What Does Height Really Mean?

Thomas H. Meyer[*] Daniel R. Roman[†]
David B. Zilkoski[‡]

[*] University of Connecticut, thomas.meyer@uconn.edu
[†] National Geodetic Survey
[‡] National Geodetic Suvey

What does *height* really mean?

Thomas Henry Meyer
Department of Natural Resources Management and Engineering
University of Connecticut
Storrs, CT 06269-4087
Tel: (860) 486-2840
Fax: (860) 486-5480
E-mail: thomas.meyer@uconn.edu

Daniel R. Roman
National Geodetic Survey
1315 East-West Highway
Silver Springs, MD 20910
E-mail: Dan.Roman@noaa.gov

David B. Zilkoski
National Geodetic Survey
1315 East-West Highway
Silver Springs, MD 20910
E-mail: Dave.Zilkoski@noaa.gov

June, 2007

The authors would like to acknowledge the careful and constructive reviews of this series by Dr. Dru Smith, Chief Geodesist of the National Geodetic Survey.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Preamble

This monograph was originally published as a series of four articles appearing in the Surveying and Land Information Science. Each chapter corresponds to one of the original papers. This paper should be cited as

Meyer, Thomas H., Roman, Daniel R., and Zilkoski, David B. (2005) What does *height* really mean? Part 1: Introduction. In *Surveying and Land Information Science*, 64(4): 223-234.

This is the first paper in a four-part series considering the fundamental question, "what does the word height really mean?" National Geodetic Survey (NGS) is embarking on a height modernization program in which, in the future, it will not be necessary for NGS to create new or maintain old orthometric height bench marks. In their stead, NGS will publish measured ellipsoid heights and computed Helmert orthometric heights for survey markers. Consequently, practicing surveyors will soon be confronted with coping with these changes and the differences between these types of height. Indeed, although "height" is a commonly used word, an exact definition of it can be difficult to find. These articles will explore the various meanings of height as used in surveying and geodesy and present a precise definition that is based on the physics of gravitational potential, along with current best practices for using survey-grade GPS equipment for height measurement. Our goal is to review these basic concepts so that surveyors can avoid potential pitfalls that may be created by the new NGS height control era. The first paper reviews reference ellipsoids and mean sea level datums. The second paper reviews the physics of heights culminating in a simple development of the geoid and explains why mean sea level stations are not all at the same orthometric height. The third paper introduces geopotential numbers and dynamic heights, explains the correction needed to account for the non-parallelism of equipotential surfaces, and discusses how these corrections were used in NAVD 88. The fourth paper presents a review of current best practices for heights measured with GPS.

1.2 Preliminaries

The National Geodetic Survey (NGS) is responsible for the creation and maintenance of the United State's spatial reference framework. In order to address unmet spatial infrastructure issues, NGS has embarked on a height modernization program whose "... most desirable outcome is a unified national positioning system, comprised of consistent, accurate, and timely horizontal, vertical, and gravity control networks, joined and maintained by the Global Positioning System (GPS) and administered by the National Geodetic Survey" (National Geodetic Survey 1998). As a result of this program, NGS is working with partners to maintain the National Spatial Reference System (NSRS).

In the past, NGS performed high-accuracy surveys and established horizontal and/or vertical coordinates in the form of geodetic latitude and longitude and orthometric height. The National Geodetic Survey is responsible for the federal framework and is continually developing new tools and techniques using new technology to more effectively and efficiently establish this framework, i.e., GPS and Continually Operating Reference System (CORS). The agency is working with partners to transfer new technology so the local requirements can be performed by the private sector under the supervision of the NGS (National Geodetic Survey 1998).

Instead of building new benchmarks, NGS has implemented a nation-wide network of continuously operating global positioning system (GPS) reference stations known as the CORS, with the intent that CORS shall provide survey control in the future. Although GPS excels at providing horizontal coordinates, it cannot directly measure an orthometric height; GPS can only directly provide ellipsoid heights. However, surveyors and engineers seldom need ellipsoid heights, so NGS has created highly sophisticated, physics-based, mathematical software models of the Earth's gravity field (Milbert 1991, Milbert & Smith 1996*a*, Smith & Milbert 1999, Smith & Roman 2001) that are used in conjunction with ellipsoid heights to infer Helmert orthometric heights (Helmert 1890). As a result, practicing surveyors, mappers, and engineers working in the United States may be working with mixtures of ellipsoid and orthometric heights. Indeed, to truly understand the output of all these height conversion programs, one must come to grips with heights in all their forms, including elevations, orthometric heights, ellipsoid heights, dynamic heights, and geopotential numbers.

According to the Geodetic Glossary (National Geodetic Survey 1986), height is defined as, "The distance, measured along a perpendicular, between a point and a reference surface, e.g., the height of an airplane above the ground." Although this definition seems to capture the intuition behind height very well, it has a (deliberate) ambiguity regarding the reference surface (datum) from which the measurement was made.

Heights fall broadly into two categories: those that employ the Earth's gravity field as their datum and those that employ a reference ellipsoid as their datum. Any height referenced to the Earth's gravity field can be called a "geopotential height," and heights referenced to a reference ellipsoid are called "ellipsoid heights." These heights are not directly interchangeable; they are referenced to different datums and, as will be explained in subsequence papers, in the absence of site-specific gravitation measurements there is no rigorous transformation between them. This is a situation analogous to that of the North American Datum of 1983 (NAD83) and the North American Datum of 1927 (NAD27) - two horizontal datums for which there is no rigorous transformation.

The definitions and relationships between elevations, orthometric heights, dynamic heights, geopotential numbers, and ellipsoid heights are not well understood by many practitioners. This is perhaps not too surprising, given the bewildering amount of jargon associated with heights. The NGS glossary contains 17 definitions with specializations for "elevation," and 23 definitions with specializations for "height," although nine of these refer to other (mostly elevation) definitions. It is the purpose of this series, then, to review these concepts with the hope that the reader will have a better and deeper understanding of what the word "height" really means.

1.2.1 The Series

The series consists of four papers that review vertical datums and the physics behind height measurements, compare the various types of heights, and evaluate the current best practices for deducing orthometric heights from GPS measurements. Throughout the series we will enumerate figures, tables, and equations with a Roman numeral indicating the paper in the series from which it came. For example, the third figure in the second paper will be numbered, “Figure II.3”.

This first paper in the series is introductory. Its purpose is to explain why a series of this nature is relevant and timely, and to present a conceptual framework for the papers that follow. It contains a review of reference ellipsoids, mean sea level, and the U.S. national vertical datums.

The second paper is concerned with gravity. It presents a development of the Earth’s gravity forces and potential fields, explaining why the force of gravity does not define level surfaces, whereas the potential field does. The deflection of the vertical, level surfaces, the geoid, plumb lines, and geopotential numbers are defined and explained.

It is well known that the deflection of the vertical causes loop misclosures for horizontal traverse surveys. What seems to be less well known is that there is a similar situation for orthometric heights. As will be discussed in the second paper, geoid undulations affect leveled heights such that, in the absence of orthometric corrections, the elevation of a station depends on the path taken to the station. This is one cause of differential leveling loop misclosure. The third paper in this series will explain the causes of this problem and how dynamic heights are the solution.

The fourth paper of the series is a discussion of height determination using GPS. GPS measurements that are intended to result in orthometric heights require a complicated set of datum transformations, changing ellipsoid heights to orthometric heights. Full understanding of this process and the consequences thereof requires knowledge of all the information put forth in this review. As was mentioned above, NGS will henceforth provide the surveying community with vertical control that was derived using these methods. Therefore, we feel that practicing surveyors can benefit from a series of articles whose purpose is to lay out the information needed to understand this process and to use the results correctly.

The current article proceeds as follows. The next section provides a review of ellipsoids as they are used in geodesy and mapping. Thereafter follows a review of mean sea level and orthometric heights, which leads to a discussion of the national vertical datums of the United States. We conclude with a summary.

1.3 Reference Ellipsoids

A **reference ellipsoid**, also called **spheroid**, is a simple mathematical model of the Earth’s shape. Although low-accuracy mapping situations might be able to use a spherical model for the Earth, when more accuracy is needed, a spherical model is inadequate, and the next more complex Euclidean shape is an ellipsoid of revolution. An ellipsoid of revolution, or simply an “ellipsoid,” is the shape that results from rotating an ellipse about one of its axes. Oblate ellipsoids are used for geodetic purposes because the Earth’s polar axis is shorter than its equatorial axis.

1.3.1 Local Reference Ellipsoids

Datums and cartographic coordinate systems depend on a mathematical model of the Earth’s shape upon which to perform trigonometric computations to calculate the coordinates of places on the Earth and in order to transform between geocentric, geodetic, and mapping coordinates. The transformation between geodetic and cartographic coordinates requires knowledge of the ellipsoid being used, e.g., see (Bugayevskiy & Snyder 1995, Qihe, Snyder & Tobler 2000, Snyder 1987).

Likewise, the transformation from geodetic to geocentric Cartesian coordinates is accomplished by Helmert's projection, which also depends on an ellipsoid (Heiskanen & Moritz 1967, pp. 181-184) as does the inverse relationship; see Meyer (2002) for a review. Additionally, as mentioned above, measurements taken with chains and transits must be reduced to a common surface for geodetic surveying, and a reference ellipsoid provides that surface. Therefore, all scientifically meaningful geodetic horizontal datums depend on the availability of a suitable reference ellipsoid.

Until recently, the shape and size of reference ellipsoids were established from extensive, continental-sized triangulation networks (Gore 1889, Crandall 1914, Shalowitz 1938, Schwarz 1989, Dracup 1995, Keay 2000), although there were at least two different methods used to finally arrive at an ellipsoid (the "arc" method for Airy 1830, Everest 1830, Bessel 1841 and Clarke 1866; and the "area" method for Hayford 1909). The lengths of (at least) one starting and ending baseline were measured with instruments such as rods, chains, wires, or tapes, and the lengths of the edges of the triangles were subsequently propagated through the network mathematically by triangulation.

For early triangulation networks, vertical distances were used for reductions and typically came from trigonometric heighting or barometric measurements although, for NAD 27, "a line of precise levels following the route of the triangulation was begun in 1878 at the Chesapeake Bay and reached San Francisco in 1907" (Dracup 1995). The ellipsoids deduced from triangulation networks were, therefore, custom-fit to the locale in which the survey took place. The result of this was that each region in the world thus measured had its own ellipsoid, and this gave rise to a large number of them; see DMA (1995) and Meyer (2002) for a review and the parameters of many ellipsoids. It was impossible to create a single, globally applicable reference ellipsoid with triangulation networks due to the inability to observe stations separated by large bodies of water.

Local ellipsoids did not provide a vertical datum in the ordinary sense, nor were they used as such. Ellipsoid heights are defined to be the distance from the surface of the ellipsoid to a point of interest in the direction normal to the ellipsoid, reckoned positive away from the center of the ellipsoid. Although this definition is mathematically well defined, it was, in practice, difficult to realize for several reasons. Before GPS, all high-accuracy heights were measured with some form of leveling, and determining an ellipsoid height from an orthometric height requires knowledge of the deflection of the vertical, which is obtained through gravity and astronomical measurements (Heiskanen & Moritz 1967, pp. 82-84).

Deflections of the vertical, or high-accuracy estimations thereof, were not widely available prior to the advent of high-accuracy geoid models. Second, the location of a local ellipsoid was arbitrary in the sense that the center of the ellipsoid need not coincide with the center of the Earth (geometric or center of mass), so local ellipsoids did not necessarily conform to mean sea level in any obvious way. For example, the center of the Clarke 1866 ellipsoid as employed in the NAD 27 datum is now known to be approximately 236 meters from the center of the Global Reference System 1980 (GRS 80) as placed by the NAD83 datum. Consequently, ellipsoid heights reckoned from local ellipsoids had no obvious relationship to gravity. This leads to the ever-present conundrum that, in certain places, water flows "uphill," as reckoned with ellipsoid heights (and this is still true even with geocentric ellipsoids, as will be discussed below). Even so, some local datums (e.g., NAD 27, Puerto Rico) were designed to be "best fitting" to the local geoid to minimize geoid heights, so in a sense they were "fit" to mean sea level. For example, in computing plane coordinates on NAD 27, the reduction of distances to the ellipsoid was called the "Sea Level Correction Factor"!

In summary, local ellipsoids are essentially mathematical fictions that enable the conversion between geocentric, geodetic, and cartographic coordinate systems in a rigorous way and, thus, provide part of the foundation of horizontal geodetic datums, but nothing more. As reported by Fischer (2004), "O'Keefe [1] tried to explain to me that conventional geodesy used the ellipsoid only

[1] John O'Keefe was the head of geodetic research at the Army Map Service.

as a mathematical computation device, a set of tables to be consulted during processing, without the slightest thought of a third dimension."

1.3.2 Equipotential Ellipsoids

In contrast to local ellipsoids that were the product of triangulation networks, globally applicable reference ellipsoids have been created using very long baseline interferometry (VLBI) for GRS 80 (Moritz 2000)), satellite geodesy for the World Geodetic System 1984 (WGS 84) (DMA 1995), along with various astronomical and gravitational measurements. Very long baseline interferometry and satellite geodesy permit high-accuracy baseline measurement between stations separated by oceans. Consequently, these ellipsoids model the Earth globally; they are not fitted to a particular local region. Both WGS 84 and GRS 80 have size and shape such that they are a best-fit model of the geoid in a least-squares sense. Quoting Moritz (2000, p.128),

> The Geodetic Reference System 1980 has been adopted at the XVII General Assembly of the IUGG in Canberra, December 1979, by means of the following: ... recognizing that the Geodetic Reference System 1967 ... no longer represents the size, shape, and gravity field of the Earth to an accuracy adequate for many geodetic, geophysical, astronomical and hydrographic applications and considering that more appropriate values are now available, recommends ... that the Geodetic Reference System 1967 be replaced by a new Geodetic Reference System 1980, also based on the theory of the geocentric equipotential ellipsoid, defined by the following constants:
>
> - Equatorial radius of the Earth: $a = 6378137$ m;
> - Geocentric gravitational constant of the Earth (including the atmosphere): $GM = 3,986,005 \times 10^8 \mathrm{m}^3\mathrm{s}^{-2}$;
> - Dynamical form factor of the Earth, excluding the permanent tidal deformation: $J_2 = 108,263 \times 10^{-8}$; and
> - Angular velocity of the Earth: $\omega = 7292115 \times 10^{-11} \mathrm{rad\ s}^{-1}$.

Clearly, equipotential ellipsoid models of the Earth constitute a significant logical departure from local ellipsoids. Local ellipsoids are purely geometric, whereas equipotential ellipsoids include the geometric but also concern gravity. Indeed, GRS 80 is called an "equipotential ellipsoid" (Moritz 2000) and, using equipotential theory together with the defining constants listed above, one *derives* the flattening of the ellipsoid rather than measuring it geometrically. In addition to the logical departure, datums that employ GRS 80 and WGS 84 (e.g., NAD 83, ITRS, and WGS 84) are intended to be geocentric, meaning that they intend to place the center of their ellipsoid at the Earth's center of gravity. It is important to note, however, that NAD 83 currently places the center of GRS 80 roughly two meters away from the center of ITRS and that WGS 84 is currently essentially identical to ITRS.

Equipotential ellipsoids are both models of the Earth's shape and first-order models of its gravity field. Somiglinana (1929) developed the first rigorous formula for normal gravity (also, see Heiskanen & Moritz (1967, p. 70, eq. 2-78)) and the first internationally accepted equipotential ellipsoid was established in 1930. It had the form:

$$g_0 = 9.78046(1 + 0.0052884 \sin^2 \phi - 0.0000059 \sin^2 2\phi) \tag{1.1}$$

where

g_0 = acceleration due to gravity at a distance 6,378,137 m from the center of the idealized Earth; and

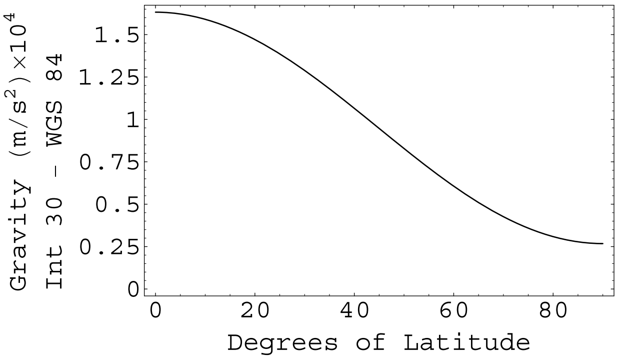

Figure 1.1: The difference in normal gravity between the 1930 International Gravity Formula and WGS 84. Note that the values on the abscissa are given 10,000 times the actual difference for clarity

ϕ = geodetic latitude (Blakely 1995, p.135).

