward the Earth's center, whereas centripetal force acts toward the center of rotation (perpendicular to the Earth's axis of rotation). Therefore, gravity and centripetal force act in the same direction only at the equator, and the difference between their directions increases from zero at the equator to 90° at the poles (Fig. CC12-7a).

Centripetal force can be resolved into two components: one parallel to the Earth's surface oriented north–south, and the other oriented toward the Earth's center (Fig. CC12-7a). The centripetal force needed to keep an object on the Earth's surface, and variations in this force needed to balance even very large changes in an object's speed relative to the Earth's surface, are extremely small in comparison with gravity. Therefore, variations in the component of centripetal force that acts toward the Earth's center (the same direction as gravity) are easily compensated by gravity and the **pressure gradient** (Fig. CC12-8). In other words, moving objects whose angular velocity is increased are immeasurably reduced in weight, whereas those whose angular velocity is reduced are immeasurably heavier.

If the Earth were perfectly spherical, the component of centripetal force parallel to the Earth's surface would not be compensated by gravity, because gravity would act perpendicular to the Earth's surface at all locations. Consequently, freely moving objects would “slide” across the surface toward the equator (Fig. CC12-7a). This explains why the Earth is not a perfect sphere, but is instead an oblate spheroid (the shape you would get if you squeezed the Earth at its poles to deform it like a squeezed basketball). The Earth's diameter as measured from the North to the South Pole is 12,714 km, slightly less than its 12,756-km

diameter measured at the equator. Because the Earth is not a perfect sphere, gravity acts at a very small angle to the surface (Fig. CC12-7b), and there is a small component of gravity parallel to the Earth's surface everywhere except at the equator and poles. This gravity component compensates for the component of centripetal force parallel to the Earth's surface (Fig. CC12-7b), so freely moving objects do not slide toward the equator.

If an object is set in motion relative to the Earth, the object's angular velocity will change just as it did for the ball on the merry-go-round. On the Earth, the vertical component of the altered centripetal force is compensated by gravity and the Earth's pressure gradient. However, if the angular velocity of an object is altered, the balance between the small component of gravity acting parallel to the Earth's surface and the component of centripetal force acting parallel to the Earth's surface is upset. The component of gravity parallel to the surface cannot compensate for changes in the corresponding component of centripetal force (unless the Earth changes its shape). Consequently, altering the angular velocity of a freely moving object on the Earth's surface causes it to move across the surface, toward or away from the pole.

Motion on the Earth: North and South

Now let's look at motions on the Earth's surface. First, consider a projectile fired toward the north in the Northern Hemisphere. This projectile has a component of motion from west to east imparted to it by the Earth's rotation. The projectile retains this west-to-east velocity, but the velocity of the Earth rotating underneath it decreases as the projectile moves north (Fig. CC12-9a). The projectile continues eastward while the Earth below it

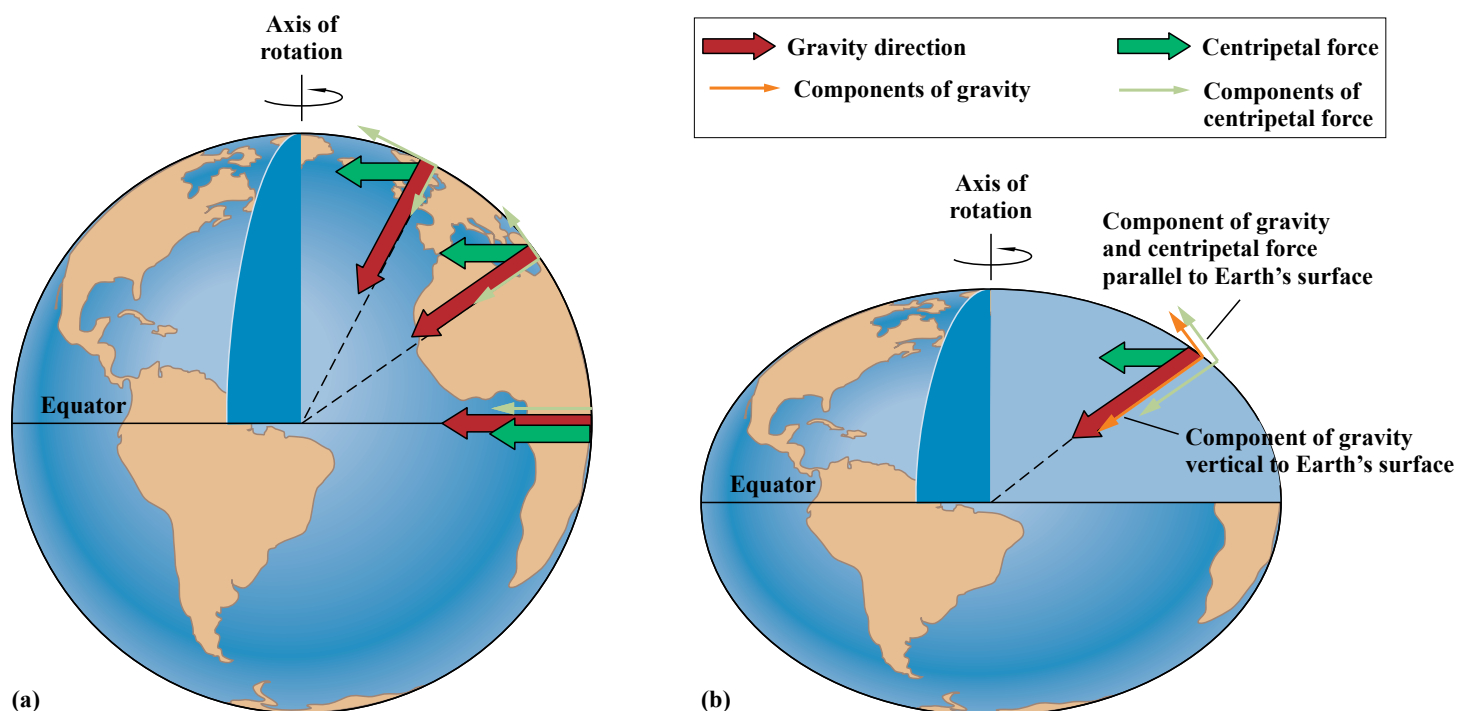


FIGURE CC12-7. (a) If the Earth were exactly spherical, there would be no component of gravity parallel to the Earth's surface at any latitude. However, to keep an object in orbit at the Earth's surface, there would have to be a component of a centripetal force parallel to the Earth's surface, everywhere except at the equator. Because there would be no gravitational force component to provide this component of centripetal force, an object anywhere on the Earth's surface except at the equator would move into a wider orbit (in which the required centripetal force was lower). Thus, the object would slide toward the equator. (b) In response to the force imbalance that would be created if the Earth were spherical, the Earth and all other planets that were at one time totally molten or gaseous were not formed as perfect spheres. Instead, the Earth is an oblate spheroid, a squashed spheroid shape, as shown (but much exaggerated) in the figure. Because the Earth is an oblate spheroid, gravity does not act exactly perpendicular to the Earth's surface, except at the equator and poles. Consequently, at all points other than these, there is a small component of gravity parallel to the Earth's surface that balances the required centripetal force for an object at that point.

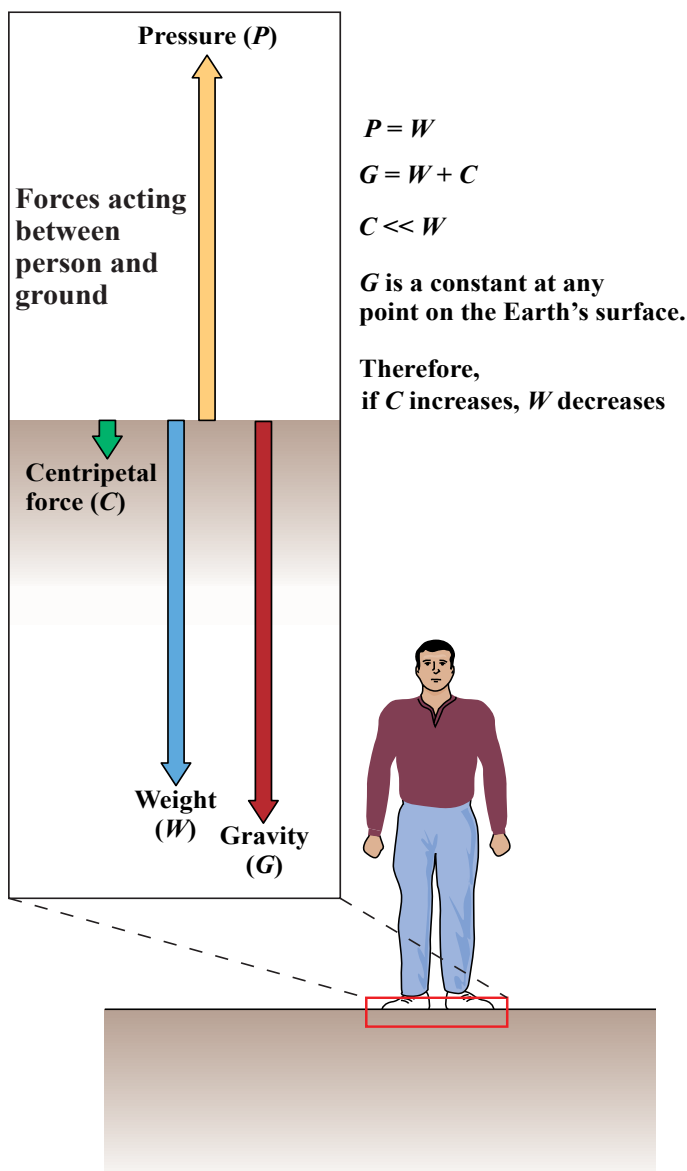


FIGURE CC12-8. There is a balance between the vertical components of gravity (which is invariable at any given location on the Earth), centripetal force, and the Earth's pressure gradient for objects moving over Earth's surface. The pressure gradient force is equal to the object's weight. The pressure gradient force and the object's weight must be very slightly smaller than the gravitational force because the gravitational force must provide the required centripetal force to maintain the object in its orbit with the rotating Earth. Thus, the object's weight equals the gravitational force minus the very much smaller centripetal force. The magnitude of the centripetal force has been greatly exaggerated in this figure.

moves eastward progressively more slowly. Thus, the projectile "leads" the Earth's rotation at an increasing rate and follows a path that appears to an observer to be a curve deflected to the right (**Fig. CC12-9b**). In addition, because gravity prevents the projectile from maintaining its original distance from the Earth's axis of rotation, the projectile's angular velocity increases as it moves north. This requires increased centripetal force and creates a force imbalance that "pushes" the projectile back to the south, away from the Earth's axis of rotation.

Similarly, a projectile fired southward in the Northern Hemisphere passes over an Earth that moves eastward progressively faster under the projectile, and it "lags" the Earth's rotation (**Fig. CC12-9c**). It also appears to be deflected to the right and is

"pushed" back north because it has a lower angular velocity. In the Southern Hemisphere, the situation is similar, but the deflection is to the left (**Fig. CC12-9b,c**).

Motion on the Earth: East and West

When an object is set in motion in a west-to-east or east-to-west direction, it moves either with or against the Earth's rotation (**Fig. CC12-10a**). Thus, the object's angular velocity is increased or decreased. When it is set in motion west to east, angular velocity is increased, and therefore the object tends to move outward away from the pole. A very small reduction in the object's weight allows gravity to compensate for the increased centripetal force and prevents the object from flying upward away from the Earth's surface. However, there is no compensating force to provide the needed increase in the component of centripetal force across the Earth's surface, and the object moves toward the equator, or to the right in the Northern Hemisphere and left in the Southern Hemisphere (**Fig. CC12-10b**).

When an object is set in motion east-to-west, the angular velocity is decreased, and the centripetal force needed to maintain its orbit is slightly decreased. The object's weight increases slightly and prevents it from moving vertically downward. However, the reduction in the component of centripetal force parallel to the Earth's surface remains unbalanced, and the object is moved across the Earth's surface toward the pole. Again, the deflection is to the right in the Northern Hemisphere and to the left in the Southern Hemisphere (**Fig. CC12-10b**).

If the object is moving either west-to-east or east-to-west exactly at the equator, there is no component of centripetal force parallel to the Earth's surface, a small adjustment in weight can totally compensate the changes in centripetal force, and there is no deflection (**Fig. CC12-10b**).

Coriolis Effect Characteristics

All objects moving freely relative to the Earth are deflected by the Coriolis effect unless they are moving directly east or west along the equator. The deflection is always to the right in the Northern Hemisphere and to the left in the Southern Hemisphere. However, it is easier to say that the direction of the deflection is *cum sole*, which means "with the sun." To an observer in the Northern Hemisphere looking toward the equator and thus toward the arc of the sun across the sky, the sun moves across the sky from left to right. In the Southern Hemisphere, the sun moves from right to left.

Freely moving objects appear to move in circular paths on the Earth's surface. These circular paths are distorted if objects move large distances across the Earth's surface. This is because the Earth is equivalent to a flat merry-go-round only at the North or South Pole, where the apparent rotation occurs across a surface that is perpendicular to the axis of rotation. At all other points on the Earth, the motion occurs on a surface that is inclined at an angle other than 90° to the axis of rotation. As a result, at all points other than the poles, a component of the Coriolis deflection is directed vertically downward and "blocked" by adjustments of the object's weight. Consequently, the rate of deflection is reduced.

