[DM-05-044] Earth's Shape, Sea Level, and the Geoid

Abstract

In 1928, C. F. Gauss set the modern definition of the shape of the Earth, being described as the shape the oceans would adopt if they were entirely unperturbed and, thus, placid—a surface now called the geoid. This surface cannot be observed directly because the oceans have waves, tides, currents, and other perturbations. Nonetheless, the geoid is the ideal datum for heights, and the science of determining the location of the geoid for practical purposes is the topic of physical geodesy. The geoid is the central concept that ties together what the various kinds of height mean, how they are measured, and how they are inter-related.

Keywords: datum, Earth geometry, ellipsoid, geodesy, geoid

Author & citation

Meyer. T. (2021). Earth's Shape, Sea Level, and the Geoid. The Geographic Information Science & Technology Body of Knowledge (2nd Quarter 2021 Edition), John P. Wilson (ed.). DOI: 10.22224/gistbok/2021.2.8.

This Topic is also available in the following editions:

DiBiase, D., DeMers, M., Johnson, A., Kemp, K., Luck, A. T., Plewe, B., and Wentz, E. (2006). Approximating the Earth's shape with geoids. The Geographic Information Science & Technology Body of Knowledge. Washington, DC: Association of American Geographers.

Explanation

As an aircraft is on approach to land, the crew must ensure that the bottom of the aircraft is above the tops of all the obstructions below it along its flight path. When civil engineers design a storm sewer, they must ensure that the entry point of the runoff is higher than the outflow point. It might seem as though these two scenarios mean the same thing—that something must be higher than another—but they do not. The aircraft must maintain a linear vertical separation between itself and the obstacles. The water must flow downhill. Vertical separation is about geometry, and flowing downhill is about gravity. Surprisingly, there are places whose linear vertical separation is zero and yet water will flow between them due to the force of gravity alone. This distinction has led to there being many types of heights, which all generally mean an offset in a vertical direction in some sense relative to a reference surface of some sort.

Pilots need accurate elevations, with elevation meaning (for the moment) a height above sea level. The altimeters in aircraft provide elevations to be compared with the elevations of the obstacles marked on an aeronautical chart of the approach. Civil engineers need accurate differences in height, with height differences meaning measurements reflecting the direction water will flow, the downhill direction. The elevations along the sewers are unneeded: the design must show the required change in height along the sewer, not the height above sea level of the design stations. Land surveyors use differential levels and

total stations to determine the change in height along the sewer, not altimeters.

Suppose someone were to say, “The height of the top of this table is fifty meters.” One might respond with the question, “Fifty meters above what?” Elevations are given relative to a reference surface, to a vertical datum, and the elevation on the datum surface is identically zero by definition. Here is the definition of elevation according to the US National Oceanic and Atmospheric Administration’s (NOAA) National Geodetic Survey (NGS):

elevation: The distance of a point above a specified surface of constant potential; the distance is measured along the direction of gravity between the point and the surface.

This rather technical definition does not mention sea level, instead, the datum for elevations is a surface of constant potential, which is short for gravitational potential energy. Being a surface of constant potential means that the potential energy is the same everywhere on this surface, it is an equipotential surface. Water does not flow across an equipotential surface due to gravity. The force of gravity arises due to a change of potential, so, since the potential energy is the same everywhere on an equipotential surface, the change of potential is zero and, therefore, there is no force available to move the water. Equipotential surfaces are level: everywhere on an equipotential surface is at the same height regarding the notions of uphill and downhill.

The definition for elevation is specialized to that for orthometric elevation (aka orthometric height):

elevation, orthometric: The distance between the geoid and a point measured along the plumb line and taken positive upward from the geoid. Appearances to the contrary notwithstanding, this definition differs from that for elevation only in that it specifies the datum: the direction of gravity between the point and the surface means the same thing as “along the plumb line.” (Plumb lines will be discussed more below.) The datum is something called the geoid. Earth’s potential energy field pervades all space. Every place in the universe has some potential energy due to its proximity to Earth’s center of mass. In the absence of all other massive bodies, this potential energy field consists of concentric surfaces of constant potential, like layers of an onion. They are all level surfaces, so which one is the best choice (in some sense) for the datum for elevations? This leads us to the definition of the geoid:

geoid: The equipotential surface of the Earth's gravity field coinciding with the mean sea level of the oceans (Gauss, 1828).