The value g_0 is called **theoretical gravity** or **normal gravity**. The dependence of this formula on geodetic latitude will have consequences when closure errors arise in long leveling lines that run mostly north-south compared to those that run mostly east-west. The most modern reference ellipsoids are GRS 80 and WGS 84. As given by (Blakely 1995, p.136), the closed-form formula for WGS 84 normal gravity is:

$$g_0 = 9.7803267714 \frac{1 + 0.00193185138639 \sin^2 \phi}{\sqrt{1 - 0.00669437999013 \sin^2 \phi}} \tag{1.2}$$

Figure 1.1 shows a plot of the difference between Equation 1.1 and Equation 1.2. The older model has a larger value throughout and has, in the worst case, a magnitude greater by 0.000163229 m/s^2 (i.e., about 16 mgals) at the equator.

1.3.3 Equipotential Ellipsoids as Vertical Datums

Concerning the topic of this paper, perhaps the most important consequence of the differences between local and equipotential ellipsoids is that equipotential ellipsoids are more suitable to be used as vertical datums in the ordinary sense than local ellipsoids and, in fact, they are used as such. In particular, GPS-derived coordinates expressed as geodetic latitude and longitude present the third dimension as an ellipsoid height. This constitutes a dramatic change from the past. Before, ellipsoid heights were essentially unheard of, basically only of interest and of use to geodesists for computational purposes. Now, anyone using a GPS is deriving ellipsoid heights.

Equipotential ellipsoids are models of the gravity that would result from a highly idealized model of the Earth; one whose mass is distributed homogeneously but includes the Earth's oblate shape, and spinning like the Earth. The geoid is not a simple surface compared to an equipotential ellipsoid, which can be completely described by just the four parameters listed above. The geoid's shape is strongly influenced by the topographic surface of the Earth. As seen in Figure 1.2, the geoid appears to be "bumpy," with apparent mountains, canyons, and valleys. This is, in fact, not so. The geoid is a convex surface by virtue of satisfying the Laplace equation, and its apparent concavity is a consequence of how the geoid is portrayed on a flat surface (Vaníček & Krakiwsky 1986). Figure 1.2 is a portrayal of the ellipsoid height of the geoid as estimated by GEOID 03 (Roman, Wang, Henning, & Hamilton 2004). That is to say, the heights shown in the figure are the distances from

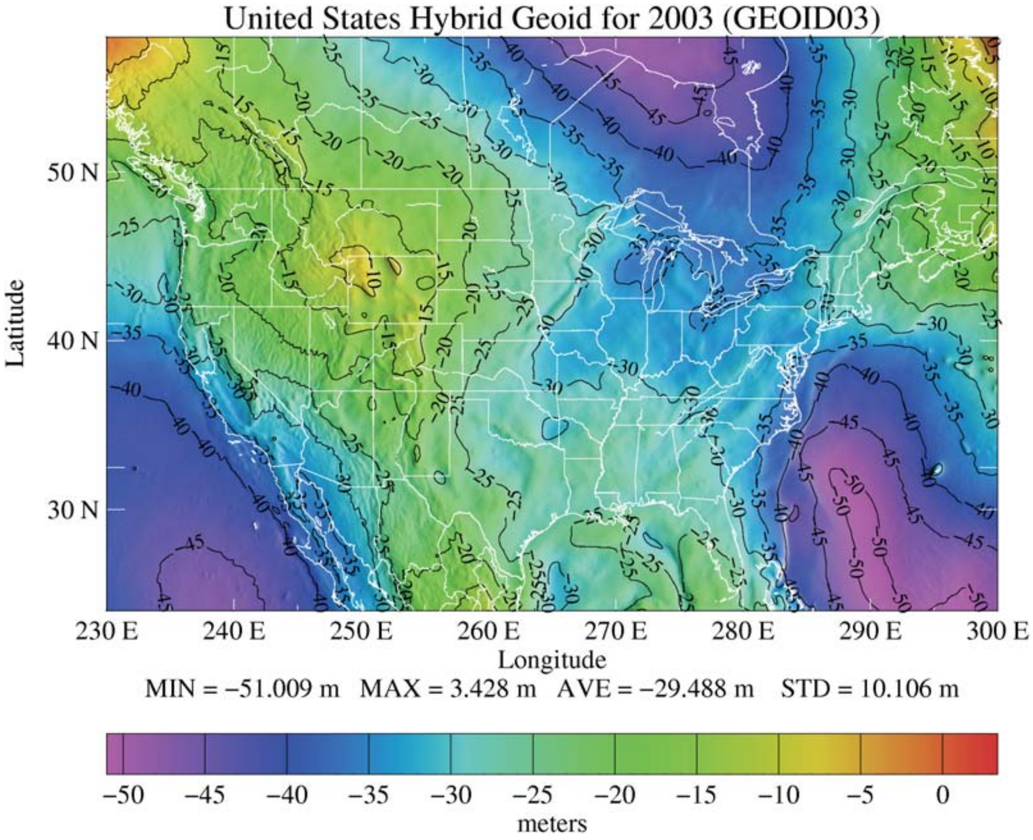
United States Hybrid Geoid for 2003 (GEOID03)
Latitude
50 N
40 N
30 N
230 E
240 E
250 E
260 E
270 E
280 E
290 E
300 E
Longitude
MIN = −51.009 m MAX = 3.428 m AVE = −29.488 m STD = 10.106 m
−50
−45
−40
−35
−30
−25
−20
−15
−10
−5
0
meters

Figure 1.2: Geoid heights with respect to NAD 83/GRS 80 over the continental United States as computed by GEOID03. Source: (NGS 2003).

GRS 80 as located by NAD 83 to the geoid; the ellipsoid height of the geoid. Such heights (the ellipsoid height of a place on the geoid) are called geoid heights. Thus, Figure 1.2 is a picture of geoid heights.

Even though equipotential ellipsoids are useful as vertical datums, they are usually unsuitable as a surrogate for the geoid when measuring orthometric heights. Equipotential ellipsoids are "best-fit" over the entire Earth and, consequently, they typically do not match the geoid particularly well in any specific place. For example, as shown in Figure 1.2, GRS 80 as placed by NAD 83 is everywhere higher than the geoid across the conterminous United States; not half above and half below. Furthermore, as described above, equipotential ellipsoids lack the small-scale details of the geoid. And, like local ellipsoids, ellipsoid heights reckoned from equipotential ellipsoids also suffer from the phenomenon that there are places where water apparently flows "uphill," although perhaps not as badly as some local ellipsoids. Therefore, surveyors using GPS to determine heights would seldom want to use ellipsoid heights. In most cases, surveyors need to somehow deduce an orthometric height from an ellipsoid height, which will be discussed in the following papers.

1.4 Mean Sea Level

One of the ultimate goals of this series is to present a sufficiently complete presentation of orthometric heights that the following definition will be clear. In the *NGS Glossary*, the term **orthometric height** is referred to **elevation, orthometric**, which is defined as, "The distance between the geoid and a point measured along the **plumb line** and taken positive upward from the geoid." For contrast, we quote from the first definition for **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.
> The surface usually specified is the geoid or an approximation thereto. Mean sea level was long considered a satisfactory approximation to the geoid and therefore suitable for use as a reference surface. It is now known that mean sea level can differ from the geoid by up to a meter or more, but the exact difference is difficult to determine.
> The terms **height** and **level** are frequently used as synonyms for elevation. In geodesy, height also refers to the distance above an ellipsoid. . .

It happens that lying within these two definitions is a remarkably complex situation primarily concerned with the Earth's gravity field and our attempts to make measurements using it as a frame of reference. The terms **geoid, plumb line, potential, mean sea level** have arisen, and they must be addressed before discussing orthometric heights.

For heights, the most common datum is mean sea level. Using mean sea level for a height datum is perfectly natural because most human activity occurs at or above sea level. However, creating a workable and repeatable mean sea level datum is somewhat subtle. The *NGS Glossary* definition of mean sea level is "The average location of the interface between ocean and atmosphere, over a period of time sufficiently long so that all random and periodic variations of short duration average to zero."

The National Oceanic and Atmospheric Administration's (NOAA) National Ocean Service (NOS) Center for Operational Oceanographic Products and Services (CO-OPS) has set 19 years as the period suitable for measurement of mean sea level at tide gauges (National Geodetic Survey 1986, p. 209). The choice of 19 years was chosen because it is the smallest integer number of years larger than the first major cycle of the moon's orbit around the Earth. This accounts for the largest of the periodic effects mentioned in the definition. See Bomford (1980, pp. 247-255) and Zilkoski (2001) for more details about mean sea level and tides. Local mean sea level is often

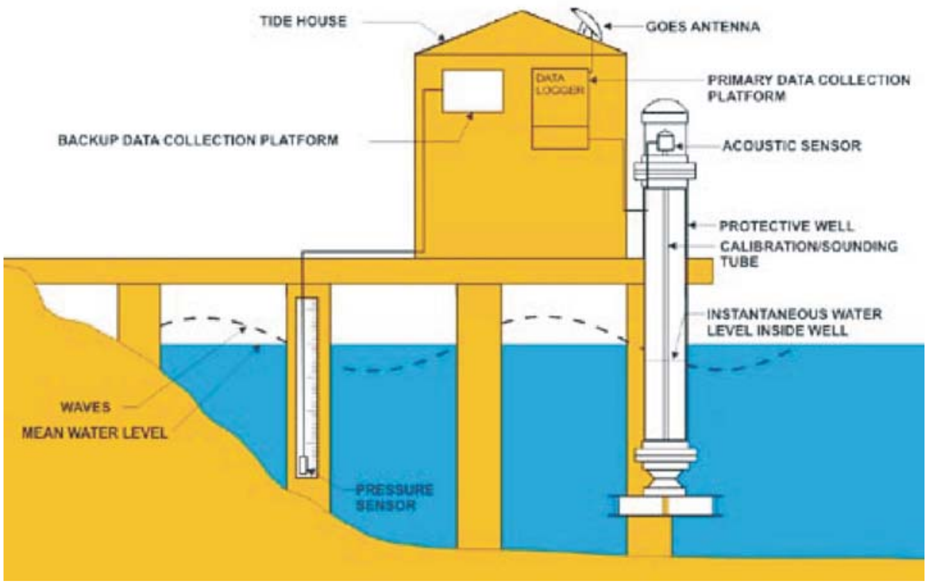
TIDE HOUSE
GOES ANTENNA
DATA LOGGER
PRIMARY DATA COLLECTION PLATFORM
BACKUP DATA COLLECTION PLATFORM
ACOUSTIC SENSOR
PROTECTIVE WELL
CALIBRATION/SOUNDING TUBE
INSTANTANEOUS WATER LEVEL INSIDE WELL
WAVES
MEAN WATER LEVEL
PRESSURE SENSOR

Figure 1.3: The design of a NOAA tide house and tide gauge used for measuring mean sea level. Source: (NOAA 2007).

measured using a tide gauge. Figure 1.3 depicts a tide house, “a structure that houses instruments to measure and record the instantaneous water level inside the tide gauge and built at the edge of the body of water whose local mean level is to be determined.”

It has been suspected at least since the time of the building of the Panama Canal that mean sea level might not be at the same height everywhere (McCullough 1978). The original canal, attempted by the French, was to be cut at sea level and there was concern that the Pacific Ocean might not be at the same height as the Atlantic, thereby causing a massive flood through the cut. This concern became irrelevant when the sea level approach was abandoned. However, the subject surfaced again in the creation of the National Geodetic Vertical Datum of 1929 (NGVD 29).

By this time it was a known fact that not all mean sea-level stations were the same height, a proposition that seems absurd on its face. To begin with, all mean sea-level stations are at an elevation of zero by definition. Second, water seeks its own level, and the oceans have no visible constraints preventing free flow between the stations (apart from the continents), so how could it be possible that mean sea level is not at the same height everywhere? The answer lies in differences in temperature, chemistry, ocean currents, and ocean eddies.

The water in the oceans is constantly moving at all depths. Seawater at different temperatures contains different amounts of salt and, consequently, has density gradients. These density gradients give rise to immense deep-ocean cataracts that constantly transport massive quantities of water from the poles to the tropics and back (Broecker 1983, Ingle 2000, Whitehead 1989). The sun’s warming of surface waters causes the global-scale currents that are well-known to mariners in addition to other more subtle effects (Chelton, Schlax, Freilich & Milliff 2004). Geostrophic effects cause large-scale, persistent ocean eddies that push water against or away from the continents, depending on the direction of the eddy’s circulation. These effects can create sea surface topographic variations of more than 50 centimeters (Srinivasan 2004). As described by Zilkoski (2001, p.40), the differences are due to “. . . currents, prevailing winds and barometric pressures, water temperature and salinity differentials, topographic configuration of the bottom in the area of the gauge site, and other physical causes . . .”

In essence, these factors push the water and hold it upshore or away-from-shore further than would be the case under the influence of gravity alone. Also, the persistent nature of these climatic factors prevents the elimination of their effect by averaging (e.g., see (Speed, Jr., Newton & Smith 1996*b*, Speed, Jr., Newton & Smith 1996*a*)). As will be discussed in more detail in the second paper, this gives rise to the seemingly paradoxical state that holding one sea-level station as a zero height reference and running levels to another station generally indicates that the other station is not also at zero height, even in the absence of experimental error and even if the two stations *are at the same gravitational potential.* Similarly, measuring the height of an inland benchmark using two level lines that start from different tide gauges generally results in two statistically different height measurements. These problems were addressed in different ways by the creation of two national vertical datums, NGVD 29 and North American Vertical Datum of 1988 (NAVD 88). We will now discuss the national vertical datums of the United States.

1.5 U.S. National Vertical Datums

The first leveling route in the United States considered to be of geodetic quality was established in 1856-57 under the direction of G.B. Vose of the U.S. Coast Survey, predecessor of the U.S. Coast and Geodetic Survey and, later, the National Ocean Service.[2] The leveling survey was needed to support current and tide studies in the New York Bay and Hudson River areas. The first leveling line officially designated as “geodesic leveling” by the Coast and Geodetic Survey followed an arc of triangulation along the 39th parallel. This 1887 survey began at benchmark A in Hagerstown, Maryland.

By 1900, the vertical control network had grown to 21,095 km of geodetic leveling. A reference surface was determined in 1900 by holding elevations referenced to local mean sea level (LMSL) fixed at five tide stations. Data from two other tide stations indirectly influenced the determination of the reference surface. Subsequent readjustments of the leveling network were performed by the Coast and Geodetic Survey in 1903, 1907, and 1912 (Berry 1976).

1.5.1 National Geodetic Vertical Datum of 1929 (NGVD 29)

The next general adjustment of the vertical control network, called the Sea Level Datum of 1929 and later renamed to the National Geodetic Vertical Datum of 1929 (NGVD 29), was accomplished in 1929. By then, the international nature of geodetic networks was well understood, and Canada provided data for its first-order vertical network to combine with the U.S. network. The two networks were connected at 24 locations through vertical control points (benchmarks) from Maine/New Brunswick to Washington/British Columbia. Although Canada did not adopt the “Sea Level Datum of 1929” determined by the United States, Canadian-U.S. cooperation in the general adjustment greatly strengthened the 1929 network. Table 1.1 lists the kilometers of leveling involved in the readjustments and the number of tide stations used to establish the datums.

It was mentioned above that NGVD 29 was originally called the “Sea Level Datum of 1929.” To eliminate some of the confusion caused by the original name, in 1976 the name of the datum was changed to “National Geodetic Vertical Datum of 1929,” eliminating all reference to “sea level” in the title. This was a change in name only; the mathematical and physical definitions of the datum established in 1929 were not changed in any way.

[2] This section consists of excerpts from Chapter 2 of Maune’s (2001) Vertical Datums.

Year of Adjustment	Kilometers of Leveling	Number of Tide Stations
1900	21095	5
1903	31789	8
1907	38359	8
1912	46468	9
1929	75159 (U.S.)	21 (U.S.)
	31565 (Canada)	5 (Canada)

Table 1.1: Amount of leveling and number of tide stations involved in previous readjustments.

1.5.2 North American Vertical Datum of 1988 (NAVD 88)

The most recent general adjustment of the U.S. vertical control network, which is known as the North American Vertical Datum of 1988 (NAVD 88), was completed in June 1991 (Zilkoski, Richards & Young 1992). Approximately 625,000 km of leveling have been added to the NSRS since NGVD 29 was created. In the intervening years, discussions were held periodically to determine the proper time for the inevitable new general adjustment. In the early 1970s, the National Geodetic Survey conducted an extensive inventory of the vertical control network. The search identified thousands of benchmarks that had been destroyed, due primarily to post-World War II highway construction, as well as other causes. Many existing benchmarks were affected by crustal motion associated with earthquake activity, post-glacial rebound (uplift), and subsidence resulting from the withdrawal of underground liquids.

An important feature of the NAVD 88 program was the re-leveling of much of the first-order NGS vertical control network in the United States. The dynamic nature of the network requires a framework of newly observed height differences to obtain realistic, contemporary height values from the readjustment. To accomplish this, NGS identified 81,500 km (50,600 miles) for re-leveling. Replacement of disturbed and destroyed monuments preceded the actual leveling. This effort also included the establishment of stable "deep rod" benchmarks, which are now providing reference points for new GPS-derived orthometric height projects as well as for traditional leveling projects. The general adjustment of NAVD 88 consisted of 709,000 unknowns (approximately 505,000 permanently monumented benchmarks and 204,000 temporary benchmarks) and approximately 1.2 million observations.