The proportion of centripetal force directed vertically downward increases from zero at the poles to 100% at the equator (**Fig. CC12-7a**). Therefore, the Coriolis deflection is at a maximum at the poles, is reduced progressively with decreasing latitude, and reaches zero at the equator. In addition, the time necessary for a

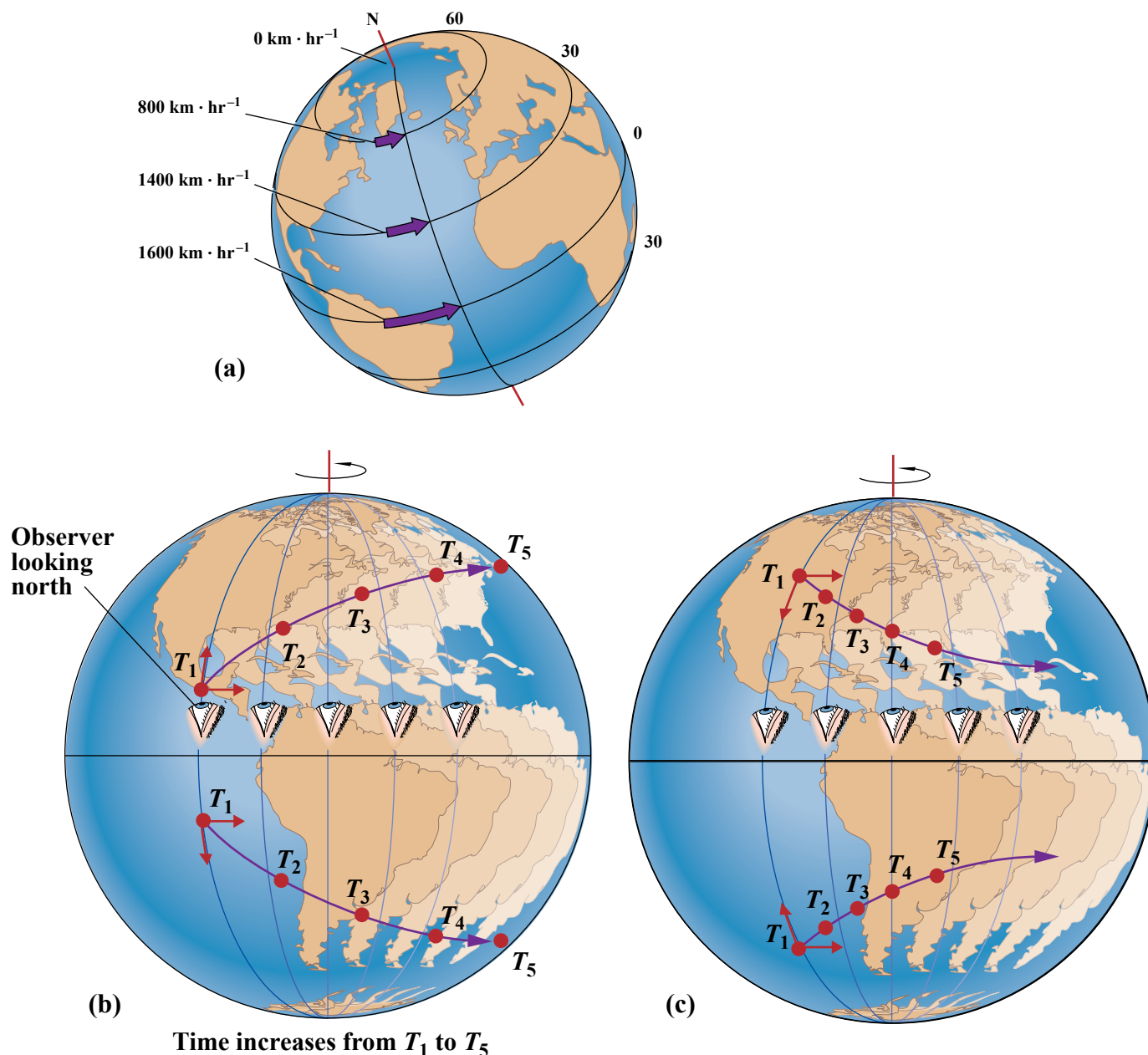


FIGURE CC12-9. (a) The orbital velocity due to the Earth's rotation of points on the Earth's surface decreases as latitude increases. (b) If an object is started in motion directly toward the pole in either hemisphere, it moves progressively into latitudes in which the Earth's surface is moving more slowly in its rotational orbit. However, if the object is moving freely (that is, if it is not attached by strong frictional forces to the solid Earth), it retains the orbital velocity that it had when it was started in motion. Thus, in addition to its motion to the north (or south in the Southern Hemisphere), the object continues to move to the east at its original orbital velocity and progressively moves ahead of the latitude from which it was set in motion. To an observer moving with the Earth's surface, the result is an apparent deflection in the path of the object, to the right in the Northern Hemisphere and to the left in the Southern Hemisphere. (c) If an object is started in motion directly toward the equator in either hemisphere, it moves progressively into latitudes in which the Earth's surface is moving more quickly in its rotational orbit. Thus, in addition to its motion to the south (or north in the Southern Hemisphere), it continues to move to the east at its original orbital velocity and progressively falls behind the latitude from which it was set in motion. The apparent deflection of the path of the object is, once again, to the right in the Northern Hemisphere and to the left in the Southern Hemisphere.

freely moving object to complete a circle increases with decreasing latitude. This time period, called the "inertial period," is 12 h at the poles, increases to 24 h at 30° latitude, and is infinite at the equator.

The diameter of the circle in which freely moving objects appear to move is determined by both the object's speed and its latitude. Because freely moving objects at the same latitude will complete a circle in the same amount of time, a faster-moving object must move in larger circles. A bullet travels in a circle of huge diameter, and along only a very tiny part of its circular path

in the extremely short time it remains airborne. Consequently, it appears to travel in a "straight" line. Air masses in the atmosphere and water masses in the oceans move much more slowly and travel in paths of much smaller radius.

For objects moving at the same speed, the inertial period and the diameter of the apparent circular path decrease with increasing latitude. Therefore, the Coriolis effect "increases" with increasing latitude. Because the diameter of the circle in which even a slowly moving object appears to travel is very large near the equator, the Coriolis deflection is often considered negligible

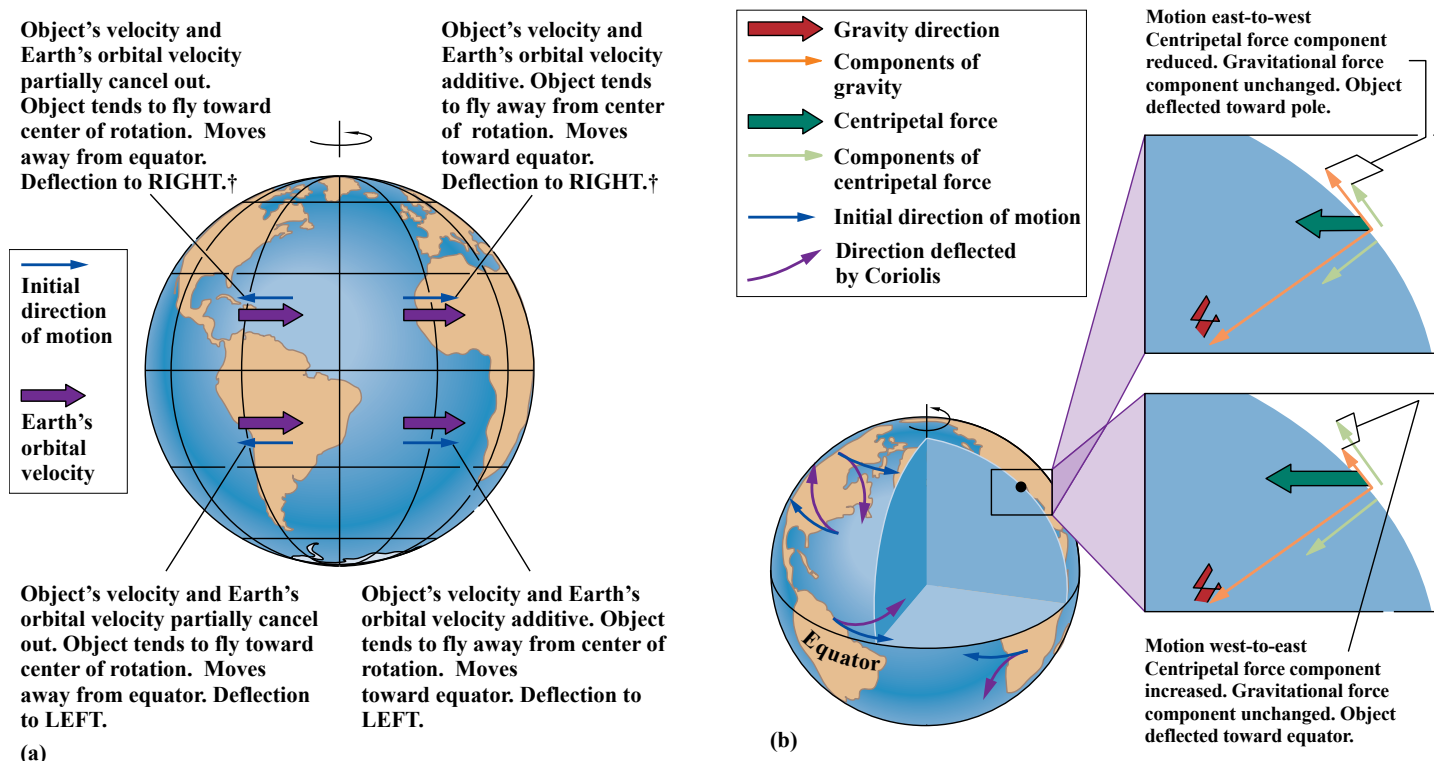


FIGURE CC12-10. (a) When objects are started in motion across the Earth's surface in an east-to-west direction, their velocity is in the opposite direction to that imparted by the orbital velocity of the Earth's surface due to the Earth's rotation. As a result, the effective orbital velocity of the object is reduced, and the centripetal force needed to maintain the body in its orbit with the Earth's spin is decreased. The slight decrease in the vertical component of centripetal force is readily compensated by a negligibly small increase in the object's weight (change in pressure gradient). (b) In contrast, no such compensation is possible for the component of centripetal force that acts parallel to the Earth's surface. The slight excess of gravitational force over the component of centripetal force parallel to the Earth's surface needed to maintain the body in its orbit, with the Earth's spin deflects the object's path toward the axis of rotation or toward the pole in each hemisphere. Thus, the deflection is to the right in the Northern Hemisphere and to the left in the Southern Hemisphere. Similarly, (a) when objects are started in motion across the Earth's surface in a west-to-east direction, their velocity is in the same direction as that imparted by the orbital velocity of the Earth's surface due to its rotation. (b) The effective orbital velocity of the object is increased, but the centripetal force provided by the component of gravity parallel to the Earth's surface is too small to maintain this orbital velocity. Therefore, the object's path is deflected away from the axis of rotation or away from the pole in each hemisphere. This deflection is also to the right in the Northern Hemisphere and to the left in the Southern Hemisphere. For ease of illustration, the magnitude of centripetal force has been greatly exaggerated relative to the force of gravity in this diagram.

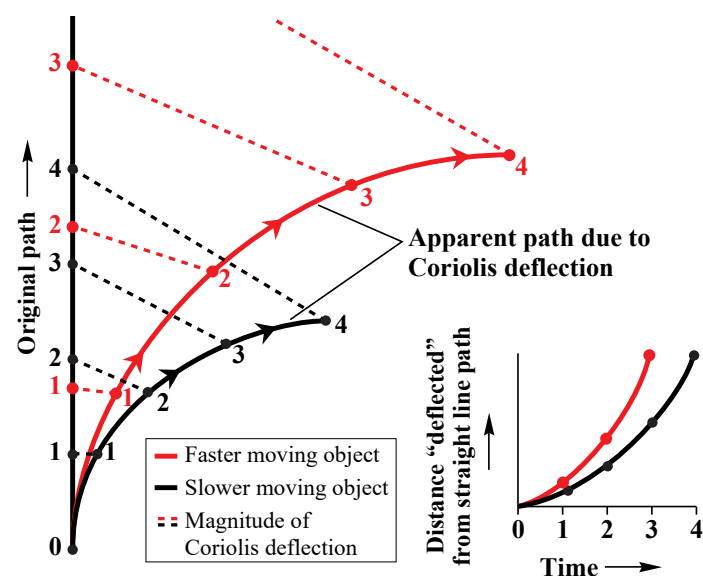


FIGURE CC12-11. The distance between the location of an object deflected by the Coriolis effect and the location it would be expected to occupy if it were not deflected (that is, if it followed a straight-line path) increases more quickly with time if the object's speed is greater. Thus, the magnitude of the Coriolis deflection is said to "increase" with increasing speed of the object.

at latitudes within 5° of the equator.

Objects moving at different speeds at the same latitude are deflected through the same angle in equal times, but the faster object is deflected farther from its original path (**Fig. CC12-11**).

Thus, the Coriolis deflection "increases" with increasing speed.

CRITICAL CONCEPT 13

Geostrophic Flow

ESSENTIAL TO KNOW

- Horizontal pressure gradients exert a force that accelerates fluid molecules in the direction of pressure decrease on the gradient. The acceleration increases as the strength of the pressure gradient increases.
- The Coriolis effect deflects fluids that flow on a pressure gradient until they flow across the gradient. The flow then continues along a line of constant pressure (isobar) as a geostrophic flow.
- Geostrophic flow conditions occur when the pressure gradient force is balanced by the Coriolis deflection.
- Geostrophic wind and current speeds are determined by the steepness of the pressure gradient. Wind or current speed increases as the steepness of the gradient increases.
- Geostrophic wind or current speed and direction can be determined from contour maps. The direction of flow is parallel to the isobars, and the speed is higher where the pressure gradient is steeper (isobars are closer together).
- Geostrophic winds and currents flow counterclockwise around low-pressure zones and clockwise around high-pressure zones in the Northern Hemisphere. In the Southern Hemisphere, they flow clockwise around low-pressure zones and counterclockwise around high-pressure zones.