This notion was first given by Carl Friedrich Gauss (Gauss, 1828) and later given its name by Listing (1872) (Torge & Müller, 2012). This definition might be better understood by its common-language explanation. Suppose, somehow, that Earth’s oceans were unperturbed so that they had no waves, no tides, and no currents; that they were altogether placid. The surface of the unperturbed oceans would be coincident with the geoid. The geoid is part of Earth’s gravity field so it envelops the whole Earth, not just the parts covered in seawater, so the explanation should be extended by allowing the oceans to freely flow through all the land masses, as well.

Every equipotential surface is associated with a single value of potential energy. The geodesy community denotes the potential value of the geoid as W0 so the geoid is chosen by determining a specific value for W0. For example, (Amin, et al., 2019) defined the geoid

to be W0= 62,636,848.102±0.004 m2s-2 based on gravity observations obtained from satellites.

Any “reasonable” value for W0 will do (Smith, 1998) and, in fact, the value might not be chosen to match sea level so long as it is “close” to it. For example, the datum surface of the North American Vertical Datum of 1988 (NAVD88) was chosen to be the level surface at a single survey marker, Father Point/Rimouski, in Quebec Canada, located at the mouth of the St. Lawrence River, a marker that is not in the tidal zone (Zilkowski, et al., 1992). (Bursa, 1999) determined the value W0=62,636,861.4±0.5 m2s-2 for the NAVD88 level surface. The offset of the NAVD88 level surface to the geoid is unknown but thought to be “small.” (Roman & Li, 2011) determined W0=62,636,856.88 m2s-2 to be, for the current models of the US and Canada, “…near optimal though some small change may be required depending pending refinement of the results along the eastern coast of the U.S.A., where the effects of the Gulf Stream complicate this analysis.” More importantly, the offset is somewhat irrelevant. First, height differences are often more important than elevations themselves: knowing the direction that liquids will flow is important to almost all civil engineering projects, whereas knowing the elevations of those projects is often largely irrelevant. The location of the datum surface vanishes in the difference of heights, so the location of the datum surface is immaterial for relative heights. Second, the geoid itself is constantly changing due to the oceans’ temperatures changing, the amount of water in the oceans changing, the changing salinity in the water, changing currents, etc. There is no single, permanent, equipotential surface that is “the geoid” (Smith, 1998), so it is somewhat pointless to try to capture it exactly, even at one moment in time. Third, nations might choose different level surfaces to be their national vertical datum to suit their own needs.

Since level surfaces are comprised of all the places at one geopotential value, and therefore at the same height in terms of how liquids flow, it follows that a change in height means a change in geopotential value; this is a change in energy, not a change in distance, per se. Potential energy has a linear change in the separation between equipotential surfaces, however, the equipotential surfaces are not parallel, it so happens. Figure 1 is an artist’s rendition of a few equipotential surfaces. Notice that the surfaces farther from the Earth have less curvature. Also notice that, since each surface has different curvature, the surfaces are not parallel. So the linear separation between two equipotential surfaces can, and does, change from place to place (Fig. 2). This logical disconnect is the underlying reason why geometric separation is not the same thing as the change in potential energy between two equipotential surfaces. Aviators need the former, and sewer designers need the latter.