Analyses indicate that the overall differences for the conterminous United States between orthometric heights referred to NAVD 88 and NGVD 29 range from 40 cm to +150 cm. In Alaska the differences range from approximately +94 cm to +240 cm. However, in most "stable" areas, relative height changes between adjacent benchmarks appear to be less than 1 cm. In many areas, a single bias factor, describing the difference between NGVD 29 and NAVD 88, can be estimated and used for most mapping applications (NGS has developed a program called VERTCON to convert from NGVD 29 to NAVD 88 to support mapping applications). The overall differences between dynamic heights referred to International Great Lakes Datum of 1985 (IGLD 85) and IGLD 55 range from 1 cm to 37 cm.

1.5.3 International Great Lakes Datum of 1985 (IGLD 85)

For the general adjustment of NAVD 88 and the International Great Lakes Datum of 1985 (IGLD 85), a minimum constraint adjustment of Canadian-Mexican-U.S. leveling observations was performed. The height of the primary tidal benchmark at Father Point/Rimouski, Quebec, Canada (also used in the NGVD 1929 general adjustment), was held fixed as the constraint. Therefore, IGLD 85 and NAVD 88 are one and the same. Father Point/Rimouski is an IGLD water-level station located at the mouth of the St. Lawrence River and is the reference station used for IGLD 85. This constraint satisfied the requirements of shifting the datum vertically to minimize the impact of NAVD 88 on U.S. Geological Survey (USGS) mapping products, and it provides the datum point desired by the IGLD Coordinating Committee for IGLD 85. The only difference between IGLD 85 and NAVD 88 is that IGLD 85 benchmark values are given in dynamic height units, and NAVD 88 values are given in Helmert orthometric height units. Geopotential numbers for individual benchmarks are the same in both systems (the next two papers will explain dynamic heights, geopotential numbers, and Helmert orthometric heights).

1.5.4 Tidal Datums

Principal Tidal Datums

A vertical datum is called a tidal datum when it is defined by a certain phase of the tide. Tidal datums are local datums and are referenced to nearby monuments. Since a tidal datum is defined by a certain phase of the tide there are many different types of tidal datums. This section will discuss the principal tidal datums that are typically used by federal, state, and local government agencies: Mean Higher High Water (MHHW), Mean High Water (MHW), Mean Sea Level (MSL), Mean Low Water (MLW), and Mean Lower Low Water (MLLW).

A determination of the principal tidal datums in the United States is based on the average of observations over a 19-year period, e.g., 1988-2001. A specific 19-year Metonic cycle is denoted as a National Tidal Datum Epoch (NTDE). CO-OPS publishes the official United States local mean sea level values as defined by observations at the 175 station National Water Level Observation Network (NWLON). Users need to know which NTDE their data refer to.

- Mean Higher High Water (MHHW): MHHW is defined as the arithmetic mean of the higher high water heights of the tide observed over a specific 19-year Metonic cycle denoted as the NTDE. Only the higher high water of each pair of high waters of a tidal day is included in the mean. For stations with shorter series, a comparison of simultaneous observations is made with a primary control tide station in order to derive the equivalent of the 19-year value (Marmer 1951).

- Mean High Water (MHW) is defined as the arithmetic mean of the high water heights observed over a specific 19-year Metonic cycle. For stations with shorter series, a computation of simultaneous observations is made with a primary control station in order to derive the equivalent of a 19-year value (Marmer 1951).

- Mean Sea Level (MSL) is defined as the arithmetic mean of hourly heights observed over a specific 19-year Metonic cycle. Shorter series are specified in the name, such as monthly mean sea level or yearly mean sea level (e.g., (Marmer 1951, Hicks 1985)).

- Mean Low Water (MLW) is defined as the arithmetic mean of the low water heights observed over a specific 19-year Metonic cycle. For stations with shorter series, a comparison of simultaneous observations is made with a primary control tide station in order to derive the equivalent of a 19-year value (Marmer 1951).

- Mean Lower Low Water (MLLW) is defined as the arithmetic mean of the lower low water heights of the tide observed over a specific 19-year Metonic cycle. Only the lower low water of each pair of low waters of a tidal day is included in the mean. For stations with shorter series, a comparison of simultaneous observations is made with a primary control tide station in order to derive the equivalent of a 19-year value (Marmer 1951).

PBM 180 1946	—	5.794 m (the Primary Bench Mark)
Highest Water Level	—	4.462 m
MHHW	—	3.536 m
MHW	—	3.353 m
MTL	—	2.728 m
MSL	—	2.713 m
DTL	—	2.646 m
NGVD 1929	—	2.624 m
MLW	—	2.103 m
NAVD 88	—	1.802 m
MLLW	—	1.759 m
Lowest Water Level	—	0.945 m

Table 1.2: Various Tidal Datums and Vertical Datums for PBM 180 1946.

Other Tidal Datums

Other tidal values typically computed include the Mean Tide Level (MTL), Diurnal Tide Level (DTL), Mean Range (Mn), Diurnal High Water Inequality (DHQ), Diurnal Low Water Inequality (DLQ), and Great Diurnal Range (Gt).

- Mean Tide Level (MTL) is a tidal datum which is the average of Mean High Water and Mean Low Water.
- Diurnal Tide Level (DTL) is a tidal datum which is the average of Mean Higher High Water and Mean Lower Low Water.
- Mean Range (Mn) is the difference between Mean High Water and Mean Low Water.
- Diurnal High Water Inequality (DHQ) is the difference between Mean Higher High Water and Mean High Water.
- Diurnal Low Water Inequality (DLQ) is the difference between Mean Low Water and Mean Lower Low Water.
- Great Diurnal Range (Gt) is the difference between Mean Higher High Water and Mean Lower Low Water.

All of these tidal datums and differences have users that need a specific datum or difference for their particular use. The important point for users is to know which tidal datum their data are referenced to. Like geodetic vertical datums, local tidal datums are all different from one another, but they can be related to each other. The relationship of a local tidal datum (941 4290, San Francisco, California) to geodetic datums is illustrated in Table 1.2.

Please note that in this example, NAVD 88 heights, which are the official national geodetic vertical control values, and LMSL heights, which are the official national local mean sea level values, at the San Francisco tidal station differ by almost one meter. Therefore, if a user obtained a set of heights relative to the local mean sea level and a second set referenced to NAVD 88, the two sets would disagree by about one meter due to the datum difference. In addition, the difference between MHW and MLLW is more than 1.5 m (five feet). Due to regulations and laws, some users relate their data to MHW, while others relate their data to MLLW. As long as a user knows which datum the data are referenced to, the data can be converted to a common reference and the data sets can be combined.

1.6 Summary

This is the first in a four-part series of papers that will review the fundamental concept of height. The National Geodetic Survey will not, in the future, create or maintain elevation benchmarks by leveling. Instead, NGS will assign vertical control by estimating orthometric heights from ellipsoid heights as computed from GPS measurements. This marks a significant shift in how the United States' vertical control is created and maintained. Furthermore, practicing surveyors and mappers who use GPS are now confronted with using ellipsoid heights in their everyday work, something that was practically unheard of before GPS. The relationship between ellipsoid heights and orthometric heights is not simple, and it is the purpose of this series of papers to examine that relationship.

This first paper reviewed reference ellipsoids and mean sea level datums. Reference ellipsoids are models of the Earth's shape and fall into two distinct categories: local and equipotential. Local reference ellipsoids were created by continental-sized triangulation networks and were employed as a computational surface but not as a vertical datum in the ordinary sense. Local reference ellipsoids are geometric in nature; their size and shape were determined by purely geometrical means. They were also custom-fit to a particular locale due to the impossibility of observing stations separated by oceans. Equipotential ellipsoids include the geometric considerations of local reference ellipsoids, but they also include information about the Earth's mass and rotation. They model the mean sea level equipotential surface that would result from both the redistribution of the Earth's mass caused by its rotation, as well as the centripetal effect of the rotation. It is purely a mathematical construct derived from observed physical parameters of the Earth. Unlike local reference ellipsoids, equipotential ellipsoids are routinely used as a vertical datum. Indeed, all heights directly derived from GPS measurements are ellipsoid heights.

Even though equipotential ellipsoids are used as vertical datums, most practicing surveyors and mappers use orthometric heights, not ellipsoid heights. The first national mean sea level datum in the United States was the NGVD 29. NGVD 29 heights were assigned to fiducial benchmarks through a least-squares adjustment of local height networks tied to separate tide gauges around the nation. It was observed at that time that mean sea level was inconsistent through these stations on the order of meters, but the error was blurred through the network statistically. The most recent general adjustment of the U.S. network, which is known as NAVD 88, was completed in June 1991. Only a single tide gauge was held fixed in NAVD 88 and, consequently, the inconsistencies between tide gauges were not distributed through the network adjustment, but there will be a bias at each mean sea level station between NAVD 88 level surface and mean sea level.

Chapter 2

Physics and Gravity

2.1 Preamble

This monograph was originally published as a series of four articles appearing in the Surveying and Land Information Science. Each chapter corresponds to one of the original papers[1]. This paper should be cited as

Meyer, Thomas H., Roman, Daniel R., and Zilkoski, David B. (2005) What does *height* really mean? Part II: Physics and Gravity. In *Surveying and Land Information Science*, 65(1): 5-15.

This is the second paper in a four-part series considering the fundamental question, "what does the word *height* really mean?" The first paper in this series explained that a change in National Geodetic Survey's policy, coupled with the modern realities of GPS surveying, have essentially forced practicing surveyors to come to grips with the myriad of height definitions that previously were the sole concern of geodesists. The distinctions between local and equipotential ellipsoids were considered, along with an introduction to mean sea level. This paper brings these ideas forward by explaining mean sea level and, more importantly, the geoid. The discussion is grounded in physics from which gravitational force and potential energy will be considered, leading to a simple derivation of the shape of the Earth's gravity field. This lays the foundation for a simplistic model of the geoid near Mt. Everest, which will be used to explain the undulations in the geoid across the entire Earth. The terms **geoid, plumb line, potential, equipotential surface, geopotential number**, and **mean sea level** will be explained, including a discussion of why mean sea level is not everywhere the same height; why it is not a level surface.

2.2 Introduction: Why Care About Gravity?

Any instrument that needs to be leveled in order to properly measure horizontal and vertical angles depends on gravity for orientation. Surveying instruments that measure gravity-referenced heights depend upon gravity to define their datum. Thus, many surveying measurements depend upon and are affected by gravity. This second paper in the series will develop the physics of gravity, leading to an explanation of the geoid and geopotential numbers.

The direction of the Earth's gravity field stems from the Earth's rotation and the mass distribution of the planet. The inhomogeneous distribution of that mass causes what are known as geoid undulations, the geoid being defined by the National Geodetic Survey (1986) as 'The equipotential surface of the Earth's gravity field which best fits, in a least squares sense, global mean sea level."

[1] Throughout the series we will enumerate figures, tables, and equations with an Arabic numeral indicating the paper in the series from which it came. For example, the third figure in the second paper will be numbered, "Figure 2.3".

The geoid is also called the "figure of the Earth." Quoting Shalowitz (1938, p. 10), "The true figure of the Earth, as distinguished from its topographic surface, is taken to be that surface which is everywhere perpendicular to the direction of the force of gravity and which coincides with the mean surface of the oceans." The direction of gravity varies in a complicated way from place to place. Local vertical remains perpendicular to this undulating surface, whereas local normal remains perpendicular to the ellipsoid reference surface. The angular difference of these two is the **deflection of the vertical**.

The deflection of the vertical causes angular traverse loop misclosures, as do instrument setup errors, the Earth's curvature, and environmental factors introducing errors into measurements. The practical consequence of the deflection of the vertical is that observed angles differ from the angles that result from the pure geometry of the stations. It is as if the observing instrument were misleveled, resulting in traverses that do not close. This is true for both plane and geodetic surveying, although the effect for local surveys is seldom measurable because geoid undulations are smooth and do not vary quickly over small distances. Even so, it should be noted that the deflection of the vertical can cause unacceptable misclosures even over short distances.

For example, Shalowitz (1938, pp. 13,14) reported deflections of the vertical created discrepancies between astronomic coordinates and geodetic (computed) coordinates up to a minute of latitude in Wyoming. In all cases, control networks for large regions cannot ignore these discrepancies, and remain geometrically consistent, especially in and around regions of great topographic relief. Measurements made using a gravitational reference frame are reduced to the surface of a reference ellipsoid to remove the effects of the deflection of the vertical, skew of the normals, topographic enlargement of distances, and other environmental effects (Meyer 2002).

The first article in this series introduced the idea that mean sea level is not at the same height in all places. This fact led geodesists to a search for a better surface than mean sea level to serve as the datum for vertical measurements, and that surface is the geoid. Coming to a deep understanding of the geoid requires a serious inquiry (Blakely 1995, Bomford 1980, Heiskanen & Moritz 1967, Kellogg 1953, Ramsey 1981, Torge 1997, Vaníček & Krakiwsky 1986), but the concepts behind the geoid can be developed without having to examine all the details. The heart of the matter lies in the relationship between gravitational force and gravitational potential. Therefore, we review the concepts of force, work, and energy so as to develop the framework to consider this relationship.

2.3 Physics

2.3.1 Force, Work, and Energy

Force is what makes things go. This is apparent from Newton's law, $\mathbf{F} = m\mathbf{a}$, which gives that the acceleration of an object is caused by, and is in the direction of, a force $\mathbf{F}$ and is inversely proportional to the object's mass m. Force has magnitude (i.e., strength) and direction. Therefore, a force is represented mathematically as a vector whose length and direction are set equal to those of the force. We denote vectors in bold face, either upper or lower case, e.g., $\mathbf{F}$ or $\mathbf{f}$, and scalars in standard face, e.g., the speed of light is commonly denoted as c. Force has units of mass times length per second squared and is named the "newton," abbreviated N, in the meter-kilogram-second (mks) system.

There is a complete algebra and calculus of vectors (e.g., see (Davis & Snider 1979) or (Marsden & Tromba 1988)), which will not be reviewed here. However, we remind the reader of certain key concepts. Vectors are ordered sets of scalar components, e.g., (x, y, z) or $\mathbf{F} = (F_1, F_2, F_3)$, and we take the magnitude of a vector, which we denote as $|\mathbf{F}|$, to be the square root of the sum of the components: For example, if $F = (1, -4, 2)$, then $|\mathbf{F}| = \sqrt{1^2 + (-4)^2 + 2^2} = \sqrt{21}$.

Vectors can be multiplied by scalars (e.g., $c\ \mathbf{A}$) and, in particular, the negative of a vector is defined as the scalar product of minus one with the vector: -$\mathbf{A}$ = -1 $\mathbf{A}$. It is easy to show that -$\mathbf{A}$ is a vector of magnitude equal to $\mathbf{A}$ but oriented in the opposite direction. Division of vectors by scalars is simply scalar multiplication by a reciprocal: $\mathbf{F}/c = (1/c)\ \mathbf{F}$. A vector $\mathbf{F}$ divided by its own length results in a unit vector, being a vector in the same direction as $\mathbf{F}$ but having unit length-a length of exactly one. We denote a unit vector with a hat: $\hat{\mathbf{F}} = \mathbf{F}/|\mathbf{F}|$.

Vectors can be added (e.g., $\mathbf{A} + \mathbf{B}$) and subtracted, although subtraction is defined in terms of scalar multiplication by -1 and vector addition (i.e., $\mathbf{A} - \mathbf{B} = \mathbf{A} + (-\mathbf{B})$). The result of adding/subtracting two vectors is another vector; likewise with scalar multiplication. By virtue of vector addition (the law of superposition), any vector can be a composite of any finite number of vectors: $\mathbf{F} = \sum_{i=1}^{n} \mathbf{f}_i,\ n < \infty$.

The **inner** or **scalar** product of two vectors **a.b** is defined as:

$$\mathbf{a}.\mathbf{b} = |\mathbf{a}| \times \mathbf{b}| \cos\theta \tag{2.1}$$

where θ is the angle between **a** and **b** in the plane that contains them. In particular, note that if **a** is perpendicular to **b**, then **a.b** = 0 because cos 90° = 0. We will make use of the fact that the inner product of a force vector with a unit vector is a scalar equal to the magnitude of the component of the force that is applied in the direction of the unit vector.

Newton's law of gravity specifies that the gravitational force exerted by a mass M on a mass m is:

$$\mathbf{F}_g = -\frac{GMm\hat{\mathbf{r}}}{|\mathbf{r}|^2} \tag{2.2}$$

where:
G = universal gravitational constant; and
$\mathbf{r}$ = a vector from M's center of mass to m's center of mass.
The negative sign accounts for gravity being an attractive force by orienting $\mathbf{F}_g$ in the direction opposite of $\hat{\mathbf{r}}$ (since $\hat{\mathbf{r}}$ is the unit vector from M to m, $\mathbf{F}_g$ needs to be directed from m to M). In light of the discussion above about vectors, Equation 2.2 is understood to indicate that the magnitude of gravitational force is in proportion to the masses of the two objects, inversely proportional to the square of the distance separating them, and is directed along the straight line joining their centroids.