UNDERSTANDING THE CONCEPT

The atmosphere and ocean waters are stratified fluids (CC1), in each of which density decreases with increasing distance from the Earth's center. The only exceptions are in limited areas where stratification is unstable. It is in these areas of unstable stratification that density-driven vertical motions of the fluid occur.

If the oceans or atmosphere were at equilibrium, density would be uniform at any one depth or altitude. The vertical density gradient would be the same everywhere, so the total weight of the overlying water and/or air column at a specific depth or altitude would be the same everywhere. Because this total weight determines atmospheric or water pressure, pressure would be uniform at any depth or altitude.

Neither the atmosphere nor the oceans is at equilibrium, because the density of atmospheric gases and ocean water is altered locally by changes in such factors as air or water temperature, dissolved salt concentration in the water, and water vapor pressure in the air (Chaps. 5, 7). Consequently, the vertical distribution of density in both oceans and atmosphere varies from place to place, and there are horizontal variations of pressure at any given height in the atmosphere or depth in the oceans.

Surfaces that consist of points of equal altitude or depth are referred to as horizontal or level surfaces. However, these surfaces are actually spherical because the Earth is a sphere. Horizontal pressure gradients develop in the oceans as a result of density differences between water masses and also because winds tend to move ocean surface waters, causing the water to pile up in some locations (Chap. 8).

Where there is a horizontal pressure gradient, the fluid is subject to a force that tends to accelerate molecules from high-pres-

sure areas toward low-pressure areas. The acceleration is greater when the pressure gradient is steeper. We perform one of the simplest demonstrations of acceleration along a pressure gradient every time we open a soda bottle or can. Once they are free to do so, the gas molecules in the high-pressure zone within the bottle are accelerated toward the lower pressure outside.

In our soda bottle experiment, the density gradient is extremely steep because the pressure difference between the air outside the bottle and the gas in the bottle is large, and the distance between the high-pressure zone inside the bottle and the low-pressure zone of the surrounding air is very short. When we open the bottle, gas molecules in the high-pressure zone are accelerated very rapidly and must move only a short distance to reach the low-pressure zone. As a result, the pressure equalizes almost instantaneously. In contrast, the pressure differential between horizontally separated high- and low-pressure zones in the oceans and atmosphere is very small, and these zones are separated by much greater distances. Consequently, the accelerations produced by atmospheric and oceanic horizontal pressure gradients are small. In addition, air or water does not flow directly from the high-pressure zone to the low-pressure zone, because the air and water molecules are subject to the Coriolis effect (CC12) once they have been set in motion.

To understand how motions induced by the pressure gradient and Coriolis effect interact, consider the following facts. Freely moving objects, including air or water molecules, actually travel in straight paths unless acted on by another force, such as the pressure gradient. The Coriolis deflection is only a perceived deflection seen by an observer on the rotating Earth. However, from our rotating frame of reference, the freely moving object is deflected *cum sole*. In Figure CC12-11, we can see that the magnitude of the deflection and the rate of increase both increase with time. These are the characteristics of an acceleration in the direction 90° *cum sole* to the direction of motion, and this “acceleration” increases with the object's speed (Fig. CC12-11).

We can now examine what happens when a fluid begins to flow in response to a horizontal pressure gradient. The fluid is accelerated down the pressure gradient from the high-pressure region toward the low-pressure region. As it begins to move, it is deflected by the Coriolis effect. The faster it moves, the greater is the Coriolis deflection (CC12). The direction of motion thus turns away from the direct path down the pressure gradient, to the right in the Northern Hemisphere and to the left in the Southern Hemisphere. As the moving fluid is deflected, the pressure gradient acceleration continues to act toward the low-pressure region while the Coriolis deflection continues to act at 90° *cum sole* to the changed direction of flow. The speed increases as the molecules are accelerated, thus increasing the Coriolis deflection.

The molecules continue to accelerate while the direction of flow is progressively turned *cum sole* until the flow is directed across the pressure gradient along a contour of equal pressure (Fig. CC13-1). When the flow is in this direction, the pressure gradient, acceleration, and Coriolis deflection act in opposite directions and balance each other. The speed of the fluid at this

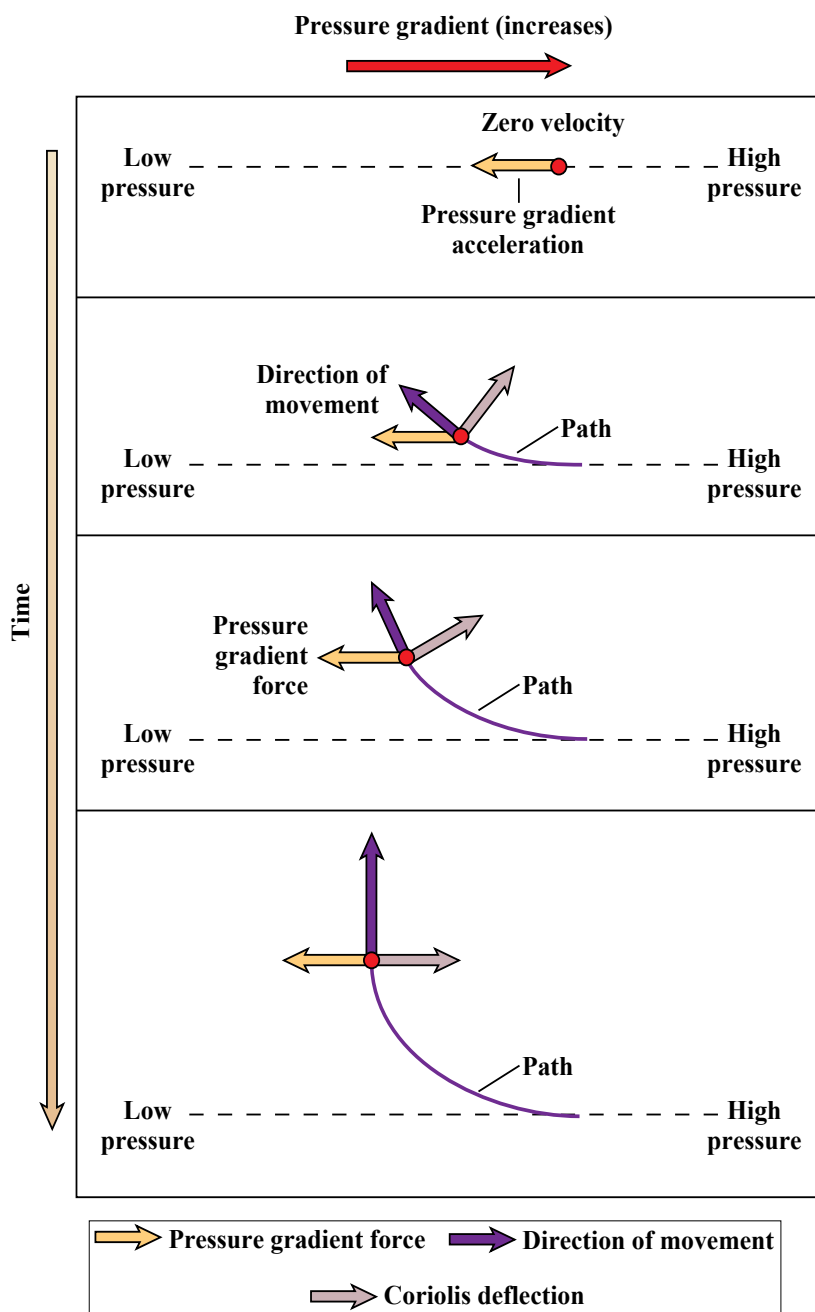


FIGURE CC13-1. When a current is initiated on a horizontal pressure gradient, the initial direction of motion is directly down the gradient. However, as it moves, the water mass is deflected by the Coriolis effect. It accelerates and is deflected until it flows directly across the pressure gradient, and the pressure gradient force is balanced by the Coriolis deflection. This is a “geostrophic” current. If the pressure gradient is steeper, the acceleration is greater, and the geostrophic current is faster, but it is still balanced by the greater Coriolis deflection associated with the higher speed.

balance point is determined by the steepness of the pressure gradient because stronger pressure gradients cause greater accelerations that must be balanced by greater Coriolis deflections, which increase with increasing speed (**CC12**).

This type of flow, in which the pressure gradient and Coriolis deflection are balanced, is called **geostrophic** flow, and the moving air or water masses are geostrophic winds or geostrophic **currents**. The most important features of geostrophic winds or currents are that the flow is directed along contours of equal pressure within a pressure gradient, and that wind or current speed is determined by the steepness of the pressure gradient.

Geostrophic flows are almost never aligned exactly along the contours of equal pressure on the pressure gradient. The reason is that pressure gradients are continuously changing in response to changes in the factors that create them, such as wind stress on the oceans and **convection** in the atmosphere (**CCS3**). In addition, **friction** between moving air masses and the ground or ocean surface, or between ocean water and the seafloor, reduces the Coriolis deflection. Thus, geostrophic winds at the surface and currents near the seafloor do not flow exactly along contours of equal pressure. Instead, they are aligned generally along these contours but offset toward the center of low-pressure zones and away from the center of high-pressure zones.

Horizontal pressure gradients are usually shown as maps of lines of equal pressure called **isobars**. Most **weather** maps in newspapers and on television are of this type. The isobaric maps in **Figure CC13-2**, on the following page, show high- and low-pressure zones and the pressure gradients between these zones. The pressure difference between adjacent isobars is the same for all adjacent isobars at all locations on each map. The gradient is steeper where the isobars are closer together. Therefore, the spacing of the isobars reveals the steepness of the pressure gradient (the change in pressure per unit distance).

Most upper-air wind and ocean current systems are geostrophic, and their flow direction is nearly parallel to the pressure contours. As a result, pressure contour maps can be used to estimate both the speed and direction of winds or currents. The wind or current direction is parallel to the isobars, and the speed can be estimated from the pressure gradient. Because the Coriolis deflection is to the right in the Northern Hemisphere, winds and currents in this hemisphere flow counterclockwise around low-pressure zones (**Fig. CC13-2a**) and clockwise around high-pressure zones (**Fig. CC13-2b**). In the Southern Hemisphere, the deflection is to the left, and winds and currents flow clockwise around low-pressure zones (**Fig. CC13-2c**) and counterclockwise around high-pressure zones (**Fig. CC13-2d**). The wind or current speed is determined by the steepness of the pressure gradient. Accordingly, wind or current speeds are higher where isobars are closer together and slower where they are more widely separated (**Fig. CC13-2**).

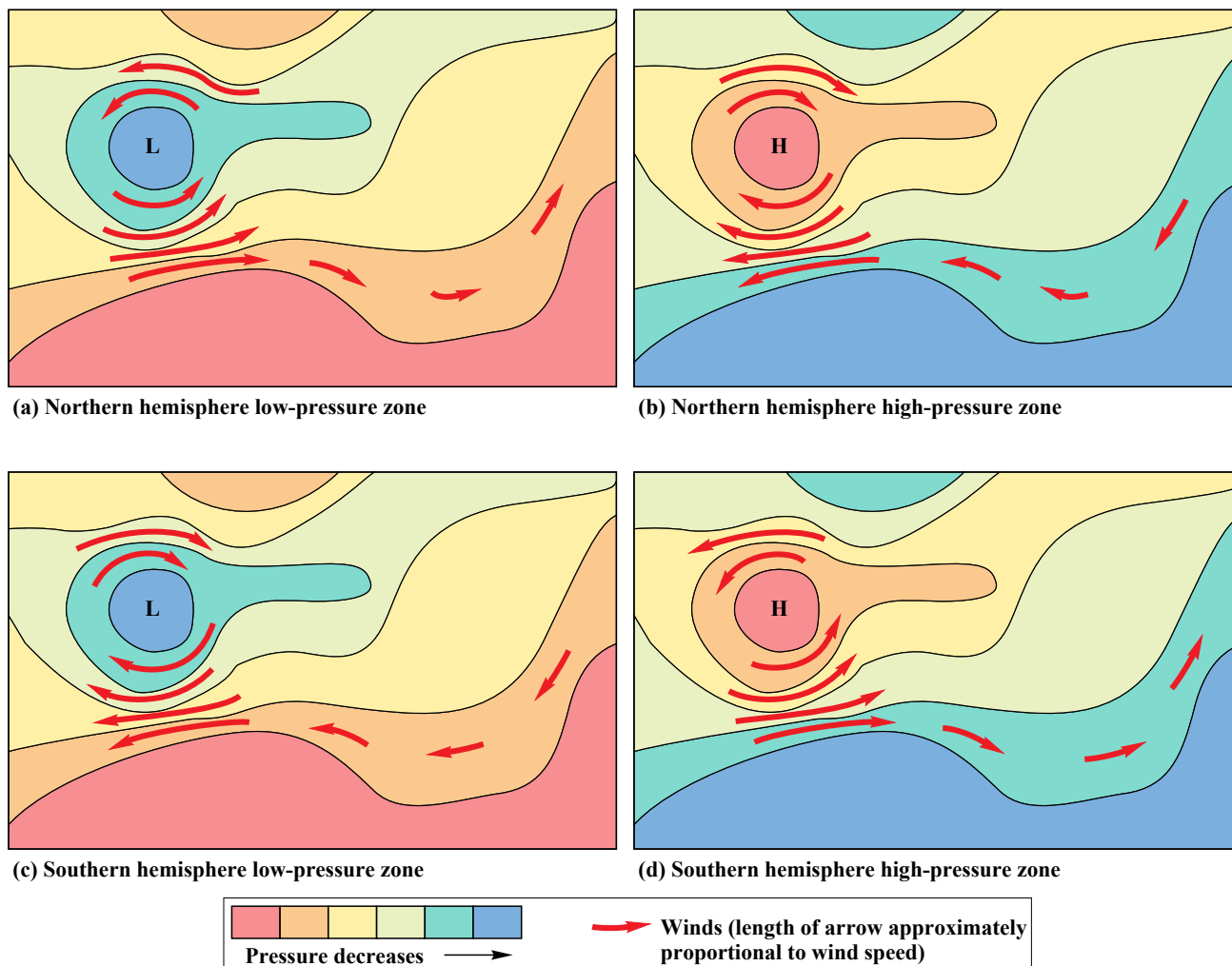


FIGURE CC13-2. Isobaric contour maps can be used to deduce the wind (or current) speeds and directions associated with the isobars. The contour map examples in this figure show circulation at (a) a low-pressure zone in the Northern Hemisphere, (b) a high-pressure zone in the Northern Hemisphere, (c) a low-pressure zone in the Southern Hemisphere, and (d) a high-pressure zone in the Southern Hemisphere. Note that the directions of rotation are opposite for the two hemispheres. Winds blow counterclockwise around a low-pressure zone and clockwise around a high-pressure zone in the Northern Hemisphere, and in the reverse directions in the Southern Hemisphere.