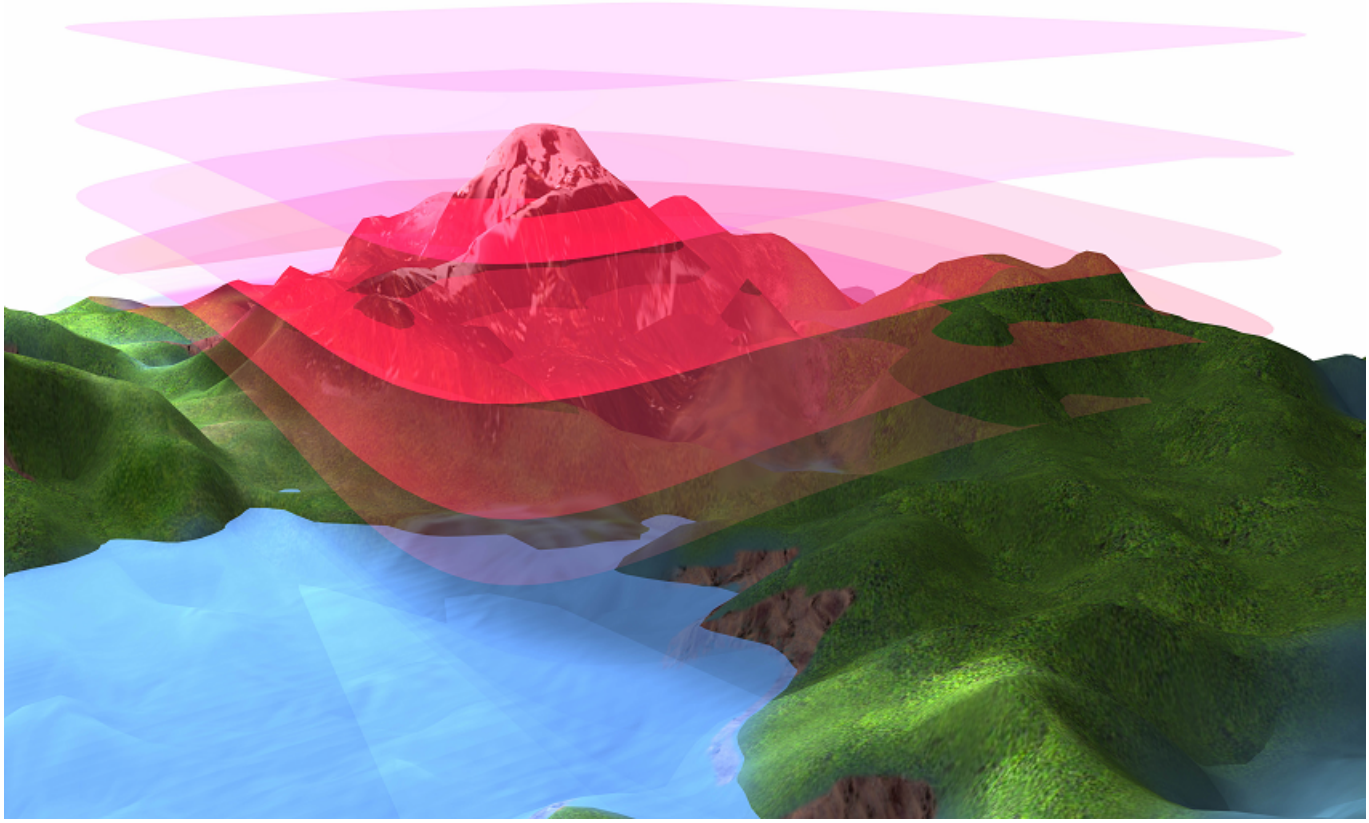


Figure 1. An artist's rendition of several equipotential surfaces have different curvatures and, thus, are not parallel. Source: author; adapted from Meyer (2010).

We can now define a **contour:**

contour: An imaginary line on the ground, all points of which are at the same elevation above or below a specified reference surface.

The image of a contour on a (topographical) map is called a contour line. By definition, all places on the ground on the same contour are at the same elevation. One might then conclude that a contour is the intersection of a level surface (equipotential surface) with the ground, but this is not true in general. Figure 2 shows two equipotential surfaces: the higher one intersects the ground along the path B-C, and the lower one intersects the ground through A-D. Let us suppose that the lower equipotential surface is the geoid. The path B-C is on an equipotential surface, so that path is level; same for the path A-D. However, the equipotential surfaces are not parallel so the orthometric height at B is different, in fact it is lower, than the orthometric height at C. The definition of orthometric height imparts an unintuitive and unwelcome property to contours: curves of constant elevation (contours) are generally not level curves.

However, this is largely, but not entirely, academic. Equipotential surfaces diverge slowly over distance, so contours are nearly level, but generally not exactly level. This can become an issue in places with very little topographic relief, which tend to be subject to flooding. In such places even minute changes in topography can move a fiducial boundary, like the 100-year flood line, by a great horizontal distance. The geopotential notion of height, called

dynamic height, can be called for. For example, Canada uses dynamic heights for the International Great Lakes Datum 1985. In Fig. 2, the dynamic height along each path is constant.