In geodesy, M usually denotes the mass of the Earth and, consequently, the product GM arises frequently. Although the values for G and M are known independently (G has a value of approximately 6.67259×10^{-11} m^3 s^{-2} kg^{-1} and M is approximately 5.9737×10^{24} kg), their product can be measured as a single quantity and its value has been determined to have several, nearly identical values, such as $GM = 398600441.5 \pm 0.8 \times 10^6$ m^3 s^2 (Groten 2004).

Gravity is a force field, meaning that the gravity created by any mass permeates all of space. One consequence of superposition is that gravity fields created by different masses are independent of one another. Therefore, it is reasonable and convenient to consider the gravitational field created by a single mass without taking into consideration any objects within that field. Equation 2.2 can be modified to describe a gravitational field simply by omitting m. We can compute the strength of the Earth's gravitational field at a distance equal to the Earth's equatorial radius (6,378,137 m) from the center of M by:

$$\begin{aligned}\mathbf{E}_g &= -\frac{GM\hat{\mathbf{r}}}{|\mathbf{r}|^2} && (2.3)\\ &= -\frac{398600441.5\ \mathrm{m}^3\mathrm{s}^2\hat{\mathbf{r}}}{(6378137\ \mathrm{m/s})^2} \\ &= 9.79829\ \mathrm{m/s}^2(-\hat{\mathbf{r}}) && (2.4)\end{aligned}$$

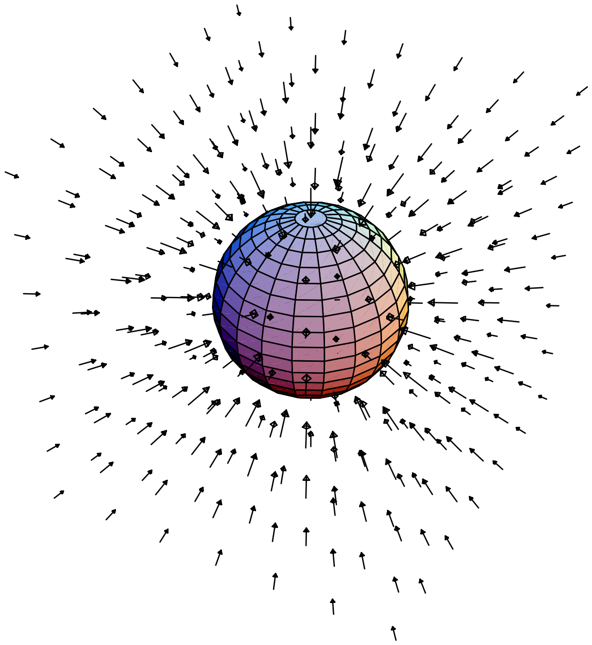

Figure 2.1: The gravitational force field of a spherical Earth. Note that the magnitude of the force decreases with separation from the Earth.

This value is slightly larger than the well-known value of 9.78033 m/s^2 because the latter includes the effect of the Earth's rotation.[2] We draw attention to the fact that Equation 2.3 has units of acceleration, not a force, by virtue of having omitted m.

It is possible to use Equation 2.3 to draw a picture that captures, to some degree, the shape of the Earth's gravitational field (see Figure 2.1. The vectors in the figure indicate the magnitude and direction of force that would be experienced by unit mass located at that point in space. The vectors decrease in length as distance increases away from the Earth and are directly radially toward the Earth's center, as expected. However, we emphasize that the Earth's gravitational field pervades all of space; it is not discrete as the figure suggests. Furthermore, it is important to realize that, in general, any two points in space experience a different gravitational force, if perhaps only in direction.

We remind the reader that the current discussion is concerned with finding a more suitable vertical datum than mean sea level, which is, in some sense, the same thing as finding a better way to measure heights. Equation 2.3 suggests that height might be inferred by measuring gravitational force because Equation 2.3 can be solved for the magnitude of r, which would be a height measured using the Earth's center of gravity as its datum. At first, this approach might seem to hold promise because the acceleration due to gravity can be measured with instruments that carefully measure the acceleration of a standard mass, either as a pendulum or free falling (Faller & Vitouchkine 2003). It seems such a strategy would deduce height in a way that stems from the physics that give rise to water's downhill motion and, therefore, would capture the primary motivating concept behind height very well. Regrettably, this is not the case and we will now explain why.

[2] The gravity experienced on and around the Earth is a combination of the gravitation produced by the Earth's mass and the centrifugal force created by its rotation. The force due solely to the Earth's mass is called **gravitational** and the combined force is called **gravity**. For the most part, it will not be necessary for the purposes of this paper to draw a distinction between the two. The distinction will be emphasized where necessary.

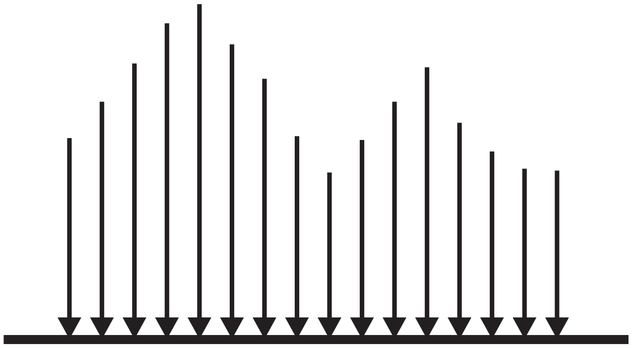

Figure 2.2: A collection of force vectors that are all normal to a surface (indicated by the horizontal line) but of differing magnitudes. The horizontal line is a level surface because all the vectors are normal to it; they have no component directed across the surface.

Suppose we use gravitational acceleration as a means of measuring height. This implies that surfaces of equal acceleration must also be level surfaces, meaning a surface across which water does not run without external impetus. Thus, our mean sea level surrogate is that set of places that experience some particular gravitational acceleration; perhaps the acceleration of the normal gravity model, g_0, would be a suitable value. The fallacy in this logic comes from the inconsideration of gravity as a vector; it is not just a scalar. In fact, the heart of the matter lies not in the *magnitude* of gravity but, rather, in its *direction*.

If a surface is level, then water will not flow across it due to the influence of gravity alone. Therefore, a level surface must be situated such that all gravity force vectors at the surface are perpendicular to it; none of the force vectors can have any component directed across the surface. Figure 2.2 depicts a collection of force vectors that are mutually perpendicular to a horizontal surface, so the horizontal surface is level, but the vectors have differing magnitudes. Therefore, it is apparent that choosing a surface of equal gravitational acceleration (i.e., magnitude) does not guarantee that the surface will be level. Of course, we have not shown that this approach necessarily would not produce level surfaces. It might be the case that it happens that the magnitude of gravity acceleration vectors just happen to be equal on level surfaces. However, as we will show below, this is not the case due to the inhomogeneous distribution of mass within the Earth.

We can use this idea to explain why the surface of the oceans is not everywhere the same distance to the Earth's center of gravity. The first article in this series noted several reasons for this, but we will discuss only one here. It is known that the salinity in the oceans is not constant. Consequently, the density of the water in the oceans is not constant, either, because it depends on the salinity. Suppose we consider columns of water along a coast line and suppose that gravitational acceleration is constant along the coasts (see Figure 2.3). In particular, consider the columns A and B. Suppose the water in column A is less dense than in column B; perhaps a river empties into the ocean at that place. We have assumed or know that:

- The force of gravity is constant,
- The columns of water must have the same weight in order to not flow, and
- The water in column A is less dense than that in column B.

It takes more water of lesser density to have the same mass as the amount of water needed of greater density. Water is nearly incompressible, so the water column at A must be taller than the

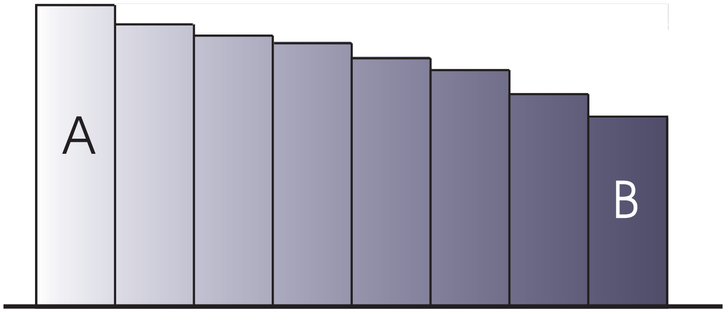

Figure 2.3: A collection water columns whose salinity, and therefore density, has a gradient from left to right. The water in column A is least dense. Under constant gravity, the height of column A must be greater than B so that the mass of column A equals that of column B.

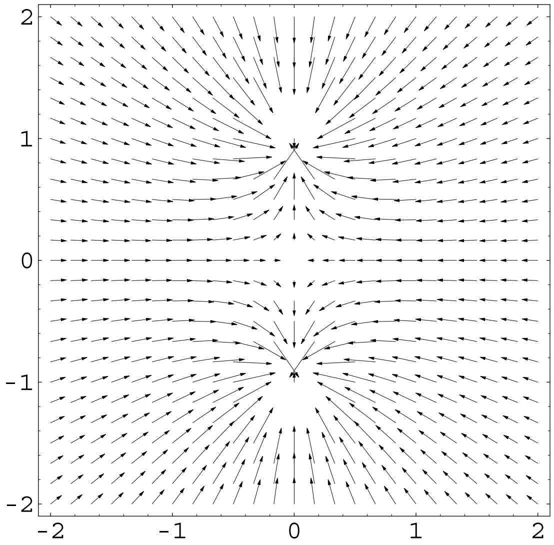

Figure 2.4: The force field created by two point masses.

column of water at B. Therefore, a mean sea level station at A would not be at the same distance from the Earth's center of gravity as a mean sea level station at B.

As another example showing why gravitational force is not an acceptable way to define level surfaces, Figure 2.4 shows the force field generated by two point-unit masses located at (0,1) and (0,-1). Note the lines of symmetry along the x and y axes. All forces for places on the x-axis are parallel to the axis and directed towards (0,0). Above or below the x-axis, all force lines ultimately lead to the mass also located on that side. Figure 2.5 shows a plot of the magnitude of the vectors of Figure 2.4. Note the local maxima around $x = \pm 1$ and the local minima at the origin. Figure 2.6 is a plot of the "north-east" corner of the force vectors superimposed on top of an isoforce plot of their magnitudes (i.e., a "contour plot" of Figure 2.5). Note that the vectors are not perpendicular to the isolines. If one were to place a drop of water anywhere in the space illustrated by the figure, the water would follow the vectors to the peak and would both follow and cross isoforce lines, which is nonsensical if we take isoforce lines to correspond to level surfaces. This confirms that equiforce surfaces are not level.

These three examples explain why gravitational acceleration does not lead to a suitable vertical

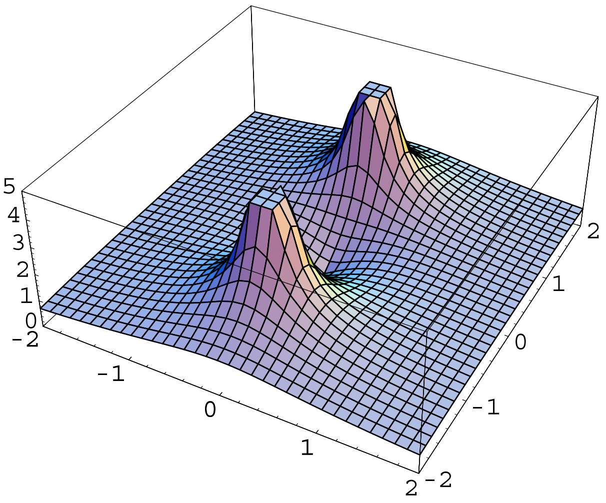

Figure 2.5: The magnitude of the force field created by two point masses.

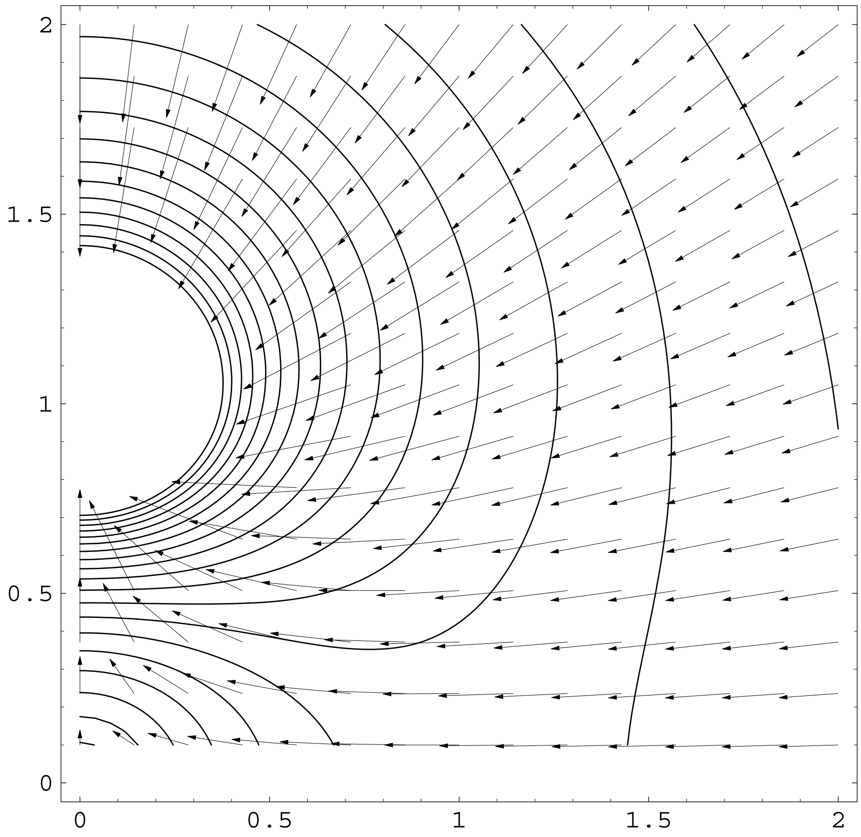

Figure 2.6: The force field vectors shown with the isoforce lines of the field. Note that the vectors are not perpendicular to the isolines thus illustrating that equiforce surfaces are not level.

datum, but they also provide a hint where to look. We require that water not flow between two points of equal height. We know from the first example that level surfaces have gravity force vectors that are normal to them. The second example illustrated that the key to finding a level surface pertains to *energy* rather than force, because the level surface in Figure 2.3 was created by equalizing the weight of the water columns. This is related to potential energy, which we will now discuss.

2.4 Work and Gravitational Potential Energy

Work plays a direct role in the definition of the geoid because it causes a change in the potential energy state of an object. In particular, when work is applied against the force of gravity causing an object to move against the force of gravity, that object's potential energy is increased, and this is an important concept in understanding the geoid. Therefore, we now consider the physics of work.

Work is what happens when a force is applied to an object causing it to move. It is a scalar quantity with units of distance squared times mass per second squared, and it is called the "joule," abbreviated J, in the mks system. Work is computed as force multiplied by distance, but only the force that is applied in the direction of motion contributes to the work done on the object.

Suppose we move an object in a straight line. If we denote a constant force by $\mathbf{F}$ and the displacement of the object by a vector $\mathbf{s}$, then the work done on the object is $W = \mathbf{F} \cdot \mathbf{s}$ (2.1). This same expression would be correct even if $\mathbf{F}$ is not directed exactly along the path of motion, because the inner product extracts from $\mathbf{F}$ only that portion that is directed parallel to $\mathbf{s}$. Of course, in general, force can vary with position, and the path of motion might not be a straight line. Let C denote a curve that has been parameterized by arc length s, meaning that $\mathbf{p} = C(s)$ is a point on C that is s units from C's starting point. Let $\hat{\mathbf{t}}(s)$ denote a unit vector tangent to C at s. Since we want to allow force to vary along C, we adopt a notion that the force is a function of position $\mathbf{F}(s)$. Then, by application of the calculus, the work expended by the application of a possibly varying force along a possibly curving path C from $s = s_0$ to $s = s_1$ is:

$$W = \int_{s_0}^{s_1} \mathbf{F}(s) \cdot \hat{\mathbf{t}}(s) ds. \tag{2.5}$$

Equation 2.5 is general so we will use it as we turn our attention to motion within a gravitational force field. Suppose we were to move some object in the presence of a gravitational force field. What would be the effect? Let us first suppose that we move the object on a level surface, which implies that the direction of the gravitational force vector is everywhere normal to that surface and, thus, perpendicular to $\hat{\mathbf{t}}(s)$, as well. Since by assumption $\mathbf{F}_g$ is perpendicular to $\hat{\mathbf{t}}(s)$, $\mathbf{F}_g$ plays no part in the work being done because $\mathbf{F}_g(s) \cdot \hat{\mathbf{t}}(s) = 0$. Therefore, moving an object over a level surface in a gravity field is identical to moving it in the absence of the field altogether, as far as the work done against gravity is concerned.