CRITICAL CONCEPT 14

Phototrophy, Light, and Nutrients

ESSENTIAL TO KNOW

- Organic matter is created from inorganic substances (primary production) by phototrophy and chemosynthesis.
- Phototrophy is performed by organisms that use light as their energy source to fuel the primary production process. There are two known mechanisms: photosynthesis and a recently discovered metabolic process that uses compounds called rhodopsins to capture light energy.
- Most primary production in Earth's ecosystems is created by oxygenic photosynthesis, but too little is known about the use of the rhodopsin-based mechanism in ocean ecosystems to assess its relative importance.
- Plants and algae are eukaryotes that perform oxygenic photosynthesis through a complex series of reactions that occur in pigmented cells called "chloroplasts."
- Photosynthesis originated and still takes place in certain Prokaryotes called Cyanobacteria. Chloroplasts are thought to have evolved from photosynthetic symbiotic cyanobacteria that became a permanent resident in a eukaryotic cell.
- A nitrogen-containing pigment called "chlorophyll a" plays the central role in transferring light energy to the oxygenic photosynthesis process. Chloroplasts contain accessory pigments that capture light energy and transfer it to chlorophyll a.
- Accessory pigments collect light energy at wavelengths at which chlorophyll a absorbs light poorly. Each species has a different suite of accessory pigments optimized to collect light energy available in its specific habitat.
- Oxygenic photosynthesis uses light energy, carbon dioxide, and water to produce oxygen (which is released)

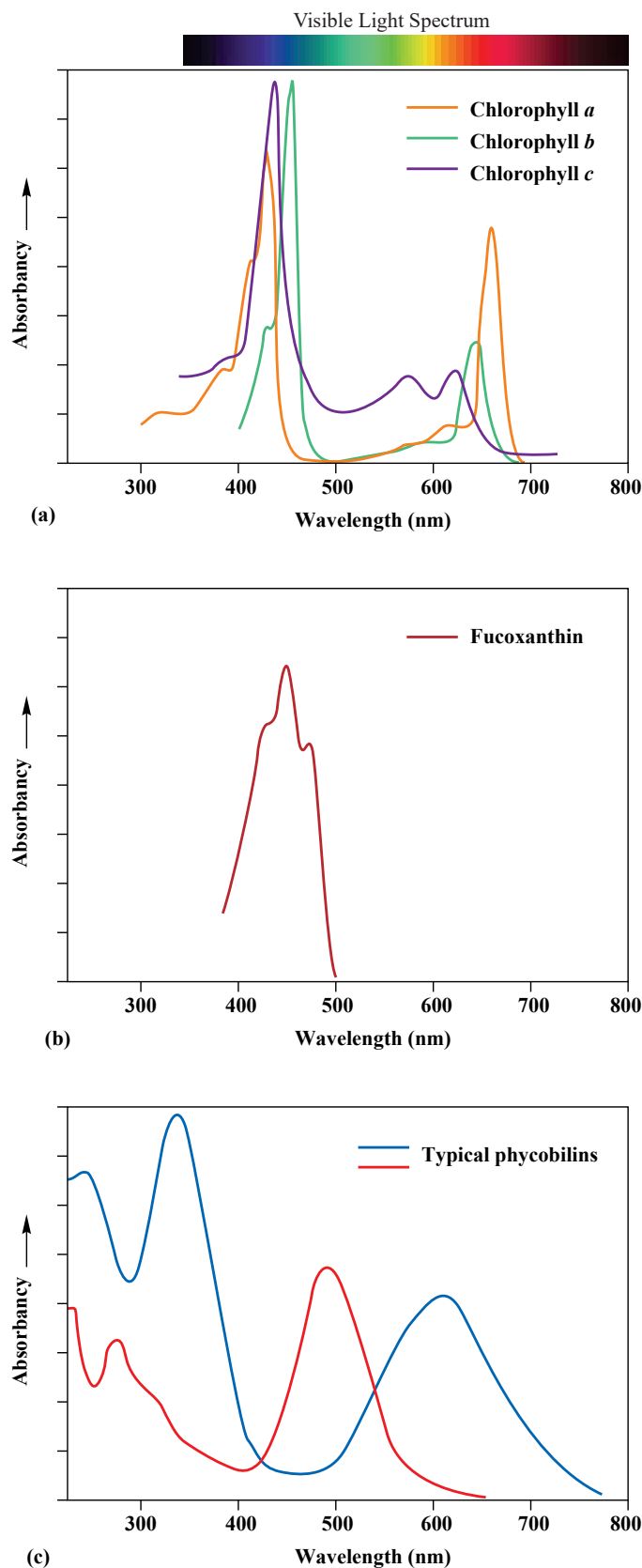


FIGURE CC14-1. Absorption spectra of photosynthetic pigments. (a) Chlorophylls *a*, *b*, and *c*. (b) Fucoxanthin (a typical xanthophyll). (c) Typical phycobilins. Note that the other pigments absorb energy in parts of the spectrum that are not absorbed efficiently by chlorophyll *a*.

and relatively simple organic compounds. Most organic compounds produced by oxygenic photosynthesis are used for cellular respiration. Only a small proportion enters biochemical cycles and is converted to more complex organic molecules.

- Only a very small percentage of the light energy used in photosynthesis is used to produce biomass.
- Nitrogen and phosphorus, among other elements, are present in key compounds involved in photosynthesis. Hence, depletion of available nitrogen and phosphorus stops production of these compounds and limits primary productivity.
- Anoxygenic photosynthesis is an alternate pathway employed by anaerobic bacteria. Anoxygenic photosynthesis uses light energy, carbon dioxide, and a hydrogen donor, such as H_2S or H_2 , to create relatively simple organic compounds and produces a very small amount of organic matter in comparison with algal and cyanobacterial photosynthesis in most present-day ocean ecosystems.
- Chemosynthetic bacteria and archaea synthesize organic compounds by using chemical energy from the oxidation of compounds such as H_2S and H_2 instead of light energy.

UNDERSTANDING THE CONCEPT

For life to exist, carbon, hydrogen, oxygen, nitrogen, phosphorus, and many other essential elements must be synthesized into organic compounds. It is believed that most organic matter on the Earth today was initially converted from inorganic matter to organic compounds by **photosynthesis**. Exceptions include a few simple organic compounds created by electrical discharges in the atmosphere, and the **chemosynthetic communities** described in **Chapters 12 and 15**; the importance of **rhodopsin-based phototrophy** is uncertain.

Photosynthesis is a complex process in which plants, algae, and certain **bacteria** and **archaea** use light energy to convert carbon dioxide into organic compounds. There are two major known photosynthetic pathways: oxygenic and anoxygenic photosynthesis. There are also variations of these pathways, such as those used by certain archaea. Because their photosynthetic mechanisms differ from those of plants and algae, organisms that use these variations are often not considered to use photosynthesis by biologists who prefer to restrict the definition of photosynthesis to the specific pathway used in plants. In this text, we use a broader definition of photosynthesis that includes all pathways used by any organisms that use chlorophylls to capture light energy and carbon dioxide to produce organic matter. Organisms that use light energy to provide sufficient organic matter to generate new biomass are called **phototrophs** or photoautotrophs.

The first step in all variations of photosynthesis is the capture of light energy by the photosynthetic organism. In oxygenic photosynthesis by plants and algae, light is captured in complex cell structures called “chloroplasts” located near the organism’s surface. The chloroplasts contain **chlorophyll *a*** and, in many species, the related pigments chlorophyll *b* and chlorophyll *c*. There are also a number of “accessory pigments,” including carotenoids, xanthophylls, phycobilins, and chlorophyll *d* and chlorophyll *f*. Each photosynthetic species has its own unique combination of pigments. Accessory pigments capture light energy and transfer it to chlorophyll *a*, which then performs the next step in the synthesis. Chlorophylls *b* and *c* can be characterized as accessory pigments because they, too, collect light energy and

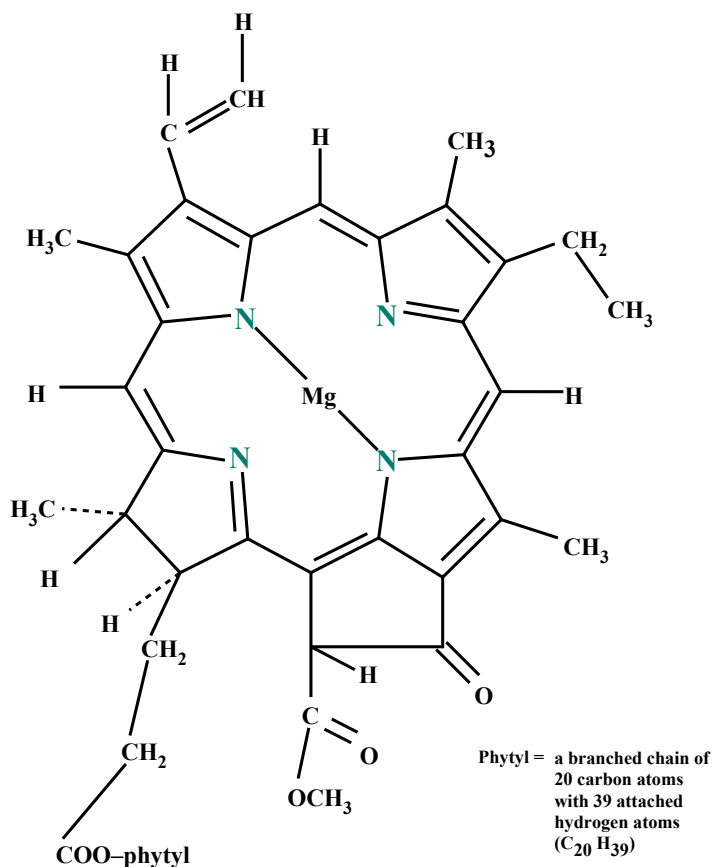


FIGURE CC14-2. The molecular structure of chlorophyll *a*, which is needed for oxygenic photosynthesis. The molecule is complex and contains four nitrogen atoms.

pass it on to chlorophyll *a*. The transfer of energy from accessory pigments to chlorophyll *a* is an extremely efficient process. In some cases, transfer efficiency approaches 100%.

Why are accessory pigments needed? Chlorophyll *a* absorbs light effectively only toward the red and violet ends of the spectrum (**Fig. CC14-1a**). To capture light energy from the central blue–green–yellow part of the spectrum, other pigments are needed. Carotenoids, which include the group of pigments called xanthophylls, absorb energy primarily in the blue and blue-green parts of the spectrum (**Fig. CC14-1b**), whereas phycobilins absorb energy primarily in the green, yellow, and ultraviolet parts (**Fig. CC14-1c**). Chlorophylls *b* and *c* also absorb in different parts of the visible spectrum than chlorophyll *a* (**Fig. CC14-1a**). Chlorophyll *d* and chlorophyll *f* absorb in the infrared wavelengths. Thus, the chloroplast contains an array of accessory pigments to capture as much of the available light energy as possible. Accessory pigments transfer light energy that they capture to chlorophyll *a*, which performs the first steps in photosynthesis.

Red and blue-violet light are absorbed more effectively by seawater than are green and blue light (**Fig. 5-17**). Hence, accessory pigments are especially important in marine algae that must photosynthesize below the immediate surface layer. **Absorption spectra** of seawater in the open ocean and coastal waters are different (**Fig. 5-18**), and the **wavelength** distribution of light changes with depth. Species of marine algae that live at different locations and depths have different accessory pigment compositions to optimize their capture of available light energy.

Photosynthetic cyanobacteria also use chlorophyll or a closely related pigment, bacteriochlorophyll, to perform oxygenic pho-

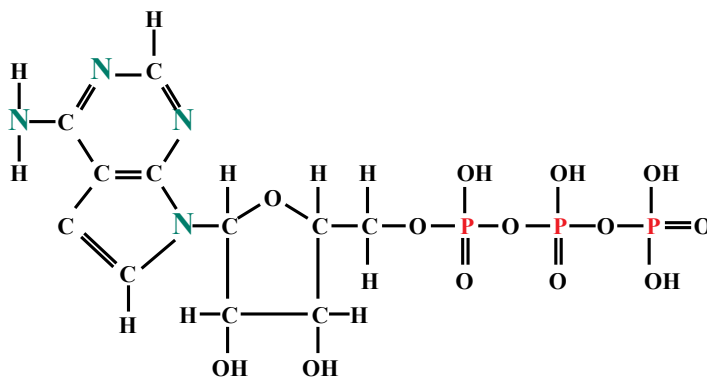


FIGURE CC14-3. The molecular structure of ATP, a chemical essential to life, is complex. The molecule contains both phosphorus (P) and nitrogen (N) atoms.

tosynthesis, but they do not have chloroplasts. Photosynthesis is believed to have originated in cyanobacteria and, over a period of about 1 billion years. They transformed Earth's atmosphere from one with almost no oxygen to the oxygen-rich atmosphere that now exists. Eukaryotes developed only toward the end of this transformation process, and chloroplasts are thought to have originated from a symbiotic cyanobacteria that became permanently located within its host Eukaryote cells.