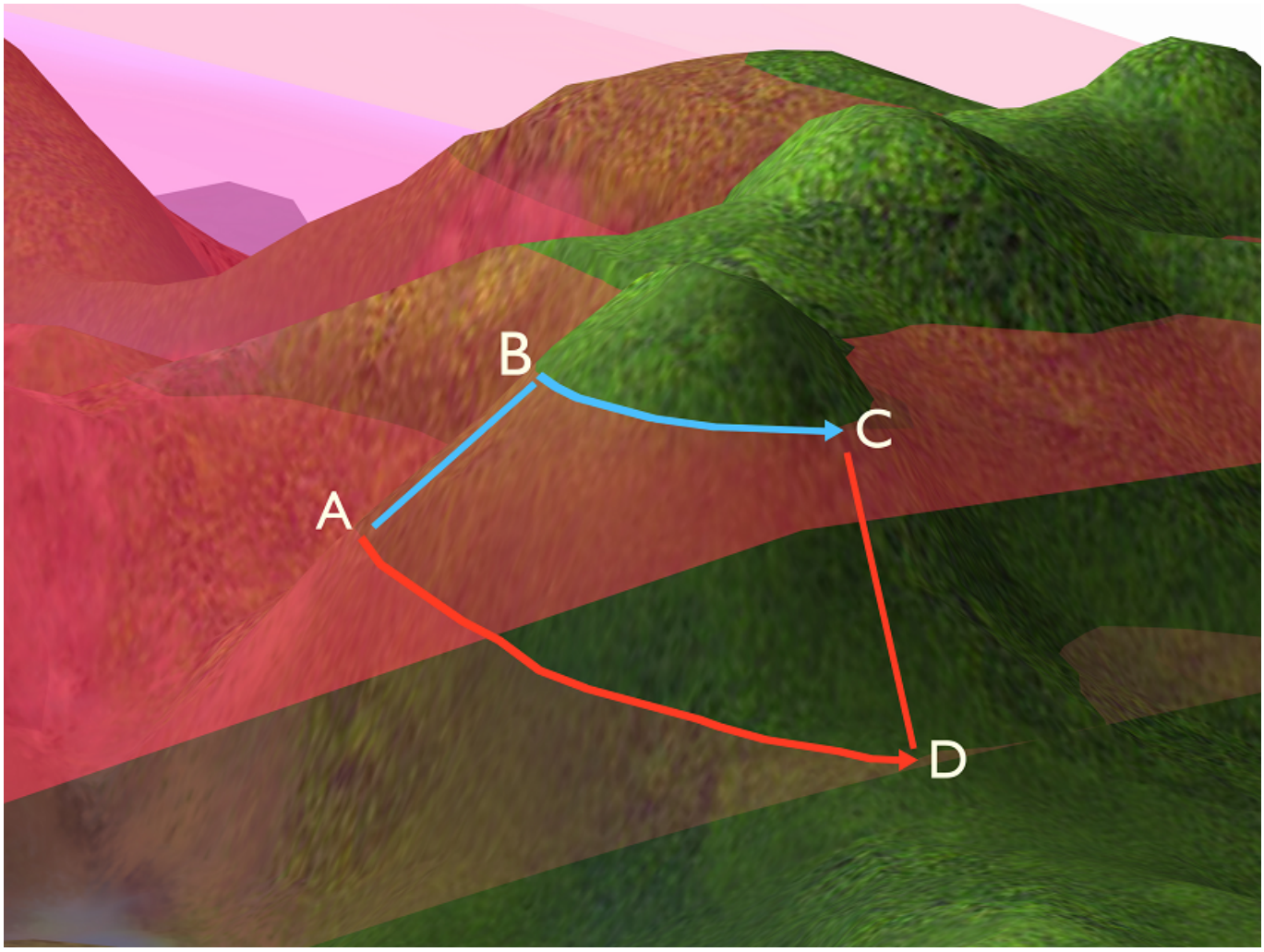


Figure 2. A close-up highlighting the non-parallelism of two equipotential surfaces. Source: author; adapted from Meyer (2010).

Nowadays, many aircraft use global navigation satellite system (GNSS) positioning technologies to determine their position and velocity. GNSS positions are inherently three-dimensional, but they refer to a Cartesian (XYZ) coordinate system whose Z-axis is parallel to, and nearby, Earth's rotational axis. Earth's axis is constantly moving in a roughly circular path. This motion is called precession, and it includes fine-scale motions such as the Chandler Wobble. This motion prohibits Earth's actual pole from being used for the Z-axis, so the Z-axis is chosen to be near to the actual axis but it is created so as to not move (as much as possible). There is no notion of "up" inherent in this coordinate system. The idea of height is closely associated with directions up and down. **Down** is the direction something falls under the influence of gravity alone, and **up** is the direction opposite of down. A GNSS receiver does not produce coordinates (directly) that provide height; however, there are well-known formulas to convert from XYZ coordinates to geodetic

longitude, latitude, and height (Claessens, 2019), so geodetic height is readily available from GNSS receivers. Geodetic heights are purely geometric, having nothing to do with gravitational potential energy, and they do not use an equipotential surface for their datum. Their datum is a simple mathematical model of Earth's macroscopic shape (the geoid) that is called a **reference ellipsoid**.