Now, suppose that we move the object along a path such that the gravitational force is not everywhere normal to the direction of motion. From Equation 2.5 it is evident that either more or less work will be needed due to the force of gravity, depending on whether the motion is against or with gravity, respectively. The gravity force will simply be accounted for by adding it to force we apply; the object can make no distinction between them. Indeed, we can use superposition to separate the work done in the same direction as gravity from the work done to move laterally through the gravity field; they are orthogonal. We now state, without proof, a critical result from vector calculus: the work done by gravity on a moving body does not depend on the path of motion, apart from the starting and ending points. This is a consequence of gravity being a conservative field (Blakely 1995, Schey 1992). As a result, the work integral along the curve defining the path of motion can be simplified to consider work only in the direction of gravity. This path is called a **plumb line** and, over short distances, can be considered to be a straight line, although the force field lines shown in Figure 2.6 show that plumb lines are not straight, in general. Therefore, from Equation 2.5, the work needed to, say, move some object vertically through a gravity field is given by:

$$W = \int_{h_0}^{h_1} \mathbf{F}_g(h) \cdot \hat{\mathbf{t}}(h) dh, \tag{2.6}$$

where
h = height (distance along the plumbline); and
$\hat{\mathbf{t}}(h)$ = the direction of gravity.
However, $\mathbf{F}_g(h)$ is always parallel to $\hat{\mathbf{t}}(h)$, so $\mathbf{F}_g(h) \cdot \hat{\mathbf{t}}(h) = \pm F_g(h)$, depending on whether the motion is with or against gravity. If we assume $\mathbf{F}_g(h)$ is constant, Equation 2.6 can be simplified as:

$$\begin{aligned} W &= \int_{h_0}^{h_1} \mathbf{F}_g(h) \cdot \hat{\mathbf{t}}(h) dh, && \text{(Eq.2.6)} \\ &= \int_{h_0}^{h_1} m\mathbf{E}_g(h) \cdot \hat{\mathbf{t}}(h) dh, && \text{(Eq.2.3)} \\ &= m\mathbf{E}_g(h) \int_{h_0}^{h_1} dh, && \text{(assuming } E_g \text{ is constant)} \\ &= mg\Delta h, && (2.7) \end{aligned}$$

where we denote the assumed constant magnitude of gravitational acceleration at the Earth's surface by g, as is customary. The quantity mgh is called **potential energy**, so Equation 2.7 indicates that the release of potential energy will do work if the object moves along gravity force lines. The linear dependence of Equation 2.7 on height (h) is a key concept.

2.5 The Geoid

2.5.1 What is the Geoid?

Although Equation 2.7 indicates a fundamental relationship between work and potential energy, we do not use this relationship directly because it is not convenient to measure work to find potential. Therefore, we rely on a direct relationship between the Earth's potential field and its gravity field that we state without justification:

$$\mathbf{E}_g = \triangledown U, \tag{2.8}$$

where
U = the Earth's potential field; and
$\triangledown$ = the gradient operator. [3] Written out in Cartesian coordinates, Equation 2.8 becomes:

$$\mathbf{E}_g = \frac{\partial U}{\partial x}\hat{\imath} + \frac{\partial U}{\partial y}\hat{\jmath} + \frac{\partial U}{\partial z}\hat{k}$$

where $\hat{\imath}, \hat{\jmath}, \hat{k}$ are unit vectors in the x, y, and z directions, respectively. In spherical coordinates, Equation (II.8) becomes:

$$\mathbf{E}_g = \frac{\partial U}{\partial r}\hat{\mathbf{r}}. \tag{2.9}$$

[3] Other authors write Equation 2.8 as $\mathbf{E}_g = -\triangledown U$, but the choice of the negative sign is essentially one of perspective: if the negative sign is included, the equation describes work done to overcome gravity. We prefer the opposite perspective because Equation 2.8 follows directly from Equation 2.3, in which the negative sign is necessary to capture the attractive nature of gravitational force.

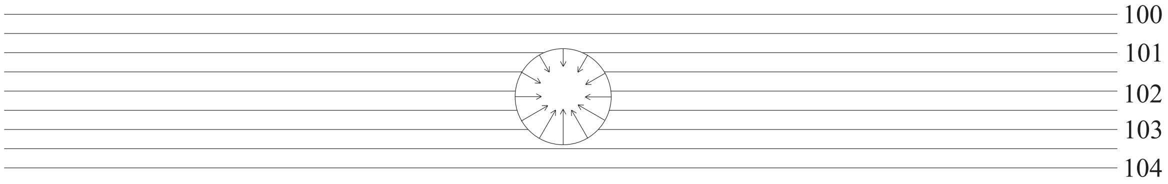

Figure 2.7: The force experienced by a bubble due to water pressure. Horizontal lines indicate surfaces of constant pressure, with sample values indicated on the side.

Equation 2.8 means that the gravity field is the gradient of the potential field. For full details, the reader is referred to the standard literature, including (Blakely 1995, Heiskanen & Moritz 1967, Ramsey 1981, Torge 1997, Vaníček & Krakiwsky 1986). Although Equation 2.8 can be proven easily (Heiskanen & Moritz 1967, p.2), the intuition behind the equation does not seem to be so easy to grasp.

We will attempt to clarify the situation by asking the reader to consider the following, odd, question: why do air bubbles go upwards towards the surface of the water? The answer that is usually given is because air is lighter than water. This is surely so but $\mathbf{F} = m\mathbf{a}$, so if bubbles are moving, then there must be a force involved. Consider Figure 2.7, which shows a bubble, represented by a circle, which is immersed in a water column. The horizontal lines indicate water pressure. The pressure exerted by a column of water increases nearly linearly with depth (because water is nearly incompressible). The water exerts a force inwards on the bubble from all directions, which are depicted by the force vectors. If the forces were balanced, no motion would occur. It would be like a rope in a tug-of-war in which both teams are equally matched. Both teams are pulling the rope but the rope is not moving: equal and opposite forces cause no motion.

However, the bubble has some finite height: the depth of the top of the bubble is less than the depth of the bottom of the bubble. Therefore, the pressure at the top of the bubble is less than the pressure at the bottom, so the force on the top of the bubble is less than that at the bottom. This pressure gradient creates an excess of force from below that drives the bubble upwards. Carrying the thought further, the difference in magnitude between any two lines of pressure is the gradient of the force field; it is the potential energy of the force field. The situation with gravity is exactly analogous to the situation with water pressure. Any surface below the water at which the pressure is constant might be called an “equipressure” surface. Any surface in or around the Earth upon which the gravity potential is constant is called an **equipotential** surface. Thus, a gravity field is caused by the difference in the gravity potential of two infinitely close gravity equipotential surfaces.

By assuming a spherical, homogeneous, non-rotating Earth, we can derive its potential field from Equation 2.9 and by denoting $|\mathbf{r}|$ by r:

$$\begin{aligned} \frac{\partial U}{\partial r}\hat{\mathbf{r}} &= \mathbf{E}_g \\ \int dU &= -\int \frac{GM}{r^2} dr \\ U &= \frac{GM}{r} + c. \end{aligned} \tag{2.10}$$

The constant of integration in Equation 2.10 can be chosen so that zero potential resides either infinity far away or at the center of M. We choose the former convention. Consequently, potential increases in the direction that gravity force vectors point and the absolute potential of an object

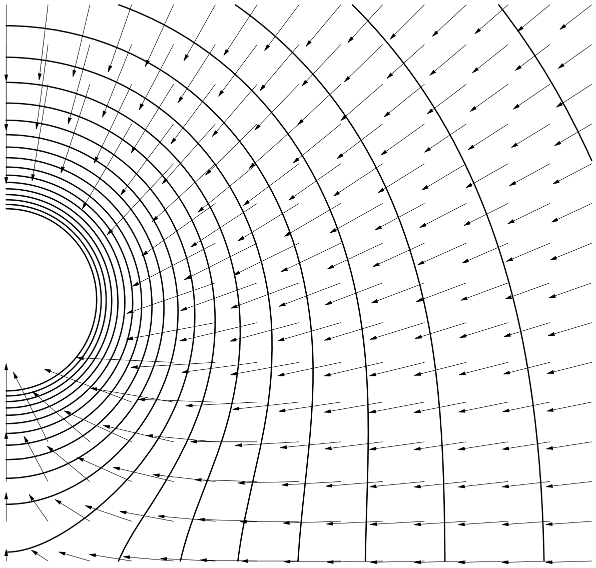

Figure 2.8: The gravity force vectors created by a unit mass and the corresponding isopotential field lines. Note that the vectors are perpendicular to the field lines. Thus, the field lines extended into three dimensions constitute level surfaces.

of mass m located a distance h from M is:

$$\begin{aligned} U &= -\int_{\infty}^{h} \frac{GMm}{r^2} dr \\ &= \frac{GMm}{r}\Big|_{\infty}^{h} \\ &= \frac{GMm}{h} - \frac{GMm}{\infty} \\ &= \frac{GMm}{h}. \end{aligned} \tag{2.11}$$

We now reconsider the definition of the geoid, being the equipotential surface of the Earth's gravity field that nominally defines mean sea level. From Equation 2.10, the geoid is some particular value of U and, furthermore, if the Earth were spherical, homogeneous, and not spinning, the geoid would also be located at some constant distance from the Earth's center of gravity. However, none of these assumptions are correct, so the geoid occurs at various distances from the Earth's center - it undulates.

One can prove mathematically that $\mathbf{E}_g$ is perpendicular to U. To illustrate this, see Figure 2.8. The figure shows the force vectors as seen in Figure 2.6 but superimposed over the potential field computed using Equation 2.10 instead of the magnitude of the force field. Notice that the vectors are perpendicular to the isopotential lines. Water would not flow along the isopotential lines; only across them. In three dimensions, the isopotential lines would be equipotential surfaces, such as the geoid.

2.5.2 The Shape of the Geoid

We now consider the shape of the geoid as it occurs for the real Earth. It is evident from Equation 2.10 that the equipotential surfaces of a spherical, homogeneous, non-rotating mass would be

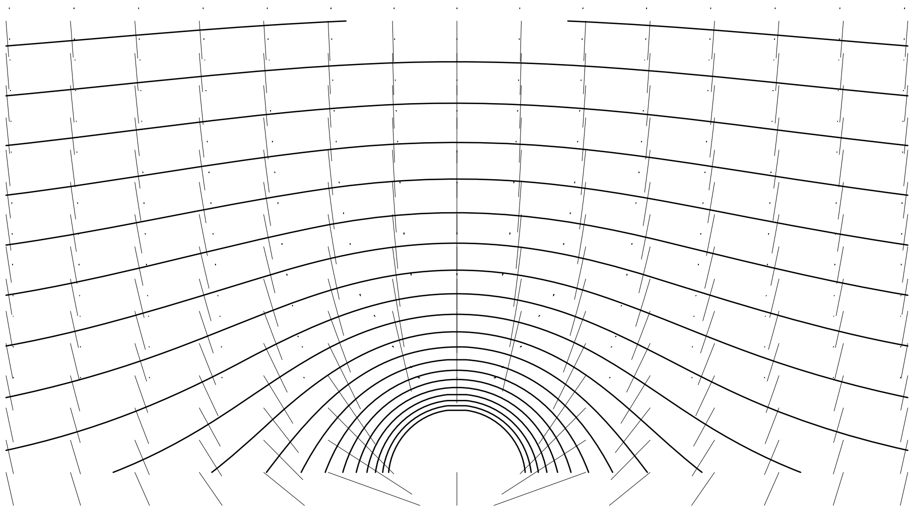

Figure 2.9: The gravity force vectors and isopotential lines created at the Earth's surface by a point with mass roughly equal to that of Mt. Everest. The single heavy line is a plumb line.

concentric, spherical shells-much like layers of an onion. If the sphere is very large, such as the size of the Earth, and we examined a relatively small region near the surface of the sphere, the equipotential surfaces would almost be parallel planes.

Now, suppose we add some mass to the sphere in the form of a point mass roughly equal to that of Mt. Everest positioned on the surface of the sphere. The resulting gravity force field and isopotential lines are shown in Figure 2.9. The angles and magnitudes are exaggerated for clarity; the deflection of the vertical is very apparent. In particular, we draw attention to the shape of the isopotential lines which run more-or-less horizontally across the figure. Notice how they bulge up over the mountain. This is true in general: the equipotential surfaces roughly follow the topographic shape of the Earth in that they bow up over mountains and dip down into valleys. Also, any one of the geopotential lines shown in Figure 2.9 can be thought of as representing the surface of the ocean above an underwater seamount. Water piles up over the top of subsurface topography to exactly the degree that the mass of the additional water exactly balances the excess of gravity caused by the seamount. Thus, one can indirectly observe seafloor topography by measuring the departure of the ocean's surface from nominal gravity (Hall 1992). The geoid, of course, surrounds the Earth, and Figure 1.2 on page 7 shows the ellipsoid height of the geoid with respect to NAD 83 over the conterminous United States as modeled by GEOID03 (Roman et al. 2004). At first glance, one could mistake the image for a topographic map. However, closer examination reveals numerous differences.

2.6 Geopotential Numbers

The geoid is usually considered the proper surface from which to reckon geodetic heights because it honors the flow of water and nominally resides at mean sea level. Sea level, itself, does not exactly match the geoid because of the various physical factors mentioned before. Therefore, actually finding the geoid in order to realize a usable vertical datum is currently not possible from mean sea level measurements. Ideally, one would measure potential directly in some fashion analogous to measuring gravity acceleration directly. If this were possible, the resulting number would be a **geopotential number**. In other words, a geopotential number is the potential of the Earth's gravity field at any point in space. Using geopotential numbers as heights is appealing for several reasons:

- Geopotential defines hydraulic head. Therefore, if two points are at the same geopotential number, water will not flow between them due to gravity alone. Conversely, if two points are not at the same geopotential number, gravity will cause the water to flow between them if the waterway is unobstructed (ignoring friction).
- Geopotential decreases linearly with distance from the center of the Earth (Equation (II.10)). This makes it a natural measure of distance.
- Geopotential does not depend on the path taken from the Earth's center to the point of interest. This makes a geopotential number stable.
- The magnitude of a geopotential number is less important than the relative values between two places. Therefore, one can scale geopotential numbers to any desirable values, such as defining the geoid to have a geopotential number of zero.

Equation 2.11 gives hope of determining height by measuring a gravity-related quantity, namely, absolute potential. Regrettably, potential cannot be measured directly. This is understandable because the manifestation of potential (the force of gravity) is created by potential differences, not in the potential itself. That is, two pairs of potential energies, say (150, 140) and (1000, 990) result in a force of the same magnitude. This is true because the difference of the two pairs is the same, namely, 10 newtons. In light of this, one might ask how images of the geoid, such as Figure 1.2 on page 7, came into being. The image in Figure 1.2 is the result of a sophisticated mathematical model based on Stokes' formula, which we take from Heiskanen & Moritz (1967, p.94) equation 2-163b, and present here for completeness:

$$N = \frac{R}{4\pi G} \int_{\sigma} \Delta g \; S(\psi) d\sigma, \tag{2.12}$$

where
N = geoid height at a point of interest;
R = mean radius of the Earth;
G = the universal gravitational constant;
Δg = the reduced, observed gravity measurements around the Earth;
ψ = the spherical distance from each surface element $d\sigma$ to the point of interest, and $S(\psi)$, which is known as Stokes' function, given by Heiskanen & Moritz (1967, p.94), equation 2-164:

$$S(\psi) = \frac{1}{\sin(\psi/2)} - 6\sin(\psi/2) + 1 - 5\cos\psi - 3\cos\psi \ln(\sin(\psi/2) + \sin^2(\psi/2))$$

The model is calibrated with, and has boundary conditions provided by, reduced gravity measurements taken in the field-the Δg's in Equation 2.12. These measurements together with Stokes' formula permit the deduction of the potential field that must have given rise to the observed gravity measurements.

In summary, in spite of their natural suitability, geopotential numbers are not practical to use as heights because practicing surveyors cannot easily measure them in the field.[4] They are, however, the essence of what the word height really means, and subsequent papers in this series will come to grips with how orthometric and ellipsoid heights are related to geopotential numbers by introducing **Helmert orthometric heights** and **dynamic heights**.

[4] Geopotential numbers have units of energy, not length. We suspect that most practicing surveyors would object to using heights that don't have length units, as well.

2.7 Summary

This second paper in a four-part series that reviews the fundamental concept of *height* presented simple derivations of the physics concepts needed to understand the force of gravity, since mean sea level and the Earth's gravity field are strongly interrelated. It was shown that one cannot use the magnitude of the force of gravity to define a vertical datum because equiforce surfaces are not level surfaces. However, it was observed that gravity potential gives rise to gravity force and, furthermore, gravity force is normal to equipotential surfaces. The practical consequence of this is that water will not flow along an equipotential surface due to the force of gravity alone. Therefore, equipotential surfaces are level surfaces and suitable to define a vertical datum. In particular, although there is an infinite number of equipotential surfaces, the geoid is often chosen to be the equipotential surface of the Earth's gravity field that best fits mean sea level in a least squares sense, and the geoid has thus become the fundamental vertical datum for mapping. It was shown that mean sea level itself is not a level surface, therefore, one cannot deduce the location of the geoid by measuring the location of mean sea level alone. Furthermore, one cannot measure gravity potential directly. Therefore, we model the geoid mathematically, based on gravity observations.