Some archaea (and perhaps bacteria) use a complex protein, such as archaerhodopsin or halorhodopsin, to collect light energy and produce organic compounds for their metabolic needs, so these organisms employ phototrophy, but they are not photosynthesizers. At present, it is not known how important this rhodopsin-based phototrophy is in the overall primary production in the oceans, but its contribution is thought to be comparatively small. In contrast, cyanobacteria are estimated to produce about one-half of all ocean primary production. They do so because they appear to be the principal primary producers in the large areas of the oceans that are nutrient-limited. Thus, although their population and primary productivity per unit area of oceans are low, their total numbers and total primary production rivals the total primary production by algae and plants in the ocean.

The key molecule in oxygenated photosynthesis, chlorophyll *a* (**Fig. CC14-2**), is a complex molecule that contains not only carbon (C), hydrogen (H), and oxygen (O) atoms, but also a magnesium atom (Mg) and, most importantly, nitrogen atoms (N). The other forms of chlorophyll and rhodopsins also contain nitrogen. When dissolved nitrogen is depleted in ocean waters, nitrogen-containing compounds, including chlorophyll *a*, cannot be synthesized by algae and **primary production** is limited or stopped (**Chap. 12**). Thus, nitrogen as nitrate (NO_3^-), nitrite (NO_2^-), or ammonia (ammonium ion $[\text{NH}_4^+]$ and NH_3) is essential to photosynthesis. The exception to this is that certain cyanobacteria can convert molecular nitrogen to ammonia, although the conversion process is energy intensive and takes place very slowly, especially in oxygenated environments such as most ocean water.

After chlorophyll *a* has captured light energy, it can transfer an excited electron (an electron to which excess energy has been added) to other molecules, which transfer the energy through yet other molecules in a complex series of steps. During these transfers, the raw materials of photosynthesis—water and carbon dioxide—are brought together and converted into organic compounds.

Some reactions in photosynthesis occur only in the presence

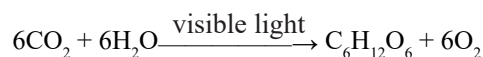
of light. The complex process of oxygenated photosynthesis can be simplified and summarized as follows. Light energy is used to split a water molecule into a hydrogen atom with an excited electron and a hydroxyl ion, each of which immediately reacts with other molecules in the chloroplast. The hydroxyl ion is then converted to oxygen and water, which are released, and the hydrogen atom with its excited electron is combined with NADP (nicotinamide adenine dinucleotide phosphate) to form NADP-H. At the same time, additional light energy is transferred to ADP (adenosine diphosphate) to form ATP (adenosine triphosphate) (Fig. CC14-3). ATP is a versatile energy-storing molecule that can donate the considerable energy associated with one of its phosphate bonds by releasing or transferring this phosphate group, which changes ATP back into ADP. ATP is an energy provider in both photosynthesis and **respiration**. Both ATP and ADP contain phosphorus, which is therefore also an essential **nutrient** for photosynthesis.

Once light energy is stored in ATP and NADP-H, the remaining steps in photosynthesis can continue in the dark. However, if light is removed for an extended period, the store of these molecules in the chloroplast is used up. The photosynthesis steps that can occur in the dark use hydrogen from NADP-H and energy from ATP to reduce carbon dioxide. Oxygen is removed from carbon dioxide and released, whereas the carbon is combined with other carbon atoms to form organic compounds.

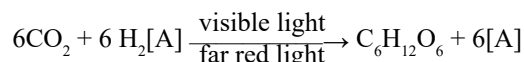
There are several different pathways for these reactions, each producing a different organic compound or compounds. One of the most common pathways leads to the production of glucose ($C_6H_{12}O_6$). Glucose and other organic molecules formed by photosynthesis are subsequently used by the organism as basic building blocks for the vast array of other organic compounds produced by living organisms in their **biochemical** cycles. Large

amounts of glucose and the other compounds created by photosynthesis are also converted back into carbon dioxide in the respiration processes of the plant. The overall efficiency of transferring light energy through photosynthesis to create new **biomass** is extremely low, a few percent at most.

The process of oxygenic primary production can be depicted by the following, much simplified equation:



Cyanobacteria photosynthesize in a somewhat different way. First, cyanobacteria contain a unique form of chlorophyll called “bacteriochlorophyll.” In addition, they do not use water to provide hydrogen for photosynthesis, and they do not release oxygen. Cyanobacterial photosynthesis can be depicted by the following simplified equation:



$H_2[A]$ is a hydrogen donor and can be H_2S , H_2 , or various organic compounds. Cyanobacterial photosynthesis does not make a major contribution to ocean primary production in most ocean areas, because **environments** where both light and a suitable hydrogen donor are available are severely limited.

In cyanobacterial photosynthesis, the needed energy is provided by light. However, it is noteworthy that many of the molecules that replace water as the hydrogen donor in cyanobacterial photosynthesis can also combine with oxygen to provide the energy needed to synthesize organic compounds in chemosynthesis (Chap. 12).

CRITICAL CONCEPT 15

Food Chain Efficiency

ESSENTIAL TO KNOW

- All organisms require food for respiration, growth, and reproduction.
- All organisms lose some of the food as waste products.
- The rate of biomass production by an organism that is available as food for the next trophic level equals the total amount of food ingested, minus the proportion used for respiration and reproduction, and the proportion lost as waste products.
- The percentage of ingested food converted to new biomass is the food chain (or trophic) efficiency. It is usually approximately 10% (ranging from 1% to 40%), regardless of species or food source.
- Approximately 90% of the available food biomass is used in respiration and reproduction or lost as waste products at each trophic level in a food chain.
- Long marine food chains that lead to consumers at higher trophic levels, such as tuna, utilize primary production hundreds of times less efficiently than short marine food chains, such as those that support anchovies or sardines, or most terrestrial food chains used to produce food for human consumption.

UNDERSTANDING THE CONCEPT

All organisms require food, which they can obtain by performing **photosynthesis** or **chemosynthesis** or by ingesting organic matter. In all species, food is distributed and used in four ways: (1) to provide energy through respiration for the life processes of the organism, (2) to provide the basic materials for production of the myriad compounds contained in any organism's body, (3) to provide the basic materials needed to produce offspring (including eggs and sperm), and (4) as waste products that are **excreted**. Hence, not all food is incorporated into new body tissues or **biomass**. The amount converted to new biomass equals the total food intake minus the amounts used for **respiration** and reproduction and lost as waste. Food used for reproductive processes does create new biomass, but the vast majority of eggs, **larvae**, and juveniles of most species are consumed by predators. Consequently, most new organic matter produced during reproduction does not contribute to the species' adult biomass.

The percentage of food ingested by a particular species and converted to new biomass is often referred to as **food chain efficiency** or “trophic efficiency.” The total biomass of a species can be estimated if the total amount of available food and the

food chain efficiency are both known. For example, studies of Loch Ness have determined that it would be virtually impossible for a viable population of “monsters” to live there, because the total amount of food available in the loch would not be enough to sustain even one very large animal, unless it converted all available food into new biomass at an impossibly high food chain efficiency.

The efficiency of converting food to biomass is an important parameter in managing food supplies for human populations. For example, much of the corn fed to beef cattle is used by the cattle in respiration, reproduction, and waste generation. Hence, although beef has a higher protein content than corn and therefore is a desirable food, the same amount of corn could feed more people if eaten directly than if first fed to cattle, which are then eaten. This concept is particularly important in many parts of the world where famine persists, but where cattle are raised on corn and other grains that could be consumed directly by people.

Terrestrial **food chains** used for human food are generally very short. Humans consume primarily plant material, which is at the **primary production** or first **trophic level**, or plant-fed animals, which are at the second trophic level. Ocean food chains that lead to human foods are, in many cases, much longer. For example, tuna generally feed at the fifth trophic level (**Fig. 12-11**). The overall efficiency of the food chain from primary production (**phytoplankton**) to tuna is very low because each step is inefficient.

Food chain efficiency is difficult to measure precisely because it varies with many factors, such as the amount of food available, food quality, the nature and timing of the reproductive cycle, and the level of physical activity required to avoid predators. As an example of this variability, consider that two human children of the same height who ingest approximately the same amount of food, but one may be much thinner than the other. We might conclude that the thinner child has a “faster metabolism.” What we mean is that the thinner child uses more food for respiration (because of either greater physical activity or a genetic disposition to utilize food less efficiently in respiration) or that the thinner child

sends a greater proportion of food to waste (because of **genetic** differences in the ability to digest and assimilate food). Of course, if the food ingested by these two children was not equal in quantity, the thinner child could also be thinner simply because of a limited food supply and eating less. Another illustration of the variability of food chain efficiency is the difference in efficiency between children, whose biomass is increasing, and adults, whose biomass stays relatively constant, even though they may ingest approximately the same amount of food as children.

Food chain efficiency is variable. It has been determined that, if averaged over a population and over time, this efficiency ranges from about 1% to 40% and, in most instances, is approximately 10% between any two trophic levels. Hence, on average, all species above the primary producer trophic level convert only about 10% of their food into new biomass, which then is available to be consumed at the next trophic level. The following relationships summarize marine food chains:

- 1 kg of biomass at trophic level 1** (phytoplankton)
- produces 0.1 kg of biomass at trophic level 2** (zooplankton)
- produces 0.01 kg of biomass at trophic level 3** (e.g., small fishes, baleen whales)
- produces 0.001 kg of biomass at trophic level 4** (e.g., larger fishes such as mackerel, squid)
- produces 0.0001 kg of biomass at trophic level 5** (e.g., predatory fishes such as tuna)
- produces 0.00001 kg of biomass at trophic level 6** (e.g., killer whales)

In this example, certain types of organisms are designated as feeding at a particular trophic level. However, some of them, such as killer whales, may feed at more than one trophic level.

Many seafood species captured and eaten by people feed at a much higher trophic level than terrestrial animals used as human food. Consequently, human use of ocean primary production is grossly inefficient compared with terrestrial primary production.

CRITICAL CONCEPT 16

Maximum Sustainable Yield

ESSENTIAL TO KNOW

- Fishing initially reduces the size of a fish stock. However, because the reduction results in greater food availability for the fishes remaining, the rate of production normally increases.
- The additional biomass produced represents an excess over that needed to maintain the population. This excess can be harvested safely.
- As fishing increases, stock size is reduced to a critical level known as the “maximum sustainable yield,” at which the production and reproduction rates of the population are just sufficient to balance the removal rate from predation and fishing. If fishing yield is increased and continued beyond the maximum sustainable yield, the population can no longer sustain itself and collapses.
- Maximum sustainable yield is difficult to establish because fish stocks vary as a result of year-to-year climate-induced changes and changes caused by other factors, such as diseases.
- Maximum sustainable yield also depends on the age structure of the population and on the degree of age selectivity in fishing methods used. Harvesting older fishes tends to increase the sustainable yield because the remaining younger fishes are faster-growing, but it also tends to reduce the breeding population because the younger fishes are sexually immature.
- Maximum sustainable yield is usually established by using one year’s data for stock size and reproductive success to project the survival of adults and young into the next year. Unexpected events, such as disease outbreaks, can render such estimates inaccurate and inadequately protective.
- Most fisheries are managed at a yield 20% to 40% below

the estimated maximum sustainable yield to allow a safety margin. This safety margin may not always be adequate, but any safety factor means that fishes that could be harvested will not be.

UNDERSTANDING THE CONCEPT

Although fisheries and **shellfisheries** provide only a small fraction (about 1%) of the world's total human food supply, they provide a much larger percentage of its required protein. In many areas, the oceans are the only significant source of protein.

The world's total fish catch is currently approximately 90 million tonnes per year, about one-third of the estimated global annual fish production. Some optimistic biologists believe that the global fishery catch could be raised by a factor of as much as 10. These optimistic estimates are based on the assumption that existing fisheries can be exploited to the maximum possible extent, that many new stocks of fishes will be discovered, and that many currently underutilized exotic **species**, including **invertebrates** such as **sea cucumbers** and **sea urchins**, will be fully utilized. Unfortunately, about 85% of major marine fish stocks are currently depleted, overexploited, or fished at their biological limits. Consequently, the global total fish catch has begun to decline from its previous maximum level. Many fisheries have been closed or severely restricted in attempts to reverse this damage.

The oceans do not appear to have the potential to help solve the world's food supply problem. However, if fisheries are to continue sustaining human populations in areas that have historically depended on them, and if seafood is to continue supplying the same proportion of human food, the world's fisheries must be managed carefully. The goal of such management must be to catch the maximum amount of seafood that can be taken from the oceans without damaging individual species or marine ecosystems. To fully meet this objective, we would need to harvest species selectively at low **trophic levels** instead of the currently consumer-desired species, such as tuna. Furthermore, we would have to learn to farm the sea as we do the land, eliminating undesirable plants and animals from ocean farms.