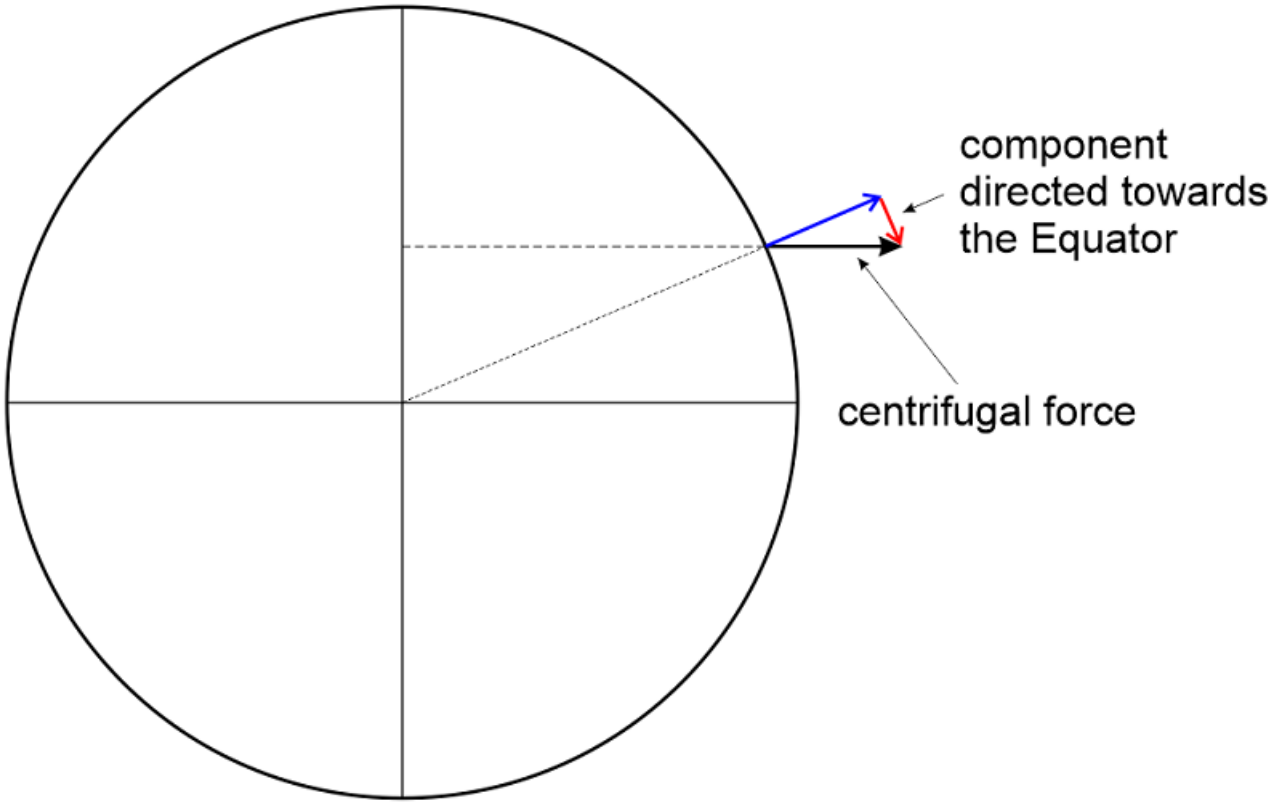


Figure 3. The decomposition of a centrifugal force vector into components directed upwards and towards the Equator. Source: author.