A geopotential number was defined to be a number proportional to the gravity potential at that place. Geopotential numbers capture the notion of height exactly because they vary linearly with vertical distance and define level surfaces. However, they are usually unsuitable for use as distances themselves because they cannot be measured directly and have units of energy rather than length.

Chapter 3

Height Systems

3.1 Preamble

This monograph was originally published as a series of four articles appearing in the Surveying and Land Information Science. Each chapter corresponds to one of the original papers. This paper should be cited as

Meyer, Thomas H., Roman, Daniel R., and Zilkoski, David B. (2005) What does *height* really mean? Part III: Height Systems. In *Surveying and Land Information Science*, 66(2): 149-160.

This is the third paper in a four-part series considering the fundamental question, "what does the word *height* really mean?" The first paper reviewed reference ellipsoids and mean sea level datums. The second paper reviewed the physics of heights culminating in a simple development of the geoid and explained why mean sea level stations are not all at the same orthometric height. This third paper develops the principle notions of height, namely measured differentially-deduced changes in elevation, orthometric heights, Helmert orthometric heights, normal orthometric heights, dynamic heights, and geopotential numbers. We conclude with a more in-depth discussion of current thoughts regarding the geoid.

3.2 Introduction

There are two general visions of what the word *height* means: a geometric separation versus hydraulic head. For Earth mensuration, these visions are not the same thing and this discrepancy has lead to many formulations of different types of heights. In broad strokes there are orthometric heights, purely geometric heights and heights that are neither. None of these are inferior to the others in all respects. They all have strengths and weaknesses, so to speak, and this has given rise to a number of competing height systems. We begin by introducing these types of heights, then examine the height systems in which they are measured and conclude with some remarks concerning the geoid.

3.3 Heights

3.3.1 Uncorrected Differential Leveling

Leveling is a process by which the geometric height difference along the vertical is transferred from a reference station to a forward station. Suppose a leveling line connects two stations A and B as depicted in Figure III.1 (c.f. Heiskanen & Moritz (1967, p. 161)). If the two stations are far enough apart, the leveling section will contain several turning points, the vertical geometric separation

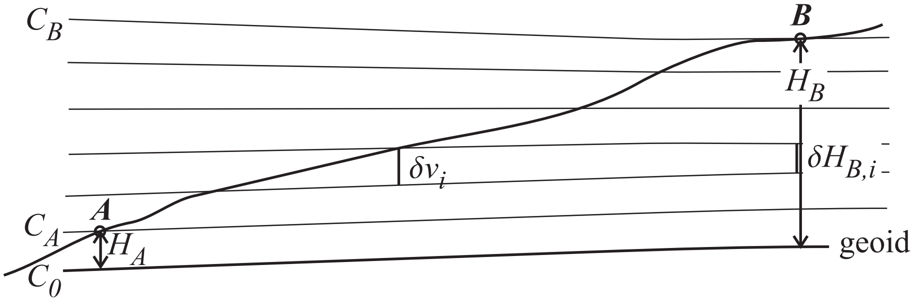

Figure 3.1: A comparison of differential leveling height differences δv_i with orthometric height differences $\delta H_{B,i}$. The height determined by leveling is the sum of the δv_i whereas the orthometric height is the sum of the $\delta H_{B,i}$. These two are not the same due to the non-parallelism of the equipotential surfaces whose geopotential numbers are denoted by C.

between which we denote as δv_i. Any two turning points are at two particular geopotential numbers, the difference of which is the potential gravity energy available to move water between them; hydraulic head. We also consider the vertical geometric separation of those two equipotential surfaces along the plumb line for B, $\delta H_{B,i}$. We will now argue that differential leveling does not, in general, produce orthometric heights. The figure depicts two stations A and B, indicated by open circles, with geopotential numbers C_A and C_B, and at orthometric heights H_A and H_B, respectively. The geopotential surfaces, shown in cross section as lines, are not parallel; they converge toward the right. Therefore, it follows that $\delta v_i \neq \delta H_{B,i}$. The height difference from A to B as determined by uncorrected differential leveling is the sum of the δv_i. Therefore, because $\delta v_i \neq \delta H_{B,i}$ and the orthometric height at B can be written as $H_B = \sum_i \delta H_{B,i}$, it follows that $\sum_i \delta v_i \neq H_B$. We now formalize the difference between differential leveling and orthometric heights so as to clarify the role of gravity in heighting. In the bubble "*gedanken* experiment" in the second paper of this series (Meyer, Roman & Zilkoski 2005*b*, pp. 11,12), we argued that the force moving the bubble was the result of a change in water pressure over a finite change in depth. By analogy we claimed that gravity force is the result of a change in gravity potential over a finite separation

$$g = -\delta W/\delta H \tag{3.1}$$

where g is gravity force, W is geopotential and H is orthometric height. Simple calculus allows rearranging to give $-\delta W = g\delta H$. Recall that δv_i and $\delta H_{B,i}$ are, by construction, across the same potential difference so $-\delta W = g\delta v_i = g'\delta H_{B,i}$, where g' is gravity force at the plumb line. Now, $\delta v_i \neq \delta H_{B,i}$ due to the non-parallelism of the equipotential surfaces but δW is the same for both, so gravity must be different on the surface where the leveling took place than at the plumb line. This leads us to Heiskanen & Moritz (1967, p.161, Eq. 4-2)

$$\delta H_{B,i} = \frac{g}{g'}\delta v_i \neq \delta v_i \tag{3.2}$$

which indicates that **differential leveling height differences differ from orthometric height differences by the amount that surface gravity differs from gravity along the plumb line at that geopotential**. An immediate consequence of this is that two different leveling lines starting and ending at the same station will, in general, provide different values for the height of final station. This is because the two lines will run through different topography and, consequently, geopotential surfaces with disparate separations. **Uncorrected differential leveling heights are not single-valued**, meaning the result you get depends on the route you took to get there.

In summary, heights derived from uncorrected differential leveling

- are readily observed by differential leveling,
- are not single-valued by failing to account for the variability in gravity,
- will not, in theory, produce closed leveling circuits, and
- do not define equipotential surfaces. Indeed, they do not define surfaces in the mathematical sense at all.

3.3.2 Orthometric Heights

According to Heiskanen & Moritz (1967, p.172), "Orthometric heights are the natural 'heights above sea level,' that is, heights above the geoid. They thus have an unequalled geometrical and physical significance." National Geodetic Survey (1986) defines **orthometric height** (ibid.) as, "The distance between the geoid and a point measured along the plumb line and taken positive upward from the geoid," with **plumb line** defined (ibid.) as, "A line perpendicular to all equipotential surfaces of the Earth's gravity field that intersect with it." In one sense, orthometric heights are purely geometric: they are the length of a particular curve (a plumb line). However, that curve depends on gravity in two ways. First, the curve begins at the geoid. Second, plumb lines remain everywhere perpendicular to equipotential surfaces through which they pass so the shape of the curve is determined by the orientation of the equipotential surfaces. Therefore, orthometric heights are closely related to gravity in addition to being a geometric quantity.

How are orthometric heights related to geopotential? Eq. 3.1 gives that $g = -\delta W/\delta H$. Taking differentials instead of finite differences and rearranging leads to $dW = -gdH$. Recall that geopotential numbers are the difference in potential between the geoid W_0 and a point of interest A, $W_A : C_A = W_0 - W_A$, so

$$\begin{aligned}
\int_{W_0}^{W_A} dW &= -\int_0^{H_A} g\,dH \\
W_A - W_0 &= -\int_0^{H_A} g\,dH \\
W_0 - W_A &= \int_0^{H_A} g\,dH \\
C_A &= \int_0^{H_A} g\,dH \qquad (3.3)
\end{aligned}$$

in which it is understood that g is not a constant. Eq. 3.3 can be used to derive the desired relationship

$$C_A = \bar{g}\ H_A \qquad (3.4)$$

meaning that a geopotential number is equal to an orthometric height multiplied by the average acceleration of gravity along the plumb line. It was argued in the second paper that geopotential is single-valued, meaning the potential of any particular place is independent of the path taken to arrive there. Consequently, orthometric heights are likewise single-valued, being a scaled value of a geopotential number.

If orthometric heights are single-valued, it is logical to inquire whether surfaces of constant orthometric height form equipotential surfaces. The answer to this is, unfortunately, no. Consider the geopotential numbers of two different places with the same orthometric height. If orthometric heights formed equipotential surfaces then two places at the same orthometric height must be at the same potential. Under this hypothesis Eq. 3.4 requires that the average gravity along the plumb lines of these different places *necessarily* be equal. However, the acceleration of gravity depends on height, latitude, and the distribution of masses near enough to be of concern; it is constant in neither magnitude nor direction. There is no reason that the average gravity would be equal and, in fact, it typically is not. Therefore, **two points of equal orthometric height need not have the same gravity potential energy, meaning that they need not be on the same equipotential surface and, therefore, not at the same height from the perspective of geopotential numbers**. Consider Figure III.2 a-d.

The figure, which is essentially a three-dimensional rendering of Figs. 2.9 and 3.1, shows an imaginary mountain together with various equipotential surfaces. Panel (b) shows the mountain with just one gravity equipotential surface. Everywhere on a gravity equipotential surface is at the same gravity potential, so water would not flow along the intersection of the equipotential surface with the topography without external influence. Nevertheless, the curve defined by the intersection of the gravity equipotential surface with the topography would *not* be drawn as a contour line on a topographic map because a contour line is defined to be, "An imaginary line on the ground, all points of which are at the same *elevation* above or below a specified reference surface" (National Geodetic Survey 1986). This runs contrary to conventional wisdom that would define a contour line as the intersection of a horizontal plane with the topography. In panels (c) and (d), one can see that the equipotential surfaces undulate. In particular, notice that the surfaces do not remain everywhere the same distance apart from each other and that they "pull up" through the mountains. Panel (d) shows multiple surfaces, each having less curvature than the one below it as a consequence of increasing distance from the Earth.

Now consider Figure 3.3, which is an enlargement of the foothill in the right side of panel 3.2(c). Suppose that the equipotential surface containing A and D is the geoid. Then the orthometric height of station B is the distance along its plumb line to the surface containing A and D; the same for station C. Although neither B nor C's plumb line is shown - both plumb lines are inside the mountain - one can see that the separation from B to the geoid is different than the separation from C to the geoid even though B and C are on the same equipotential surface. Therefore, they have the same geopotential number but have different orthometric heights. This illustrates why orthometric heights are single-valued but do not create equipotential surfaces.

How are orthometric heights measured? Suppose an observed sequence of geometric height differences δv_i has been summed together for the total change in geometric height along a section from station A to B, $\Delta v_{AB} = \sum_i \delta v_i$. Denote the change in orthometric height from A to B as ΔH_{AB}. Eq. 3.4 requires knowing a geopotential number and the average acceleration of gravity along the plumb line but neither of these are measurable. Fortunately, there is a relationship between leveling differences Δv and orthometric height differences ΔH: a change in orthometric height equals a change in geometric height plus a correction factor known as the **orthometric correction** (for a derivation see Heiskanen & Moritz (1967, pp.167-168, Eqs. 4-31 and 4-33)

$$\Delta H_{AB} = \Delta v_{AB} + OC_{AB} \tag{3.5}$$

where OC_{AB} is the orthometric correction and has the form

$$OC_{AB} = \sum_{A}^{B} \frac{g_i - \gamma_0}{\gamma_0} \delta v_i + \frac{\bar{g}_A - \gamma_0}{\gamma_0} H_A - \frac{\bar{g}_B - \gamma_0}{\gamma_0} H_B \tag{3.6}$$

where g_i is the observed force of gravity at the observation stations, $\bar{g}_A, \bar{g}_B$ are the average values of gravity along the plumb lines at A and B, respectively, and γ_0 is an arbitrary constant, which is often taken to be the value of normal gravity at 45° latitude. Although Eq. 3.6 stipulates gravity be observed at every measuring station, Bomford (1980, p.206) suggested that the observation stations need to be no closer than two to three km in level country but should be as close as 0.3 km in mountainous country. Others recommended observation station separations be 15 to 25 km in level

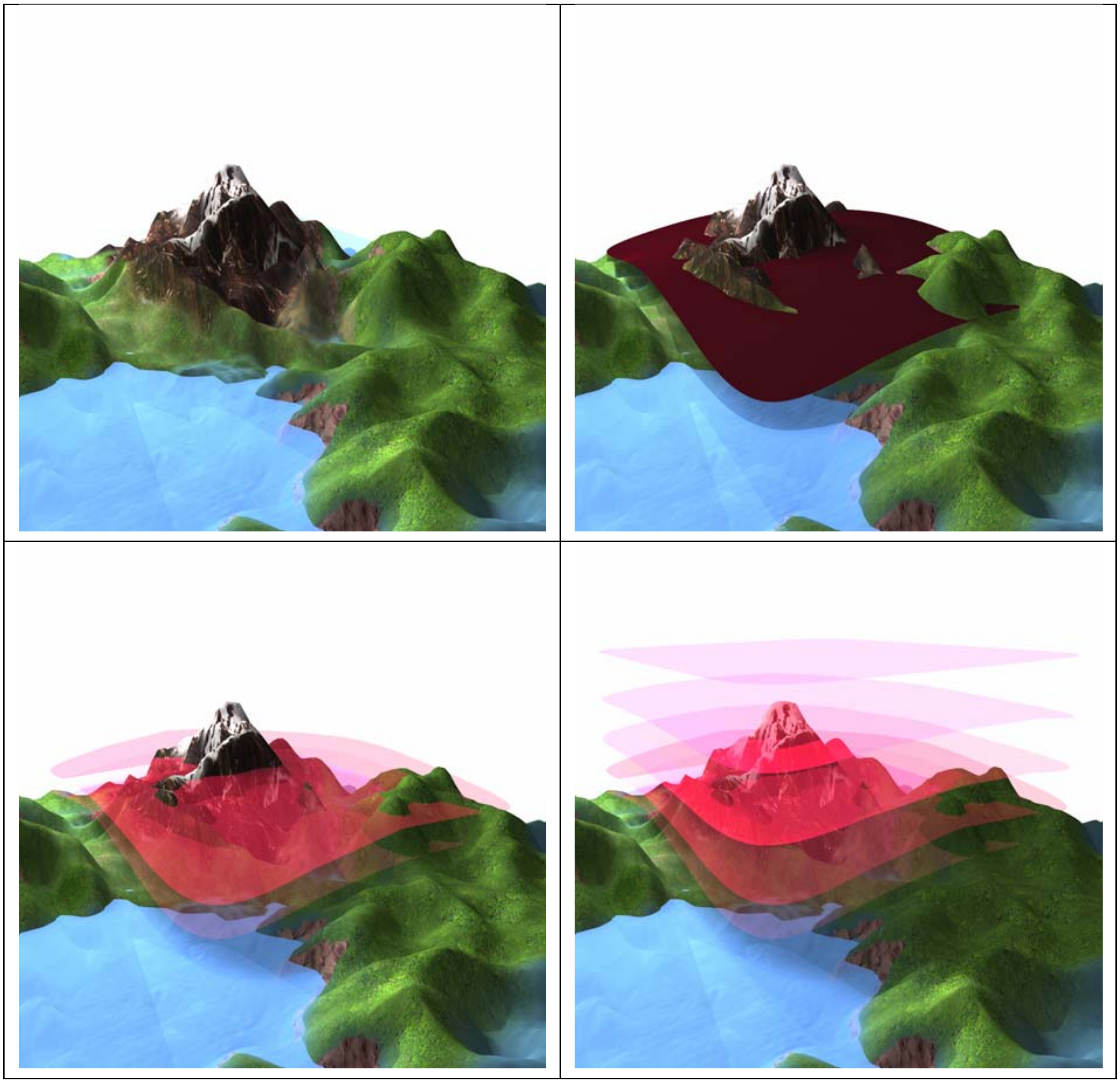

Figure 3.2: Four views of several geopotential surfaces around and through an imaginary mountain. (a) The mountain without any equipotential surfaces. (b) The mountain shown with just one equipotential surface for visual simplicity. The intersection of the surface and the ground is a line of constant gravity potential but *not* a contour line. (c) The mountain shown with two equipotential surfaces. Note that the surfaces are not parallel and that they undulate through the terrain. (d) The mountain shown with many equipotential surfaces. The further the surface is away from the Earth, the less curvature it has. (Image credit: Ivan Ortega, Office of Communication and Information Technology, UConn College of Agriculture and Natural Resources)

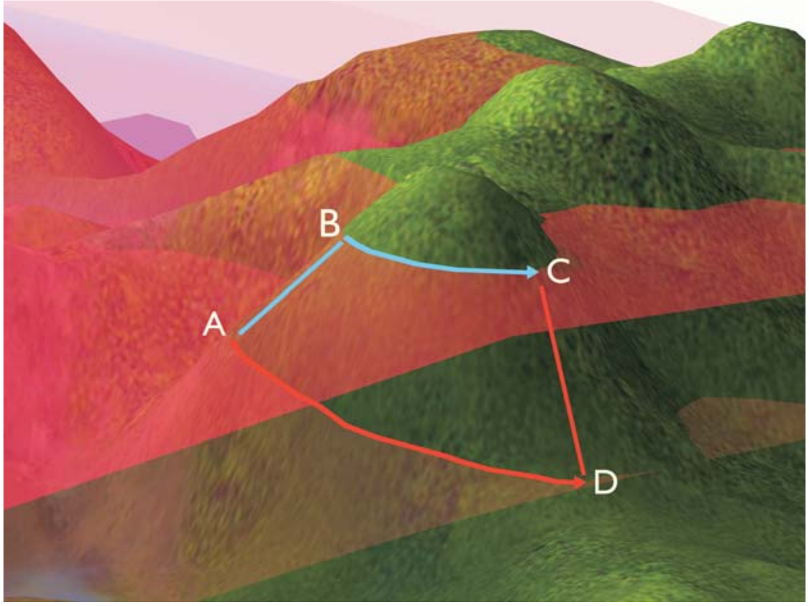

Figure 3.3: B and C are on the same equipotential surface but are at difference distances from the geoid at A-D. Therefore, they have different orthometric heights. Nonetheless, a closed leveling circuit with orthometric corrections around these points would theoretically close exactly on the starting height, although leveling alone would not.

country and 5 km in mountainous country (Strang van Hees 1992, Kao, Hsu & Ning 2000, Hwang & Hsiao 2003). There is a fair amount of literature on practical applications of orthometric corrections of which the following is a small sample: (Forsberg 1984, Strang van Hees 1992, Kao et al. 2000, Allister & Featherstone 2001, Hwang 2002, Brunner 2002, Hwang & Hsiao 2003, Tenzer, Vaníček, Santos, Featherstone & Kuhn 2005). The work described in these reports was undertaken by institutions with the resources to field gravimeters with their necessary surveying crews. Although there has been progress made in developing portable gravimeters (Faller & Vitouchkine 2003), it remains impractical to make the required gravity measurements called for by Eq. 3.6 for most surveyors. For first-order leveling, National Geodetic Survey (NGS) has used corrections that depend solely on the geodetic latitude and normal gravity at the observation stations thus avoiding the need to measure gravity (Survey 1981, pp. 5-26) although, if leveling is used to determine geopotential numbers such as in the NAVD 88 adjustment, orthometric corrections aren't used. NGS data sheets include modeled gravity at benchmarks, which provide a better estimate of gravity than normal gravity and are suitable for orthometric correction.