Neither approach is likely to be fully acceptable, at least in the foreseeable future. Consequently, the principal approach of fishery management, which will probably continue for many years, is to manage the fishing of each individual species that is targeted by fishers. The specific objective is to maximize the total amount of the species caught (the yield) while ensuring that the **standing stock** does not decline to levels that cannot sustain the yield in the future. Thus, the goal is to manage each species to ensure the **maximum sustainable yield**.

To determine the maximum sustainable yield, we must have a good understanding of the species' life cycle. Consider how fish stocks respond to fishing. The stock (or **biomass**) of a fish species is limited primarily by its food supply and its predators. If human or other predators harvest more of the species than is normally taken by its natural predators, more food will be available for others of the species. If excess food is available, the species will reproduce and grow to use this food until the population is again at a size where food availability limits further growth. Exceptions occur when the excess food is consumed by competitor species, but in most cases, if we harvest a species and reduce its stock somewhat, the total production of the species will increase. If we harvest an amount equal to this increased production each year, the population will remain stable, but at a lower number than before we began to harvest. If we harvest more than the increase

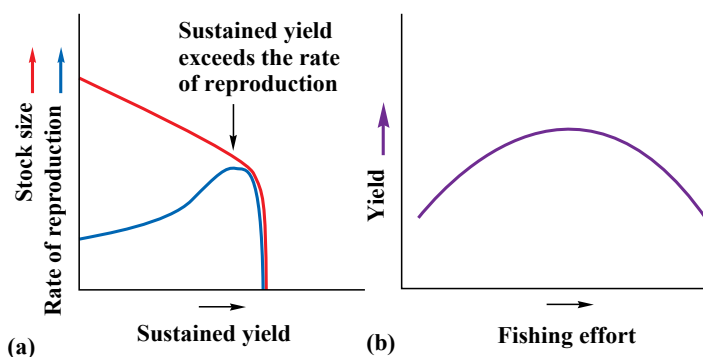


FIGURE CC16-1. Conceptual model showing (a) the changes in fish stock size with increasing fishing, and (b) fishing yield as a function of fishing effort. Note that moderate fishing can normally produce high sustained yields because the rate of reproduction initially increases as the stock size is reduced and more food becomes available to younger fishes. However, if the fishing effort and yield are increased further, the stock size is reduced to such a level that there are no longer enough individuals to sustain a high rate of reproduction, and the stock collapses.

in production, the stock will progressively decline.

As a stock declines with increased fishing, a critical level, generally 35% to 70% of the original stock, is reached at which the remaining stock becomes so small that it is barely able to grow and reproduce fast enough to replace fishes removed from the population by predation and fishing. If fishing continues to increase beyond this critical point, the stock will begin to decline precipitously (**Fig. CC16-1a**), and unless the catch is reduced, the stock will collapse to a very low abundance and will not recover. The critical level at which the species is just able to replace the stock lost to fishing is the maximum sustainable yield.

One of the most important characteristics of a fishery is the "fishing effort," which is the number of boat or person days of fishing expended. As a typical fishery develops, fishing effort increases as more boats and fishers target the resource species. At first, yield increases rapidly, but as the maximum sustainable yield is approached, fishing effort increases faster than yield because stock size is reduced (**Fig. CC16-1a**). Each boat must fish longer to catch the same amount of fish. Consequently, either the cost of each fish caught rises, or the catch and income of each fisher declines. Therefore, for economic reasons, the optimum harvest level of a species may be well below its maximum sustainable yield. Fishery management often controls fishing effort to control yield (**Fig. CC16-1b**).

If the maximum sustainable yield is exceeded in an uncontrolled fishery, fishing effort often rises dramatically as stocks decline. Fishers targeting the species try to protect their livelihoods by increasing their efforts in an attempt to maintain their individual historical catch levels. This dynamic occurs both with technologically advanced fishers targeting a regional resource, such as North Atlantic cod, and with subsistence fishers in island communities, whose growing populations lead to increased fishing effort for local **reef** fishes.

Changes in the age structure of populations subject to fishing also affect maximum sustainable yield because age structure influences the population's reproductive capacity. Like children, young fishes increase their biomass with time faster than adults do, even if they consume similar amounts of food. If larger adult fishes of a particular stock are preferentially targeted by fishers, the excess food and decreased competition from larger fishes will enable a greater number of young fishes to grow successfully.

Because young fishes gain weight or biomass faster than adult fishes do, a greater number of young fishes in the population will cause an increase in the rate of biomass production, even if the available food supply remains the same. If large fishes are preferentially removed, the stock gains more biomass per year, and the maximum sustainable yield is increased. Hence, fishery management commonly attempts to control the size of harvested fishes by establishing minimum mesh sizes for nets, requiring the release of captured small individuals, or other methods.

Although it is generally advantageous to reduce the average age of a fish population to increase the maximum sustainable yield, the stock can be adversely affected if too many large fish are removed. Only the larger fish of most species are of reproductive age. Therefore, reducing the number of large fishes also reduces the breeding population.

Clearly, establishing maximum sustainable yield is a difficult task that requires knowledge of stock size, age structure, and reproductive process for each species. It is further complicated by the natural year-to-year variability of fish populations and the requirement that maximum sustainable yield be established before or early in a given year. The natural variability of fish populations from year to year is very large, at least for some species, due to **climate** variation, disease, and other factors. For example, a relatively small change in water temperature may alter the timing of a **phytoplankton bloom**. In turn, this change may cause an entire year class of **larvae** to die if they depend on the timely availability of this food. In addition, because diseases are present in fish populations, the equivalent of epidemics can occur and decimate

the stock. Therefore, fish stocks can vary dramatically from year to year and on even shorter timescales, and the maximum sustainable yield will vary accordingly.

If the size and condition of the stock were continuously monitored and known, the maximum sustainable yield could be continuously adjusted to accommodate changes. However, data cannot be gathered and analyzed quickly enough to do this. Consequently, maximum sustainable yields generally are estimated from the previous year's data. Estimates of the stock of adults and number of juveniles that enter the population are made each year. In addition, estimates are made of the survival rates of both adults and juveniles through the coming winter. These data are used to project what the stock will be in the following year, and the projection is then used to estimate maximum sustainable yield and set fishing limits for the following year. If an unexpected event occurs that adversely affects the population after the maximum sustainable yield has been estimated and fishing limits established, the permitted yield may be high enough to damage the stocks. If the unexpected decrease in the stock is recognized early enough in the fishing season, emergency measures can be taken to reduce fishing efforts. However, the stock size is often not well known until after the fishing season ends.

To account for uncertainties and variability in maximum sustainable yield, most fisheries set the permissible catch 20% to 40% below the estimated maximum sustainable yield. This practice leads to conflicts because some people feel that this safety margin is not enough to ensure protection of the stock, and others feel that part of the resource is being wasted because the maximum sustainable yield is not fully used.

CRITICAL CONCEPT 17

Species Diversity and Biodiversity

ESSENTIAL TO KNOW

- **Species diversity** is a well-defined scientific term that expresses a combination of species richness (number of species) and evenness (degree to which the community has balanced populations with no dominant species).
- **High species diversity** is generally equated with healthy ecosystems, but there are exceptions.
- **Biodiversity** is a poorly defined term that refers to a combination of genetic diversity (genetic variation within populations of a species), species diversity (species richness and evenness within a community), ecosystem diversity (variations in communities of species within an ecosystem), and physiological diversity (variations in feeding, reproduction, and predator avoidance strategies within a community or ecosystem).

UNDERSTANDING THE CONCEPT

The term **biodiversity** entered common usage when the global community of nations recognized the need to preserve species from extinction and unique **habitats** or **ecosystems** from destruction. The term has no precise definition but is based on the much more precisely defined technical term **species diversity**, which has been in use for many years.

Species diversity is a measure of species richness and evenness. *Richness* describes the number of individual species,

whereas *evenness* expresses the degree to which a community has balanced populations in which there is not a small number of numerically dominant species. Several statistical indices have been developed to express both species richness and evenness as a single species diversity number.

Species richness is important because the larger the number of species present in a community, the more robust the community is considered to be. If there are many species, a community disturbance (such as a change in temperature or **salinity**) may lead to the exclusion or even extinction of one or a few species, but most species are likely to survive. Evenness is important because if the community is dominated by only a few species and these species are excluded or rendered extinct by a disturbance, competition between the remaining species to become the new dominant species may cause severe ecosystem instability. Alternatively, an uneven community may reflect effects of a disturbance, such as **pollution**, that favors dominance by tolerant species.

Although communities are generally thought to be desirable and stable if they contain many species and have strong evenness, those characteristics may not always be ideal. For example, **coral reefs** that are undisturbed by storm wave damage for many decades become dominated by just a few species of hard corals, and most other hard coral species are excluded by competition with these dominant species. Thus, the beautiful, untouched hard

coral communities present in a few sheltered locations have low species richness, and species are unevenly distributed. These reefs have low species diversity (at least of corals).

In contrast, a coral reef that was damaged several years earlier by a **hurricane** has more species of corals (and possibly other species) and greater evenness, or higher species diversity. Few of us would consider this damaged **environment** an ideal situation, but the periodic disturbance may be necessary to restore and preserve the diversity of coral reefs since high species diversity is generally considered a positive attribute of any ecosystem.

During the hundreds of millions of years that life has existed on the Earth, virtually all the species that have lived have become extinct and been replaced by others. Because this natural process still continues today, none of the species now on the Earth are likely to be living a few million years in the future. Human disturbances, including pollution, have caused many species to become endangered or extinct. The rate of species extinction due to human disturbances is estimated to be many times faster than natural extinction, and possibly is much faster than the rate at which new species can evolve to replace those that are lost.

Only recently have scientists begun to realize that probably the best way to control the extinction rates of species is to control losses of their habitat and disturbances of their community structure. In response to the need to consider habitat and community structure in protecting individual species, a new concept was developed called “biodiversity.” Biodiversity is still often interpreted as species richness, but it should include many attributes of natural ecosystems. For example, biodiversity can be separated into four components:

- Genetic diversity, or variation in the genes within a species or population. High genetic diversity is thought to maximize the potential for new species development. It is also thought to maximize the potential for species survival when the species is subjected to an environmental disturbance because some members of the species may be more resistant.
- Species diversity, or species richness and evenness. High species diversity is thought to maximize the stability of the ecosystem and its resistance to environmental disturbances.
- Ecosystem diversity, or variation in the communities of species within an ecosystem. High ecosystem diversity is thought to reflect the availability of a wide variety of ecological niches. If the range of niches is large, environmental disturbance is unlikely to alter the ecosystem in such a way that more than a few species lack a niche within which to survive.
- Physiological diversity, or variation in physiological adaptations to feeding, reproduction, and predator avoidance within a community or ecosystem. High physiological diversity is thought to reflect a greater ability of the community within an ecosystem to adjust to environmental disturbances and maintain its stability.

A global agreement was established in 1993 to preserve biodiversity. Because biodiversity is a poorly defined term that encompasses many different characteristics of organisms in their living environment, this agreement will continue to be very difficult to translate into management actions. Protection of biodiversity may mean very different things to different nations or individuals.

CRITICAL CONCEPT 18

Toxicity

ESSENTIAL TO KNOW

- A chemical is toxic if it can cause death or adverse sublethal effects in organisms exposed to it at a concentration above a critical threshold.
- Chemicals that are toxic can have many possible sublethal adverse effects on organisms, such as inhibition of the ability to photosynthesize or feed. The most important sublethal effects appear to be those that interfere with reproductive success.
- All chemicals that can be toxic have a concentration threshold below which they have no sublethal or lethal toxic effects. Hence, for every chemical in the ocean environment, there is a concentration below which it is environmentally safe.
- Many substances that are toxic at high concentrations are also essential to life, and the growth of marine organisms may be inhibited if they are not present at concentrations above a certain threshold. For these substances, there is an optimum range between the minimum concentration that an organism requires to meet its needs and the concentration above which the substance is toxic.
- The range of optimum concentration can be large or small. The concentration at which toxicity occurs can be many times higher than or very close to the natural range.
- Anthropogenic inputs of potentially toxic substances can be assimilated safely in the oceans if the amount introduced does not cause concentrations to exceed the threshold at which sublethal toxicity occurs. The quantity that can be safely assimilated is different for each substance and is determined in part by its sublethal toxicity threshold and background concentrations.
- Sublethal or lethal toxicity threshold concentrations are difficult to determine because they vary among species, among substances, and with other factors, such as physical stresses and synergistic and antagonistic effects of other chemical constituents.
- Marine organisms bioaccumulate most toxic substances. Bioaccumulation occurs when the concentration in the organism is higher than the environmental concentration, but the concentrations are in equilibrium.
- A few toxic substances are biomagnified in marine organisms. Biomagnification occurs when the organism retains all the toxic substance to which it is exposed in its food or environment and does not lose any of the substance, even if its environmental concentration decreases.
- Carcinogenicity, mutagenicity, and teratogenicity can be considered to be lethal or sublethal effects, but there is probably no concentration threshold below which there is

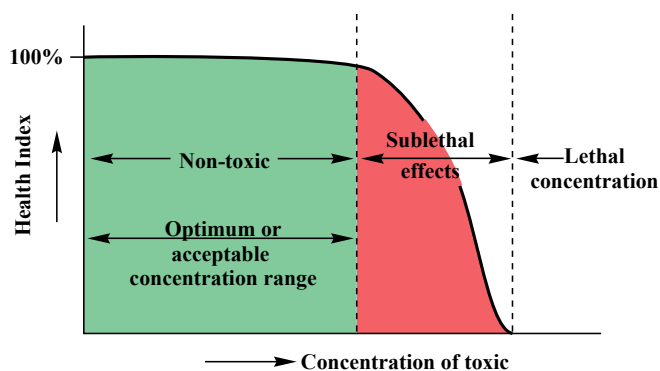
no effect. Each exposed organism has a small probability of suffering an effect at any specific concentration. More individuals within a population suffer the effects as the concentration increases.