Reference ellipsoids are **ellipsoids of revolution**, also known as **spheroids**, or just "ellipsoids." It has been known conclusively since the time of Sir Isaac Newton that Earth's macroscopic shape is that of an oblate spheroid. Newton determined this—oblate vs. prolate—using arguments about gravity: Earth's diurnal rotation imparts a centrifugal force that is perpendicular to its rotational axis. This force is perpendicular to Earth's surface only at the Equator, and, everywhere else, this force has a component directed towards the Equator (Fig. 3). If Earth were spherical, this component would cause the oceans to pile up around the Equator leaving the polar regions devoid of oceans. This does not happen, apparently, so Earth's macroscopic shape must tilt slightly towards the oblate to counter-act this Equatorial component of the centrifugal force (Stommel & Moore, 1989). This flattening of the ellipsoid can be derived using Stoke's Integral and associated formulas (Hofmann-Wellenhof & Moritz, 2006; Amin, et al., 2019). Modern geodetic reference frames are geocentric, meaning the origin of their defining XYZ coordinate system is at Earth's

center of mass, including the oceans and the atmosphere (Fig. 4). Thus, modern reference ellipsoids are placed by reference frames such that their origins are likewise geocentric and, thus, are suitable as a geometric vertical datum. We define geodetic heights as

height, geodetic––The perpendicular distance from an ellipsoid of reference to a point.

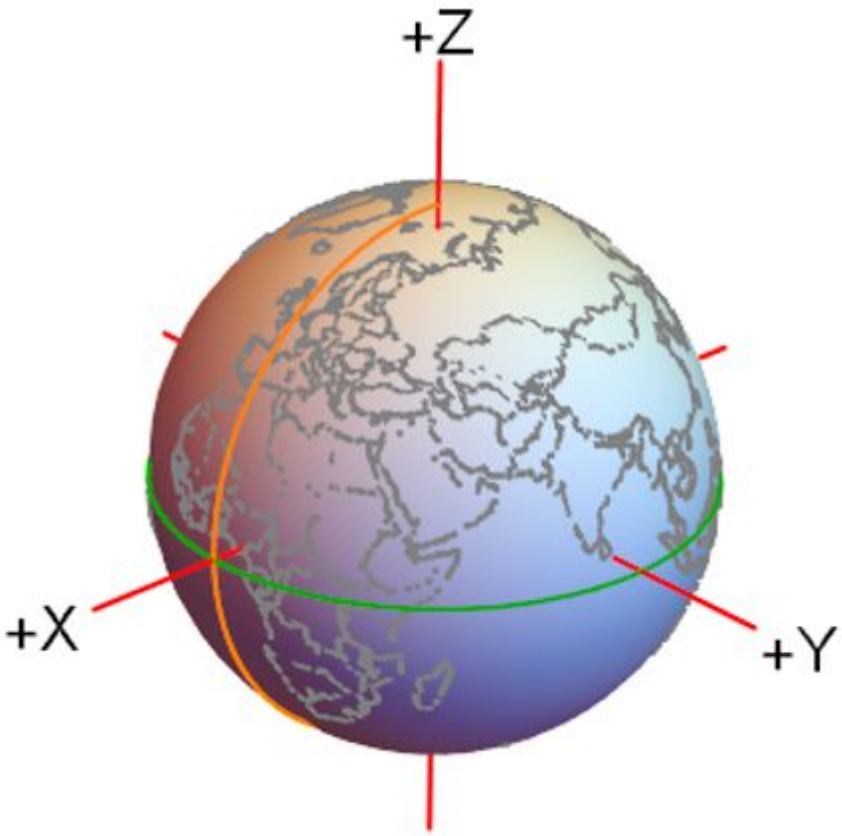


Figure 4. An XYZ coordinate system, whose origin is at Earth's center of mass including the oceans and the atmosphere. Source: author.

Being the datum surface, the orthometric elevation of the geoid itself is identically zero. However, reference ellipsoids are not the same surface as the geoid, so one can create a geodetic height map of the geoid relative to the simple geometric ellipsoid and, thus, visualize it. The geodetic heights of the geoid are found using **geoid models**, such as the Earth Gravitational Model 2008 (EGM2008) (Pavlis, et al., 2012). These models show that the World Geodetic System 1984 (WGS84) geoid departs from the WGS84 reference ellipsoid by more than 100 m in places. The geodetic height of a place on the geoid relative to a reference ellipsoid is called a geoid height.