Although exact knowledge of $\bar{g}$ is not possible at this time, its value can be estimated either using a **free-air correction** (Heiskanen & Moritz 1967, pp. 163-164), or by the reduction of Poincaré and Prey (ibid., p 165). The former depends on knowledge of normal gravity only by making assumptions regarding the mean curvature of the potential field outside of the Earth. Orthometric heights that depend upon this strategy are called **Helmert orthometric heights**. NGS publishes NAVD 88 Helmert orthometric heights. The Poincar and Prey reduction, which requires a remove-reduce-restore operation, is more complicated and only improves the estimate slightly (*ibid.*, pp. 163-165).

In summary, orthometric heights

- constitute the embodiment of the concept of "height above sea level"

- are single-valued by virtue of their relationship with geopotential numbers and, consequently, will produce closed leveling circuits, in theory,
- do not define equipotential surfaces due to the variable nature of the force of gravity. This could, in principle, lead to the infamous situation of water apparently "flowing uphill." Although possible, this situation would require a steep gravity gradient in a location with relatively little topographic relief. This can occur in places where subterranean features substantially affect the local gravity field but have no expression on the Earth's surface, and
- are not directly measurable from their definition. Orthometric heights can be determined by observing differential leveling-derived geometric height differences to which are applied a small correction, the **orthometric correction**. The orthometric correction requires surface gravity observations and an approximation of the average acceleration of gravity along the plumb line.

3.3.3 Ellipsoid Heights and Geoid Heights

Ellipsoid heights are the straight line distances normal to a reference ellipsoid produced away from (or into) the ellipsoid to the point of interest. Before GPS it was practically impossible for anyone outside the geodesy community to determine an ellipsoid height. Now, GPS receivers produce three-dimensional baselines (Meyer 2002) resulting in determinations of geodetic latitude, longitude and ellipsoid height. Therefore, today, ellipsoid heights are commonplace.

Ellipsoid heights are almost never suitable surrogates for orthometric heights because equipotential ellipsoids (Meyer, Roman & Zilkoski 2005*a*, pp. 226,227) are not, in general, suitable surrogates for the geoid (although see Kumar (2005)). Consider that nowhere in the conterminous United States is the geoid closer to a GRS 80-shaped ellipsoid centered at the ITRF origin than about two meters. Confusing an ellipsoid height with an orthometric height could not result in a blunder less than two meters but would typically be far worse, even disastrous. For example, reporting the height of an obstruction in the approach to an airport runway at New York City using ellipsoid heights instead of orthometric heights would apparently lower the reported height by around 30 m with a possible result of causing a pilot to mistakenly believe the aircraft had 30 m more clearance than what is real.

Ellipsoid heights have no relationship to gravity, they are purely geometric. It is remarkable, then, that ellipsoid heights have a simple (approximate) relationship to orthometric heights, namely

$$H \approx h - N \tag{3.7}$$

where H is orthometric height, h is ellipsoid height and N is the ellipsoid height of the geoid itself, a **geoid height** or **geoid undulation**. This relationship is not exact because it ignores the deflection of the vertical. Nevertheless, it is close enough for most practical purposes. According to Eq. 3.7, ellipsoid heights can be used to determine orthometric heights if the geoid height is known. As discussed in the previous paper, geoid models are used to estimate N thus enabling the possibility of determining orthometric heights with GPS (Meyer et al. 2005*b*, p.12). We will explore these relationships in some detail in the last paper in the series on GPS heighting.

In summary, ellipsoid heights

- are single-valued (because a normal gravity potential field satisfies Laplace's equation and is, therefore, convex),
- do not use the geoid or any other physical gravity equipotential surface as their datum,
- do not define equipotential surfaces, and

- are readily determined using GPS.

3.3.4 Geopotential Numbers and Dynamic Heights

Geopotential numbers C are defined from Eq. 2.6, (c.f. (Heiskanen & Moritz 1967, p.162, Eq. 4-8)) which gives the change in gravity potential energy between a point on the geoid and another point of interest. The geopotential number for any place is the potential of the geoid W_0 minus the potential of that place W (recall the potential decreases with distance away from the Earth so this difference is a positive number). Geopotential numbers are given in geopotential units (g.p.u.), where 1 g.p.u. = 1 kgal-meter = 1000 gal meter (Heiskanen & Moritz 1967, p.162, Eq. 4-8). If gravity is assumed to be a constant 0.98 kgal, a geopotential number is approximately equal to 0.98 H, so geopotential numbers in g.p.u. are nearly equal to orthometric heights in meters. However, geopotential numbers have units of energy, not length, and are therefore an "unnatural" measure of height.

It is possible to scale geopotential numbers by dividing by a gravity value, which will change their units from kgal-meter to meter. Doing so results in a **dynamic height**:

$$H^{dyn} = C/\gamma_0 \tag{3.8}$$

One reasonable choice for γ_0 is the value of normal gravity (Eq. 1.2) at some latitude, conventionally taken to be 45 degrees. Obviously, scaling geopotential numbers by a constant does not change their fundamental properties so dynamic heights, like geopotential numbers, are single-valued, produce equipotential surfaces and form closed leveling circuits. They are not, however, geometric like an orthometric height: two different places on the same equipotential surface have the same dynamic height but generally do not have the same orthometric height. Thus, dynamics heights are not "distances from the geoid."

Measuring dynamic heights is accomplished in a manner similar to that for orthometric heights: geometric height differences observed by differential leveling are added to a correction term that accounts for gravity,

$$\Delta H^{dyn}_{AB} = \Delta v_{AB} + DC_{AB} \tag{3.9}$$

where Δv_{AB} is the total measured geometric height difference derived by differential leveling and DC_{AB} is the dynamic correction. The dynamic correction from station A to B is given by Heiskanen & Moritz (1967, p.163, Eq. 4-11) as

$$DC_{AB} = \sum_{A}^{B} \frac{g_i - \gamma_0}{\gamma_0} \delta v_i \tag{3.10}$$

where g_i is the (variable) force of gravity at each leveling observation station, $\gamma_0 = \gamma_0(45°)$, and the δv_i are the observed changes in geometric height along each section of the leveling line. However, DC typically takes a large value for inland leveling conducted far from the defining latitude. For example, suppose a surveyor in Albuquerque, New Mexico (at a latitude of around 35 N), begins a level line at the Route 66 bridge over the downtown railroad tracks at an elevation of, say, 1510 m and runs levels to the Four Hills subdivision at an elevation of, say, 1720 m, a change in elevation of 210 m. From Eq. 3.10, $DC = \Delta v(g - \gamma_0)/\gamma_0$. So taking $\gamma_0 = \gamma_{45°} = 980.62$ gal and $\gamma_{35°} = 979.734$ gal, then $DC = 210$ m (979.734 gal - 980.62 gal)/980.62 gal = -0.189775 m, a correction of roughly two parts in one thousand. This is a huge correction compared to any other correction applied in first-order leveling with no obvious physical interpretation such as the refraction caused by the atmosphere. It's unlikely that surveyors would embrace a height system that imposed such large corrections that would often affect even lower accuracy work. Nonetheless, dynamics heights are of practical use wherever water levels are needed, such as the Great Lakes and also along ocean shores even if they are used far from the latitude of the normal gravity constant. The geoid is thought to be not more than a couple meters from the ocean surface and, therefore, shores will have geopotential near to that of the geoid. Consequently, shores have dynamic heights near to zero regardless of their distance from the defining latitude. Even so, for inland surveying, DC can have a large value, on the order of several meters at the equator.

The dynamics heights in the International Great Lakes Datum of 1985 are established by the "Vertical Control - Water Levels" Subcommittee under the Coordinating Committee on Great Lakes Basic Hydraulics and Hydrology Data (CCGLBHHD). In summary, dynamic heights

- are a scaling of geopotential numbers by a constant to endow them with units of length,
- are not geometric distances,
- are single-valued by virtue of their relationship with geopotential numbers and, consequently, will produce closed-circuits, in theory,
- define equipotential surfaces, and
- are not measurable directly from their definition. Dynamic heights can be determined by observing differential leveling-derived geometric height differences to which are applied a correction, the **dynamic correction**. The dynamic correction requires surface gravity observations and can be on the order of meters in places far from the latitude at which γ_0 was defined.

3.3.5 Normal Heights

Of heights defined by geopotential (orthometric and dynamic) Heiskanen & Moritz (1967, p.287) write,

> The advantage of this approach is that the geoid is a level surface, capable of simple definition in terms of the physically meaningful and geodetically important potential W. The geoid represents the most obvious mathematical formulation of a horizontal surface at mean sea level. This is why the use of the geoid simplifies geodetic problems and makes them accessible to geometrical intuition.
>
> The disadvantage is that the potential W inside the earth, and hence the geoid W = const., depends on [a detailed knowledge of the density of the earth]Therefore, in order to determine or to use the geoid, the density of the masses at every point between the geoid and the ground must be known, at least theoretically. This is clearly impossible, and therefore some assumptions concerning the density must be made, which is unsatisfactory theoretically, even though the practical influence of these assumptions is usually very small.

These issues lead Molodensky in 1945 to formulate a new type of height, a normal height, which supposed that the Earth's gravity field was normal, meaning the actual gravity potential equals normal gravity potential (Molodensky 1945). The result of this postulate allowed that the "physical surface of the earth can be determined from geodetic measurements alone, without using the density of the earth's crust" Heiskanen & Moritz (1967, p.288). This conceptualization of heights allowed a fully rigorous method to be formulated for their determination, a method without assumptions. The price, however, was that "This requires that the concept of the geoid be abandoned. The mathematical formulation becomes more abstract and more difficult" (*ibid.*). Normal heights are defined by

$$C = \int_0^{H^*} \gamma \, dH^* \tag{3.11}$$

and

$$C = \bar{\gamma} \, H^* \tag{3.12}$$

where H^* is normal height and γ is normal gravity. These formulae have identical forms to those for orthometric height (c.f. Eqs. 3.3 and 3.4) but their *meaning* is completely different. First, the zero used as the lower integral bound is not the geoid; it is a reference ellipsoid. Consequently, normal heights depend upon the choice of reference ellipsoid and datum. Second, normal gravity is an analytical function so its average may be computed in closed form; no gravity observations are required. Third, from its definition one finds that a normal height H^* is that ellipsoid height where the normal gravity potential equals the actual geopotential of the point of interest. Regarding this, Heiskanen & Moritz (1967, p.170) commented, "...but since the potential of the earth is evidently not normal, what does all this mean?"

Like orthometric and dynamic heights, normal heights can be determined from geometrical height differences observed by differential leveling and applying a correction. The correction term has the same structure as that for orthometric correction, being

$$NC_{AB} = \sum_A^B \frac{g_i - \gamma_0}{\gamma_0} \delta v_i + \frac{\bar{\gamma}_A - \gamma_0}{\gamma_0} H_A^* - \frac{\bar{\gamma}_B - \gamma_0}{\gamma_0} H_B^* \tag{3.13}$$

with $\bar{\gamma}_A$ and $\bar{\gamma}_B$ being the average normal gravity at A and B, respectively, and other terms defined as Eq. 3.6. Normal corrections also depend upon gravity observations g_i but do not require assumptions regarding average gravity within the Earth. Therefore, they are rigorous; all the necessarily quantities can be calculated or directly observed. Like orthometric heights, they do not form equipotential surfaces (because of normal gravity's dependence on latitude; recall that dynamic heights scale geopotential simply by a constant whereas orthometric and normal heights' scale factors vary with location). Like orthometric heights, normal heights are single valued and give rise to closed leveling circuits. Geometrically, they represent the distance from the ellipsoid up to a surface known as the telluroid (see Heiskanen & Moritz (1967) for further discussion.)

In summary, normal heights

- are geometric distances, being ellipsoid heights, but not to the point of interest,
- are single-valued and, consequently, produce closed-circuits, in theory,
- do not define equipotential surfaces, and
- are not measurable directly from their definition. Normal heights can be determined by observing differential leveling-derived geometric height differences to which are applied a correction, the **normal correction**. The normal correction requires surface gravity observations only and, therefore, can be determined without approximations.

3.4 Height Systems

The term "height system" refers to a mechanism by which height values can be assigned to places of interest. In consideration of what criteria a height system must satisfy, Hipkin (2002) suggested

Table 3.1: A comparison of height systems with respect to various properties that distinguish them.

	Single-valued	Defines level surfaces	No misclosure	Small correction	Physically meaningful	Rigorous implementation
Uncorrected Dif. Leveling	no	no	no	n/a	yes	yes
Helmert Orthometric	yes	no	yes	yes	yes	no
Ellipsoidal	yes	no	yes	n/a	yes	yes
Dynamic	yes	yes	yes	no	yes	yes
Normal	yes	no	yes	yes	no	yes

two necessary conditions:
(i.Hipkin) Height must be single-valued.
(ii.Hipkin) A surface of constant height must also be a level (equipotential) surface.
Heiskanen & Moritz (1967, p.173) held two different criteria, namely
(i.H&M) Misclosures be eliminated.
(ii.H&M) Corrections to the measured heights be as small as possible.
The first two criteria (i.Hipkin and i.H&M) are equivalent: if heights are single-valued, then leveling circuits will be closed and vice versa. The second criteria form the basis of two different philosophies about what is considered important for heights. Requiring that a surface of constant height be equipotential requires that the heights be a scaled geopotential number and excludes orthometric and normal heights. Conversely, requiring the measurement corrections to be as small as possible precludes the former, at least from a global point of view, because dynamic height scale factors are large far from the latitude of definition. No height meets all these criteria. This has given rise to the use of (Helmert) orthometric heights in the United States, dynamic heights in Canada and normal heights in Europe (Ihde & Augath 2000). Table 3.1 provides a comparison of these height systems.

3.4.1 NAVD 88 and IGLD 85

Neither NAVD 88 nor IGLD 85 attempts to define the geoid or to realize some level surface which was thought to be the geoid. Instead, they are based upon a level surface that exists near the geoid but at some small, unknown distance from it. This level surface is situated such that shore locations with a height of zero in this reference frame will generally be near the surface of the ocean. IGLD 85 had a design goal that its heights be referenced to the water level gauge at the mouth of the St. Lawrence River. NAVD 88 had a design goal that it minimize recompilation of the USGS topographic map series, which was referred to NGVD 29. The station at Father Point/Rimouski met both requirements. NAVD 88 was realized using Helmert orthometric heights whereas IGLD 85 employs dynamic heights. Quoting from IGLD85 (1995),

> Two systems, orthometric and dynamic heights, are relevant to the establishment of IGLD (1985) and NAVD (1988). The geopotential numbers for individual bench marks are the same in both height systems. The requirement in the Great Lakes basin to provide an accurate measurement of potential hydraulic head is the primary reason for adopting dynamic heights. It should be noted that dynamic heights are basically geopotential numbers scaled by a constant of 980.6199 gals, normal gravity at sea level at 45 degrees latitude. Therefore, dynamic heights are also an estimate of the hydraulic

head.