- **Carcinogenic, mutagenic, and teratogenic substances occur naturally. Anthropogenic inputs will increase the incidence of the effects of such substances. However, at least for some of these substances, anthropogenic inputs would need to be large before the increased incidence would be significant or measurable in comparison with the natural incidence of effects.**

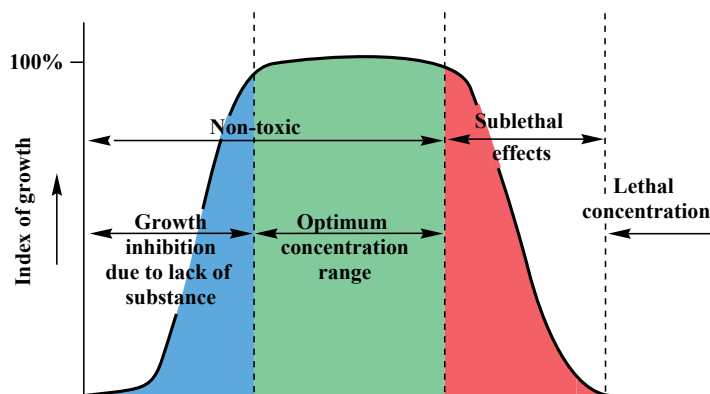
UNDERSTANDING THE CONCEPT

A chemical is toxic if it is present at high enough concentrations to cause death or adverse sublethal effects in organisms, including human beings, that are exposed to the substance. If an organism is exposed to a concentration that is lethally toxic, the organism dies. If an organism is exposed to a concentration that is sublethally toxic, the organism is not killed, but it is disadvantaged in some way.

Examples of the many possible sublethal effects are partial in-



(a) Non-essential element or compound



(b) Essential element or compound

FIGURE CC18-1. Conceptual model of toxicity, showing the general behavior of two different types of toxic substances: (a) elements or compounds that are not essential to life, and (b) elements or compounds that are essential to living organisms for their growth and reproduction, but that become toxic to the organisms if their concentration is too high. The health index plotted on the y-axis for nonessential substances in (a) can be many different factors, such as reproductive success or incidence of disease. For essential substances (b), the index of growth can similarly be such factors as primary production rate or biomass. The concentrations at which transitions occur among growth inhibition, the optimum or nontoxic range, sublethal concentrations, and lethal concentrations vary among species, among life stages (e.g., juveniles and adults), and among toxic substances. They can even vary among individuals of a species or for a single individual, depending on the level of environmental stress that the individual is experiencing as a result of other factors.

hibition of **photosynthesis** in plants and partial inhibition of feeding, hunting, or prey avoidance capabilities in animals. However, the most important sublethal effects are generally believed to be those that reduce reproductive success. Reproductive success can be reduced by partial inhibition of egg or sperm production, reduction in the probability of fertilization when egg and sperm meet, reduction in the survival ability of **larvae** or juveniles, and in other ways. Because of the wide range of possible sublethal effects caused by toxic chemicals, identifying and measuring such effects is often impossible either in the **environment** or in laboratory experiments.

The effects of any chemical that is potentially toxic vary with its concentration (**Fig. CC18-1**) and with individual species. Above a certain concentration, the chemical can cause death. At somewhat lower concentrations, the same substance may produce sublethal adverse effects. Below an even lower concentration, the substance is not toxic as it causes no sublethal effects and does not harm the organism. This concentration is the sublethal toxicity threshold (**Fig. CC18-1a**). Any concentration below the sublethal toxicity threshold could be considered an acceptable concentration. However, this is not true for naturally occurring potentially toxic substances, because some are also essential **nutrients**. These substances have an optimum concentration range below the sublethal toxicity threshold, within which they do not harm the organism (**Fig. CC18-1b**). However, below this optimum range, some essential function is limited by a lack of the substance, and the organism is disadvantaged (**Fig. CC18-1b**).

All substances have an acceptable or optimum concentration range within which they do not harm organisms or the **ecosystem**. Therefore, **anthropogenic** additions of chemicals to the oceans are not always harmful. Such inputs can be safe if the assimilative capacity or acceptable concentration range in the environment is not exceeded. The **assimilative capacity** is determined by the sublethal toxicity threshold of the most sensitive species.

The relationships shown in **Figure CC18-1** generally apply to all chemicals other than **carcinogens**, **mutagens**, and **teratogens**. However, concentrations at which the inhibition of growth ends (for essential nutrients), sublethal toxicity begins, and lethal toxicity occurs can all vary by orders of magnitude for different species, and even for different life stages of a species, exposed to a specific substance.

Most importantly, the optimum concentration range for essential nutrients may vary dramatically. For example, zinc is a toxic but essential chemical that has a wide optimum concentration range. In certain parts of the oceans, zinc may be the nutrient most likely to be depleted and hence most likely to restrict the growth of certain organisms. Because it has a large optimum concentration range, zinc becomes sublethally toxic only at concentrations that are many orders of magnitude above those normally present in the oceans. In contrast, copper has a very narrow optimum concentration range. Like zinc, copper may be growth-limiting to certain species in some parts of the oceans. However, at concentrations only a few times higher than the limiting concentration and within the range often present in **contaminated** coastal waters, copper can reduce the rate of photosynthesis by certain species of **phytoplankton**. Hence, copper is sublethally toxic at concentrations very close to those that occur naturally in the oceans.

Acceptable or optimum concentration ranges and the range of

concentrations present naturally must both be considered in assessments of whether anthropogenic discharges might create **pollution** problems. For example, much smaller quantities of copper than of zinc can be discharged safely to the oceans.

Concentrations that represent the transitions between growth inhibition, nontoxicity, sublethal toxicity, and lethal toxicity are very difficult to determine for several reasons. First, members of a species differ somewhat in susceptibility to a given toxic chemical. Second, different species and life stages of a single species may have dramatically different susceptibilities to a particular chemical. Third, complicated, antagonistic, and synergistic effects occur among potentially toxic substances and other chemicals in the environment. Synergistic effects increase toxicity. For example, high concentrations of one trace metal can increase the toxicity of another trace metal. Antagonistic effects are the opposite; high concentrations of one trace metal may reduce the toxicity of another metal. Fourth, the susceptibility of a species to a chemical may be altered by stress due to other environmental factors, including temperature, **salinity**, and the extent of competition from other species. Fifth, the toxicity of many substances is determined by their chemical form. For example, **ionic** copper is substantially more toxic than copper **complexed** with organic matter.

Because of the variability among chemicals, among species, among species' life stages, among chemical states, and with synergistic, antagonistic, and stress factors, it is not possible to determine a single concentration at which a chemical is lethal or sublethal in the environment. For this reason, toxicologists often measure lethal toxic concentrations of chemicals on selected species in the laboratory by using **bioassays**. They assume that sublethal effects will occur at concentrations below the lethal concentration. They also assume that some species are more sensitive to the toxic chemical than the tested species is. A "safe" concentration in the environment is then estimated. Generally, a concentration two orders of magnitude below the minimum lethal concentration is arbitrarily assumed to be safely below the level at which any sublethal adverse effects occur. However, in some cases, such as copper, the optimum concentration range (**Fig. CC18-1b**) is less than two orders of magnitude below the lethal concentration threshold. Therefore, this approach may set the "safe" concentration below naturally occurring levels and sometimes even below the level needed to sustain growth.

Bioassays are conceptually simple. The test species is usually exposed for 96 h (a duration chosen because one set of tests can be set up and run in a 5-day workweek) to several different concentrations of the potentially toxic substances in seawater. The test species is also exposed to control seawater with none of the potentially toxic substances added. After exposure, the number of organisms that have died or exhibited a sublethal effect (e.g., failure of an **embryo** to develop or of a larval stage to **metamorphose** to the next life stage) are counted. Generally, no death or sublethal effect occurs in the control seawater (otherwise the experiment is not acceptable, because a factor other than the tested contaminant has caused a confounding effect). If the test substance is toxic, some or all individuals will have died or shown the sublethal effect in the test with the highest contaminant concentration. A progressively smaller proportion of the test population is usually affected at lower contaminant concentrations. From these tests, the contaminant concentrations at which one-half of the test organisms are killed or suffer the observed

sublethal effect are calculated. These values are called the LC50 and EC50, respectively.

Bioassay tests must be interpreted carefully because they do not use the most sensitive organism in the ecosystem being assayed. Without testing all species, there is no way to determine which is the most sensitive. In fact, because different life stages of some species have different levels of sensitivity (juvenile stages are generally more sensitive than adults), all life stages of all species would have to be tested. In addition, most marine species are difficult to keep alive in the laboratory because this alien environment stresses the organisms. Species used for bioassays must be stress-tolerant, and hence they are also likely to be tolerant of toxicant stress.

Many factors, such as salinity, temperature, light intensity, water chemistry, **sediment** characteristics, food supply, and competition or cooperation with other individuals or species, are different in bioassays than in the environment. The additional stresses of the test environment may make test organisms more susceptible to toxicant stress.

Bioassays are often conducted with samples of wastes in the form discharged (e.g., sewage or industrial **effluent**). Although this approach takes into account synergism or antagonism between components of the waste, there remains the question of whether additional such effects occur when the waste is mixed in seawater and subject to different conditions in the environment.

Most chemicals that exhibit toxicity are more highly concentrated in tissues of marine species than in the surrounding seawater. This fact has led to a popular belief that toxic chemicals are taken up continuously by marine species and that concentrations of the toxins in their tissues increase progressively throughout their lifetime, regardless of how the concentration changes in their environment. This may be true for some species and some chemicals, but it is not true for most species and most substances. Most potentially toxic chemicals are **bioaccumulated** by most marine species, but are not **biomagnified**.

Bioaccumulation occurs when the concentration of a compound is regulated by equilibria between the organism and its food and/or surrounding water. The substance is taken in by the organism from its food and/or directly from the water, but the organism is capable of **excreting** some of the substance either directly back to the water or through its urine or feces. Hence, if the food or seawater concentration of the substance increases, the concentration in the organism will increase, but if the food or seawater concentration decreases, the concentration in the organism will also decrease. However, it may take some time for equilibrium to be reached.

Certain toxic substances are more dangerous than others in marine ecosystems because they biomagnify. Biomagnification occurs when organisms at each **trophic level** in a **food chain** retain all or almost all of a toxic substance ingested in food (**CC15**). Consequently, the concentration of the toxin increases at each trophic level.

Certain chemicals are carcinogenic (cancer-causing), mutagenic (causing **genetic** changes in the offspring by altering the parental **DNA**), or teratogenic (causing abnormal development of the embryo). Although each can be considered a sublethal effect, many scientists do not believe that these effects follow the toxicity-concentration relationships shown in **Figure CC18-1**, but rather that they may occur at any concentration of these substances.

If this hypothesis is correct, a single molecule of such a chemical could cause cancer in an exposed individual. However, if many individuals were each exposed to a single molecule, only a few (if any) would be expected to develop cancer. If the same group of individuals were each exposed to two molecules of the chemical, twice as many would be expected to develop cancer. This type of effect is said to be “probabilistic.” The probability that any one individual will contract cancer from a given concentration of the chemical is reduced proportionally as the concentration is reduced, but it never reaches zero.

There are many naturally occurring carcinogens, teratogens, and mutagens, but only a very small percentage of marine populations suffer adverse effects as a result of these naturally occurring compounds. Because such a small fraction is affected, great numbers of individuals would have to be monitored to accurately

estimate the percentage of a population that suffers such effects. Furthermore, determining the extent to which a higher incidence of such effects might be caused by human contamination of the oceans, or parts thereof, would be even more difficult.

If the probabilistic-effect hypothesis is correct, any quantity of anthropogenic input of a carcinogen, mutagen, or teratogen will increase the incidence of such effects in marine species. The increase may be extremely small in comparison with the incidence of such effects due to the naturally occurring concentrations, in which case many scientists believe adverse effects on the marine ecosystem would be negligible. Hence, some believe that it would be environmentally safe and acceptable to dispose of limited quantities of anthropogenic carcinogens, mutagens, and teratogens in the ocean.

CRITICAL CONCEPT 19

Metagenomics

ESSENTIAL TO KNOW

- Within oceans, there are diverse microbial biomes, which are ecological niches where microbes interact with one another and their surrounding environment.
- Metagenomics can tell us what organisms are there and what they are capable of doing, but can't tell us if they are actually doing it. Other related techniques are necessary to identify which species and genes are active and how this information impacts our understanding of marine ecosystems and ocean chemistry.
- Bioinformatics is used to assemble and analyze metagenomic data using powerful computers.
- Metagenomic studies have revealed a vast array of microbes in marine environments, many of which have never been identified, let alone cultured in the lab.
- Viruses are present in significantly higher numbers than any other organism in the ocean, and recent metagenomic and related studies have shed light on their impact.

UNDERSTANDING THE CONCEPT

Within the ocean, a wide array of microorganisms actively engage in biogeochemical cycling while interacting with one another as well. The complex interactions exist within smaller ecological niches known as microbial **biomes**. These biomes comprise a diverse group of organisms, including bacteria, algae, viruses, and archaea (**Fig. CC19-1**). It is important not only to understand the role these biomes play in marine ecosystems but also to identify the players involved.

Investigating how microbial niches exist within marine environments has historically been challenging, primarily because many of the microbes found there cannot be cultured (grown in a lab). It was not until the discovery of **DNA sequencing** that we truly began to understand the diversity of microbes that exist in ocean biomes. **DNA**, which is the genetic makeup (code) of an organism and contains all the information needed to perform the life processes of that species, can be sequenced so that every nucleotide that makes up that code can be identified.