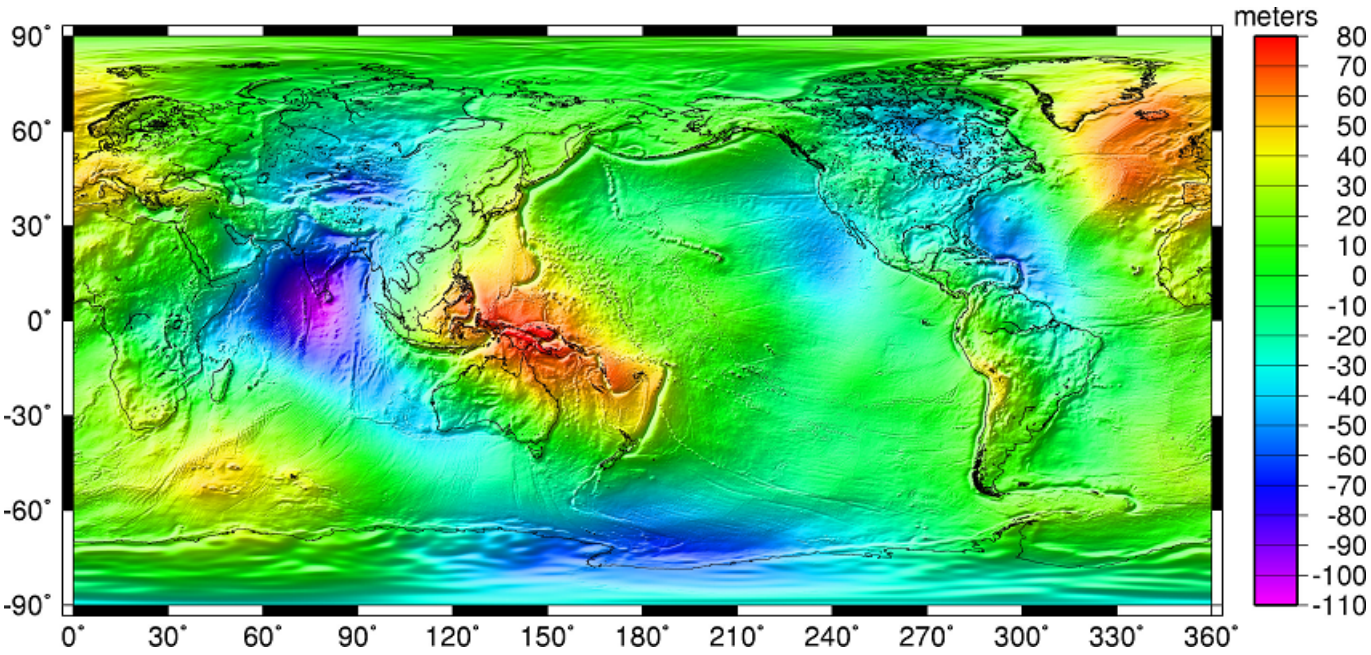


Figure 5. Geoid undulations from the EGM2008 gravity model. Source: NGA.

Geoid models provide the connection between elevations and geodetic heights:

$H = h - N$

where H is orthometric height, h is geodetic height, and N is geoid height. Figure 6 illustrates the geometric relationships. In Fig. 6, the orthometric height, H, of p is the arc length of the slightly curved blue line from the geoid up to p. The blue curved line is called a **plumb line**. The plumb line is slightly curved because it must remain perpendicular to all the equipotential surfaces from the geoid up to p, and those surfaces are not parallel in general. The geodetic height, h, of p is the length of the black straight-line segment from the ellipsoid up to p. The geoid height, N, is the length of the straight-line segment from the ellipsoid to p's projection down to the geoid along the plumb line. The Figure illustrates that the formula is only approximately equal, not exactly equal. However, the plumb-line's curvature is minute, so we may safely write the formula as an equality for practical purposes. So, a GNSS receiver determines a position in XYZ coordinates; these are converted to geodetic longitude, latitude, and height; and the geodetic height is converted to orthometric height using the formula above.