Thousands of novel organisms have been identified since DNA

sequencing technology, primarily **metagenomics**, was discovered. Despite extensive efforts to better characterize marine microbes using DNA sequencing, a large portion of microbes in ocean ecosystems have yet to be identified.

Metagenomics is the analysis of all the genomic DNA in an environmental sample that contains multiple species (**Fig. CC19-2**). Metagenomics data provides information on what organisms are present in the sample and what metabolic capabilities they have. This allows scientists to infer what they may be doing in the microbial biome, including their possible involvement in biogeochemical cycling. Time-lapse studies or comparative studies can be used to determine how communities change over time and space, as well as how they react to environmental changes. It is important to note that the limitation of metagenomics is

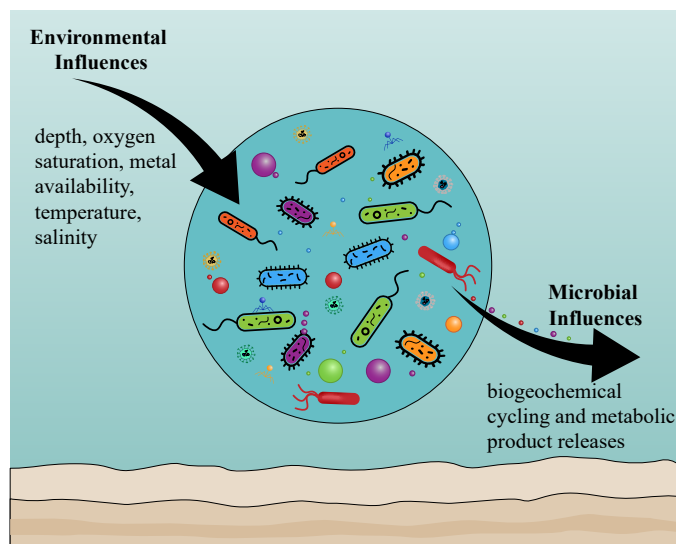


FIGURE CC19-1. Microbial biomes consist of thousands of microorganisms, including viruses, bacteria, and archaea. The community structure is shaped by the chemical and physical elements (sunlight, oxygen saturation, metal availability, temperature, salinity) within the environment. These diverse communities play a critical role in biogeochemical cycling and nutrient availability and can respond rapidly to environmental changes.

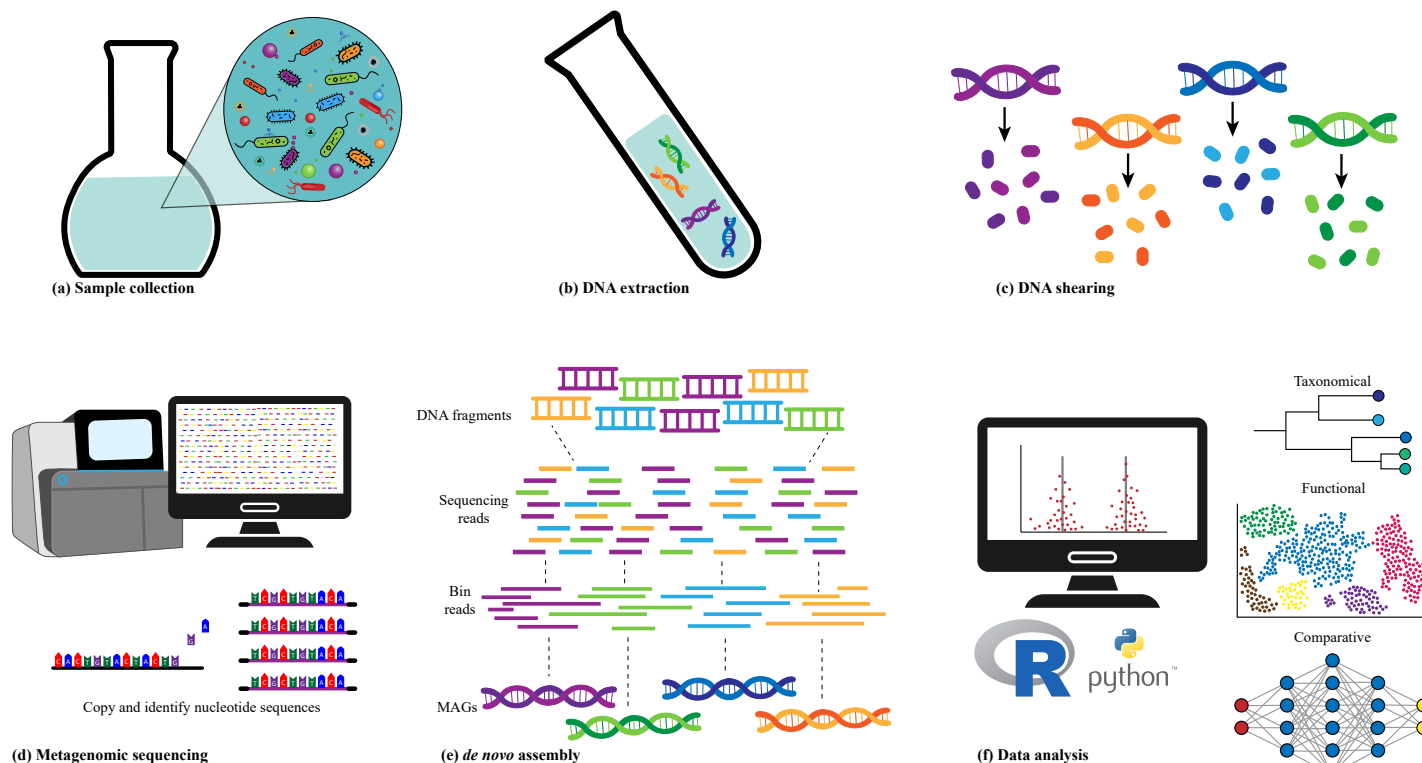


FIGURE CC19-2. The metagenomics process. (a) Marine samples are collected either via water collection or sediment coring. (b) All of the DNA in the sample is extracted. (c) The DNA is sheared, or broken into smaller pieces, for sequencing. (d) Samples are sequenced where millions of copies of the fragmented DNA are made, and the nucleotide bases of those copies are identified. Sequences are returned as reads that report the nucleotide sequence and quality (Phred score) of each DNA strand. (e) Reads are aligned based on overlapping nucleotides, and similar sequences are binned (grouped) together to form metagenome-assembled genomes (MAGs). (f) The MAGs are then analyzed to identify the organisms present (taxonomic data), their metabolic pathways, and the dominant functions of the organisms present (functional data), as well as the overall community interactions with one another and the environment (comparative data).

that while it can tell us what genes are present, it can't tell us if or how often those genes are in use in the cell's life processes. Essentially, metagenomics provides the map of that organism, similar to a map of a community, but we can't know for sure that every business on the map is open.

Samples can be taken from any type of environment (marine environments, soil, feces, gut) without any prior knowledge or understanding of the microorganisms present. Within marine environments, samples are either collected by seafloor coring or water collection. A sediment push core can be used to sample ocean floors and is typically then separated into about 2 cm sections before total DNA extraction is completed. Water sampling involves the collection of a large volume of water, at least 2 to 3 L, which can be collected at varying depths. Once collected, the water is processed in the lab using filtration methods to concentrate the microbes in the sample. The filter with the concentrated microbes is then processed for total DNA extraction. Once DNA has been extracted and confirmed to be of sufficient concentration and quality, the samples are ready for sequencing.

Metagenomics is the sequencing, or reading of the nucleotide bases (A, G, C, T) that make up DNA in the correct order, of all of the DNA present in a sample. In order to do this, the DNA is broken, typically through shearing, into pieces of approximately the same length. The size that the DNA is broken into depends on the sequencing technology, where the typical size for short-reads is around 400 nucleotides. There are two primary ways of shearing DNA: mechanical and enzymatic. Mechanical shearing is often completed using sonication, which transmits high-frequency, short-wavelength acoustic energy, breaking the DNA into smaller

pieces at random spots in the DNA. Enzymatic shearing uses specific and non-specific enzymes to nick and cut the DNA. Shearing the DNA results in millions of sequences that need to be read then reassembled and analyzed (**Fig. CC19-2**). The sequences are typically recorded in text-based files that store DNA sequences with associated quality scores.

DNA sequences generally fall into two categories: short-read and long-read. The difference between short- and long-read sequencing refers to the length of DNA that is copied in a single sequence. Short-reads are typically about 400 nucleotides in length, while long-reads range from 1500 to >10,000 nucleotides in length. The longer read lengths are used to account for long sequences of repeats in the DNA and areas where shorter runs missed DNA sequencing. Shorter reads tend to be more accurate, making them more reliable (e.g., higher quality) than some long-read technologies. While other long-read technologies are extremely accurate (99.99%), their cost makes their use prohibitive. Often, to maximize coverage, a mix of short- and long-read sequencing is used, balancing high throughput and affordability. Different sequencing technologies vary in how the sequencing is performed, but the overall output is similar.

Before sequences can be reassembled, it is important to check the quality of the sequences and discard any low-quality sequences. This is a crucial step because low-quality reads have a high error rate and can result in the incorrect identification of a gene or even an organism. Quality is measured using **Phred quality scores**, which reflect the confidence that a nucleotide base read is correct. Phred scores (Q-score) typically range from 10 to 50 on a logarithmic scale, where a lower Phred score indicates a higher

likelihood of error. For example, a Phred score of 20 (Q20) indicates that there is a 1 in 100 (1%) chance that a nucleotide is incorrect, and a Phred score of 40 (Q40) indicates that there is a 1 in 10,000 (0.01%) chance that a nucleotide is incorrect. The generally accepted benchmark for quality is a Phred score of 30 (Q30). Phred scores are reported in the final sequencing files.

Metagenomic data is assembled and analyzed using **bioinformatics**. Because the organisms in the original sample are unknown, the sequences must be reassembled without a template; this is known as **de novo assembly**. To do this, the sequences are aligned, or scaffolded, by looking for areas where the reads overlap. This process is repeated until there are no more available overlaps. The assembled reads are referred to as contigs. The contigs then undergo a process known as **binning**, which groups contigs that likely come from the same or closely related genomes. Once grouped, genomes are assembled to create **metagenome-assembled genomes (MAGs)**. The MAGs are then analyzed to identify the organisms in the sample, and identify novel metabolic pathways and the specific genes present. This allows assessment of the potential role each microbe may have in the system.

Due to the sheer number of reads, processing them requires a high amount of computational power and is most often completed using supercomputers.

Different types of analysis of the assembled MAGs can be performed in order to answer many different questions. The data analyzed can be organized into three types:

- Taxonomic data provides insight into the identities of the organisms present in a sample. This includes matching the sequences to a database of known sequences. From there, the **relative abundance** (approximation) of specific species can be determined, and the taxonomic diversity can be examined. Additionally, novel species can be identified.
- Functional data includes annotating all known genes and identifying the metabolic pathways present for each organism. Additionally, the most abundant genes (**functional abundance**) in the system and the dominant functions in the system can be determined.
- Comparative data looks at how the MAGs interact within the biome in relation to one another and the environment as a whole. The types of analysis performed include correlation plots of interactions with the environment, such as depth, oxygen saturation, metal availability, temperature, and salinity, as well as between species (co-occurrence networks).

To organize and share sequencing data from the marine biome, the KMAP Global Ocean Gene Catalog 1.0, the largest open-source database of ocean microbes, was developed. At the time of its introduction in 2024, the catalog comprised 308.6 million gene groups from over 2100 ocean samples.

As mentioned above, metagenomics can be used to identify which microorganisms comprise the sampled biome and which genes these microorganisms possess, but it cannot tell us whether these genes are active. Many genes are inactive or are only active under certain environmental conditions. DNA is the repository for all the information needed to perform the life processes of that species. Active genes in this DNA library are passed to RNA, where they act as instructions to run the cell's life processes, including the construction of essential compounds such as amino acids, other organic acids, sugars, lipids, and proteins. One particular set of RNA, called mRNA (messenger RNA), contains the

information needed for the cell to construct proteins.

Determining which genes are active in an environmental sample is difficult, but can be addressed by using one or more of three techniques: metatranscriptomics, proteomics, and metabolomics. Metatranscriptomics is performed in a similar way to metagenomics, except that it extracts and analyzes mRNA in the environmental sample, so it provides information on which genes are active in protein construction. Proteomics and metabolomics analyze environmental samples: proteomics for proteins and metabolomics for various other compounds, including amino acids, other organic acids, sugars, and lipids. Each of these three techniques involves careful sample preparation and analysis and produces a large amount of data that must be processed. Further, the results need to be carefully interpreted to obtain accurate information about which genes identified by metagenomic analysis are actually in use under the environmental conditions present when each sample was collected.

REFERENCES

- Figure opener – NASA, 2007: Blue Marble – Image of the Earth from Apollo 17. Accessed 21 November 2025, <https://www.nasa.gov/image-article/blue-marble-image-of-earth-from-apollo-17/>.
- Figure CC04-01 – Sundborg, Å. 1956: The River Klarälven a study of fluvial processes. *Geografiska Annaler*, **38**, 125–316, <https://doi.org/10.1080/20014422.1956.11880887>.
- Figure CC09-01 – Scott, R. C., 1996: *Introduction to Physical Geography*. West Publishing, 560 pp, https://openlibrary.org/works/OL2973497W/Introduction_to_physical_geography.
- Figure CC11-03 – Adapted from Gleick, J., 1987: *Chaos: Making a New Science*, Viking Penguin, 400 pp.
- Figure CC11-04 – Adapted from Gleick, J., 1987: *Chaos: Making a New Science*, Viking Penguin, 400 pp.