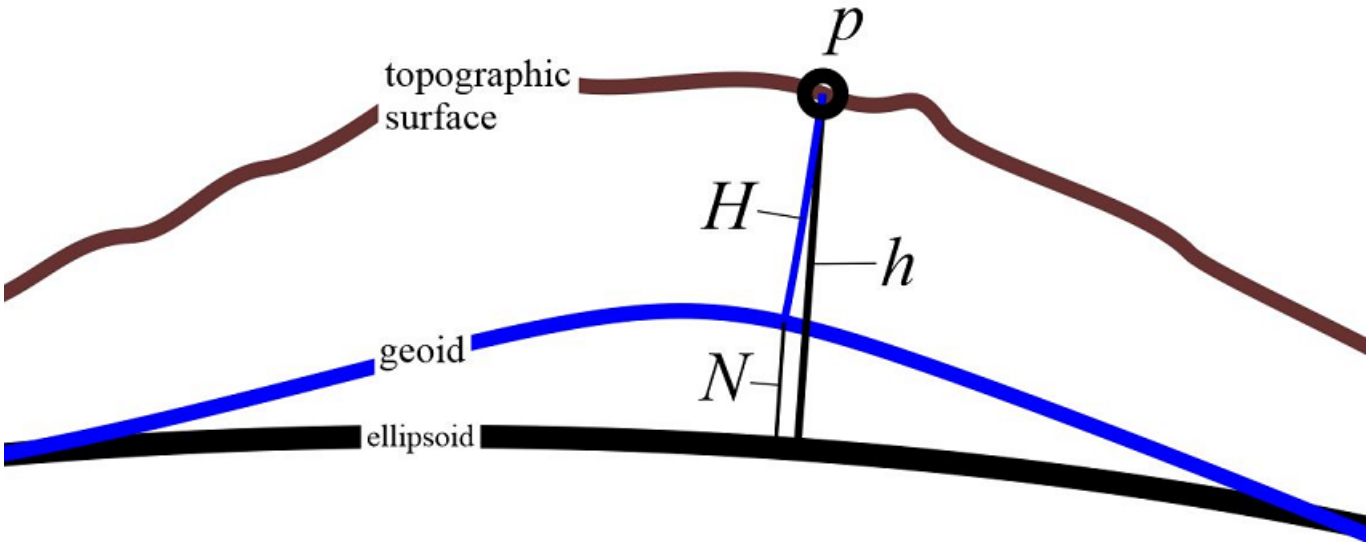


Figure 6. The geometric relationships between orthometric height H, geodetic height h, and geoid height N. Source: author.

There is a vast literature about the geoid, reference ellipsoids, heights, and heighting, together known as physical geodesy. For a more lengthy and in-depth treatment but still aimed at the general geomatics community, see Meyer (2010).

References

Amin, H., Sjöberg, L. E. & Bagherbandi, M. (2019). A global vertical datum defined by the conventional geoid potential and the Earth ellipsoid parameters. Journal of Geodesy, Volume 93, pp. 1943-1961.

Bursa, M. (1999). Determination of the geopotential at the tide gauge defining the North American Vertical Datum 1988 (NAVD88). GEOMATICA, 53(3), pp. 291-296.

Claessens, S. J., 2019. Efficient transformation from Cartesian to geodetic coordinates. Computers & Geosciences, Volume 133, p. 104307.

Gauss, C. F., 1828. Bestimmung des Breitenunterschiedes zwischen den Sternwarten von Göttingen und Altona durch Beobachtungen am Ramsdenschen Zenithsector. s.l.:Bei Vandenhoeck und Ruprecht.

Hofmann-Wellenhof, B. & Moritz, H. (2006). Physical Geodesy. Springer Science & Business Media.

Listing, J. B. (1872). Über unsere jetzige Kenntniss der Gestalt und Grösse der Erde: Aus den Nachrichten der K. Ges. der Wiss. s.l.:Dieterich.

Meyer, T. (2010). Introduction to Geometrical and Physical Geodesy: Foundations of Geomatics. Redlands, CA: Esri Press.

Pavlis, N. K., Holmes, S. A., Kenyon, S. C. & Factor, J. K. (2012). The development and

evaluation of the Earth Gravitational Model 2008 (EGM2008). Journal of Geophysical Research Solid Earth, 117(B4).

Roman, D. R. & Li, X. (2011). Determination of an Optimal Geopotential Value for the North American Geoid. AGU Fall Meeting Abstracts.

Smith, D. A. (1998). There is no such thing as "the" EGM96 geoid: subtle points on the use of a global geopotential model. IGeS Bulletin, Volume 8, pp. 17-27.

Stommel, H. M. & Moore, D. W. (1989). An Introduction to the Coriolis Force. New York: Dover.

Zilkowski, D. B., Richards, J. H. & Young, G. M. (1992). Special Report: Results of the General Adjustment of the North American Vertical Datum of 1988. Surveying and Land Information Systems, 52(3), pp. 133-149.