Also, "IGLD 85 and NAVD 88 are now one and the same... The only difference between IGLD 85 and NAVD 88 is that IGLD 85 benchmark values are given in dynamic height units, and NAVD 88 values are given in Helmert orthometric height units. The geopotential numbers of benchmarks are the same in both systems". The United States covers a large area north-to-south within which is a considerable variety of topographic features. Therefore, dynamic heights would not be entirely acceptable for the U.S. because the dynamic corrections in the interior of the country would often be unacceptably large. The U.S. is committed now and for the future to orthometric heights, which in turn implies a commitment to geoid determination.

3.5 Geoid Issues

The geoid is widely accepted as the proper datum for a vertical reference system, although this perspective has challengers (Hipkin 2002). Conceptually, the geoid is the natural choice for a vertical reference system and, until recently, its surrogate, mean sea level, was the object from which the geoid was realized. However, no modern vertical reference system, in fact, uses the geoid as its datum primarily because the geoid is difficult to realize (although Canada has recently proposed re-defining their vertical datum using GPS and a geoid model). An exact, globally-satisfactory definition of the geoid is not straightforward. Both of these issues will be explored in turn.

The reasons that the geoid is not realizable from a mean sea level surrogate were given in the second paper in the discussion regarding why the mean sea surface is not a level surface. Quoting Hipkin (2002, p.376), the "... nineteenth century approach to establishing a global vertical datum supposed that mean sea level could bridge regions not connectable by leveling. The 'geoid' was formalized into the equipotential [surface] best fitting mean sea level and, for more than a century, the concepts of mean sea level, the geoid, and the leveling datum were used synonymously." We now know this use of "geoid" for "mean sea level" and vice versa to be incorrect because the mean sea surface is not an equipotential surface. Therefore, the mean sea surface is questionable as a vertical datum. Furthermore, Hipkin argues that measuring changing sea levels is one of the most important contributions that geodesy is making today. For this particular application, it does not make sense to continually adjust the vertical datum to stay at mean sea level and, thus, eliminate the phenomena trying to be observed. In contrast, chart makers, surveyors and mappers who define flood planes and subsidence zones would probably require that the vertical datum reflect changes in sea level to ensure their products are up-to-date and not misleading. Although a valid scientific point, Hipkin's argument does not override the need for NGS to determine the geoid, or a level surface near the geoid, in order to provide a well-defined datum for orthometric heights.

The second issue asserts that it is not straightforward to produce a globally-acceptable definition of the geoid. If one searches for a physics-based definition of the geoid, one finds that, according to Smith (1998, p.17), "The Earth's gravity potential field contains infinitely many level surfaces... The geoid is one such surface with a particular potential value, W_0." W_0 is a fundamental geodetic parameter (Burša 1995, Groten 2004) and its value has been estimated by using sea surface topography models (also called dynamic ocean topography models) and spherical harmonic expansions of satellite altimetry data (e.g., (Burša 1969, Burša 1994, Nesvorny & Sima 1994, Burša, Radej, Sima, True & Vatrt 1997, Burša, Kouba, Kumar, Müller, Radej, True, V. & Vojtíšková 1999) as well as GPS + orthometric height observations (Grafarend & Ardalan 1997). More recently (summer 2005, January/February 2006), research conducted in a joint effort between NGS, the National Aeronautics and Space Administration Goddard Flight Center and Naval Research Laboratory personnel is attempting to model the geoid by coupling sea surface topography model results with airborne gravimetry and Light Detection And Ranging (LIDAR) measurements in a manner similar to the aforementioned, space-based altimetry efforts. If successful, this work will result in another solution to the ongoing problem of determining W_0 with particular focus on the coastal regions of the U.S., c.f. Smith & Roman (2001, p.472). NGS is also examining earth gravity models (EGM's) derived from the satellite-based Gravity Recovery and Climate Experiment (GRACE) (Tapley, Bettadpur, Watkins & Reigber 2004) and (soon) Gravity Field and Steady-State Ocean Circulation Explorer (GOCE) data (Rebhan, Aguirre & Johannessen 2000) to establish higher confidence in the long wavelengths in EGM's (i.e., macroscopic scale features in the geoid model). Aerogravity data are being collecting to try and bridge the gaps at the shorelines between terrestrial data and the deep ocean and altimeter-implied gravity anomalies. EGM's and aerogravity data are being used to cross-check each other and the existing terrestrial data.

Even so, there is no consensus as to which value for W_0 should be chosen. Smith (1998) suggested W_0 could be chosen at least two ways: pick a "reasonable" value or adopt a so-called "best fitting ellipsoid." Hipkin (2002) has argued for the first approach: "To me it seems inevitable that, in the near future, we shall adopt a vertical reference system based on adopting a gravity model and one that incorporates $W = W_0 \equiv U_0$ to define its datum," with the justification (*ibid.*) that, "Nowadays, when observations are much more precise, their differences [between mean sea surface heights at various measuring stations] are distinguishable and present practice leads to confusion. It is now essential that we no longer associate mean sea level with any aspect of defining the geoid." In fact, G99SSS and GEOID99 were computed by choosing to model a specific $W = W_0$ surface (Smith & Roman 2001). Defining $W_0 \equiv U_0$ is unnecessary because it is computable as the zero-order geoid undulation (Smith 2006, personal communication). Other researchers have explored the second alternative by using the altimetry and GPS+leveling methods mentioned above. However, different level surfaces fill the needs of different user groups better than others. Moreover, it is probably unsatisfactory to define a single potential value for all time because mean sea level is constantly changing due to, for example, the changing amount of water in the oceans, plate tectonics changing the shape and volume of the ocean basins and the continents, and "thermal expansion of the oceans changing ocean density resulting in changing sea levels with little corresponding displacement of the equipotential surface" (Hipkin 2002). The geoid is constantly evolving, which leads to the need for episodic datum releases, as is done in the U.S. with mean sea level. If a global vertical datum is defined, it will only be adopted if it meets the needs of those who use it. With the United States' commitment to orthometric heights comes a need to define the geoid into the foreseeable future.

3.6 Summary

Heights derived through differential spirit leveling, ellipsoid and geoid heights, orthometric heights, geopotential numbers, dynamic heights, and normal heights were defined and compared regarding their suitability as an engineering tool and to reflect hydraulic head. It was shown that differential leveling heights provide neither single-valued heights nor an equipotential surface, resulting in theoretical misclosures of leveling circuits. Orthometric heights are single-valued but do not define level surfaces and require an approximation in their determination. Geopotential numbers are single-valued and define level surfaces but do not have linear units. Dynamics heights are single-valued, define level surfaces, are not intrinsically geometric in spite of having linear units, and often have unacceptably large correction terms far away from the latitude at which they are normalized. Normal heights are geometric, single-valued, have global applicability and can be realized without assumptions but do not define level surfaces. There is, in fact, no single height system that is both geometric and honors level surfaces simultaneously because these two concepts are physically incompatible due to the non-parallelism of the equipotential surfaces of the Earth's gravity field. Two modern vertical datums in use in North America, NAVD 88 and IGLD 85, express heights as either Helmert orthometric heights or dynamic heights, respectively. It was shown that this difference is, in one sense, cosmetic because these heights amount to different scalings of the same geopotential numbers. Nevertheless, Helmert orthometric heights and dynamic heights are incommensurate. The fact that there are disparate height systems reflect the needs and, to some extent, the philosophies behind their creation. No one height system is clearly better than the others in all counts.

Different organizations and nations have chosen various potentials to be their geoid for reasons that suit their purposes best. Others have argued that the gravity potential value $W = W_0 = U_0$ could be adopted to be the geoid's potential, which is attractive for some scientific purposes, though the U_0 of GRS 80 is no better or worse choice than any other U_0. However, the United States is committed to orthometric heights and NGS is actively engaged in measurements to locate the geoid based on LIDAR observations, gravimetric geoid models and sea surface topography models.

Chapter 4

GPS Heighting

4.1 Preamble

This monograph was originally published as a series of four articles appearing in the Surveying and Land Information Science. Each chapter corresponds to one of the original papers. This paper should be cited as

Meyer, Thomas H., Roman, Daniel R., and Zilkoski, David B. (2005) What does *height* really mean? Part IV: GPS Heighting. In *Surveying and Land Information Science*, 66(3): 165-183.

This is the final paper in a four-part series examining the fundamental question, "What does the word *height* really mean?" The creation of this series was motivated by the National Geodetic Survey's (NGS) embarking on a height modernization program as a result of which NGS will publish measured ellipsoid heights and computed Helmert orthometric heights for vertical bench marks. Practicing surveyors will therefore encounter Helmert orthometric heights computed from Global Positioning System (GPS) ellipsoid heights and geoid heights determined from geoid models as their published vertical control coordinate, rather than adjusted orthometric heights determined by spirit leveling. It is our goal to explain the meanings of these terms in hopes of eliminating confusion and preventing mistakes that may arise over this change. The first paper in the series reviewed reference ellipsoids and mean sea level datums. The second paper reviewed the physics of heights culminating in a simple development of the geoid in order to explain why mean sea level stations are not all at the same orthometric height. The third paper introduced orthometric heights, geopotential numbers, dynamic heights, normal heights, and height systems. This fourth paper is composed of two sections. The first considers the stability of the geoid as a datum. The second is a review of current best practices for heights measured with the Global Positioning System (GPS), essentially taking the form of a commentary on NGS' guidelines for high-accuracy ellipsoid and orthometric height determination using GPS.

JULIET : And not impute this yielding to light love,
Which the dark night hath so discovered.
ROMEO: Lady, by yonder blessed moon I vow,
That tips with silver all these fruit-tree tops–
JULIET: O, swear not by the moon, th' inconstant moon,
That monthly changes in her circle orb,
Lest that thy love prove likewise variable.

– William Shakespeare: Romeo and Juliet-The Balcony Scene (Act 2, Scene 2)

4.2 Vertical Datum Stability

Stability is a desirable quality for a datum, meaning that a datum ought not to change with time-this is a concept well understood by surveyors. The purpose of this series of papers is to explore issues pertaining to determining orthometric heights with GPS technology at the accuracy on the order of centimeters; so if the datums to which the height systems are referred vary by this amount or more, then these effects must be taken into account and removed. Therefore, let us consider the geoid in this light: is the geoid stable or does it change with time and, if so, how quickly and by how much?

An investigation into the variability of the geoid is equivalent to an investigation into the variability of the Earth's gravity potential field; it is a subject in the field of geodynamics. Changes in Earth's gravity are caused by changes in (1) the Earth's diurnal rotation that produces the centrifugal force component of gravity; (2) the Earth's mass and its distribution; or (3) the spatial arrangement of objects massive enough and near enough that their gravitational fields have a discernible effect on the geoid.

4.2.1 Changes in the Earth's Rotation

The Earth's diurnal rotation is not constant in velocity or direction. It is known that the length of the day is decreasing by about two milliseconds per century and that there are seasonal variations (with periods on the order of a month) on the same order ((Vaníček & Krakiwsky 1986, p. 68). Consequently, the Earth's centrifugal force is likewise diminishing and variable. However, these variations are far too small (on the order of 10^{-12} radians s^{-1}) to change the Earth's centrifugal force at a discernible level in faster than a geologic time frame.

The rotational axis of the Earth slowly traces a circle on the celestial sphere, the same motion that can be observed in a spinning top. This motion is called {**precession**. The Earth's precession is caused by the equatorial bulges not aligning in the plane of the **ecliptic** (the plane in which the Earth orbits the sun), thereby giving rise to a torque from the gravitational attraction of the sun (Vaníček & Krakiwsky 1986, p.59). The Earth's precession is slow, with its axis returning to a previous orientation once in approximately 25765 years, a period known as a **Platonic year**. Likewise, the equatorial bulges are not aligned with the Moon's orbital plane, which is inclined 511' to the ecliptic. The intersection of the Moon's orbital plane with the ecliptic is known as the **nodal line**, and the nodal line rotates once in 18.6 years, the **Metonic cycle**. This constant realignment of the Moon with the Earth also affects the orientation of the Earth's rotational axis, causing a motion called **nutation** (Vaníček & Krakiwsky 1986) and (Volgyesi 2006, p.61).

There are additional, smaller perturbations as well. The motion of the Earth's rotational axis in the celestial reference frame affects astronomic and satellite observations but not gravity because, although the direction of the centrifugal force vector is changing, this change was brought about by a motion of the Earth itself, so the relative change is zero. However, actual movement of the rotational axis relative to the Earth's crust itself (known as "Polar Motion" or "Polar Wobble") does affect gravity, because the direction of the centrifugal force vector in this case is changing relative to the Earth's crust. These small changes are only on the order of a few nanoGals, well below the noise level of most gravity measurements.

4.2.2 Changes in the Earth's Mass

The Earth's mass can increase or decrease, and it can be redistributed. Concerning the former, the Earth does gain mass almost continuously due to a stream of space debris entering the atmosphere and, occasionally, striking the Earth's surface. Similarly, the Earth is constantly loosing mass as gaseous molecules too light to be bound by gravity drift off into space (e.g., helium gas). Neither the addition nor the removal of mass changes the Earth's gravity field enough to be of concern in this paper.

The Earth's mass is redistributed in various ways including post-glacial rebound, melting ice caps and glaciers, the Earth's fluid outer core, the oceans (Cazenave & Nerem 2002), and earthquakes. For example, earthquakes can be caused by the motion of tectonic plates along their margins, and this motion causes a change in the Earth's shape. Earthquakes can cause a measurable change in the Earth's rotation velocity, and thus its gravity, by changing one of its moments of inertia (Chao & Gross 1987, D.E. & Manshina 1971, Soldati & Spada 1999). The Sumatra, Indonesia, earthquake of December 26, 2004 was such an event. It decreased the length of day by 2.68 microseconds, shifted the "mean North pole" about 2.5 cm in the direction of 145 degrees East Longitude, and decreased the Earth's flattening by about one part in 10 billion (Buis 2005). The uplift of plates due to tectonic or post-glacial activities affects ellipsoidal heights, as well as having a smaller gravity-based effect which changes the geoid. The National Geodetic Survey is planning to engage in research which tracks the time-dependent changes of the geoid due to these effects.

4.2.3 Tides

People who have been at an ocean shore for half a day or more have had the opportunity to watch the ocean advance inland and then retreat back out to sea. This motion is caused primarily by the gravitational attraction of the Moon and, to a lesser degree, the Sun. Therefore, the definition of **tide** found in NGS' Geodetic Glossary may be somewhat surprising.

> **Tide** (1) Periodic changes in the shape of the Earth, other planets or their moons that relate to the positions of the Sun, Moon, and other members of the solar system.

Note that this definition is not about the oceans, *per se*. Instead, it speaks of, among other things, a change of the shape of the *Earth itself*, the **Earth tide** or **body tide**. It is commonplace knowledge that the Moon moves the oceans; it deforms them to set them in motion. But, what is probably not so well known is that the Earth's core, mantle, and crust have their shape deformed in a manner similar to the deformation of the oceans. The NGS definition continues:

> In particular, (2) those changes in the size and shape of a body that are caused by movement through the gravitational field of another body. The word is most frequently used to refer to changes in size and shape of the Earth in response to the gravitational attractions of the other members of the solar system, in particular, the Moon and sun. In such cases, three different tides are usually distinguished: the atmospheric tide, which acts on the gaseous envelope of the Earth; the earth tide, which acts on the solid Earth; and the ocean tide (usually simply called "the tide"), which acts on the hydrosphere.

The effects of the tides are numerous and complicated, so perhaps the first question to consider is whether the tides cause enough of an effect to be of concern. Is the earth tide large enough to affect the geoid in any practical way? It happens that there are two high and low earth tides each day, with the highest being on the order of a 50 cm displacement from its undeformed shape (Moritz 1980, p.477)! So, the answer is "yes;" we must take tides into consideration.

Tides on the Earth arise due to the influences from all celestial bodies. The Sun and the Moon produce the largest effects by far, but the other planets have a discernable affect, albeit too small to impact GPS positioning (Wilhelm & Wenzel 1997, p.11). All celestial bodies create tides in the same way, the only difference being the details of how these manifest themselves. Therefore, we will consider the effects created by the Moon, with the understanding that they apply to any celestial body with the appropriate change of mass and distance variables.

4.2.4 Tidal Gravitational Attraction and Potential

According to Newton, force gives rise to motion by accelerating mass. The gravitational force of the Moon on the Earth itself is found using Equation 2.2 on page 17:

$$\mathbf{F}_E = -\frac{GMm\hat{\mathbf{r}}}{|\mathbf{r}|^2}, \tag{4.1}$$

where:
$\mathbf{r}$= a vector from the Moon's center to the Earth's center (note the negative sign in Equation 4.1 reversing the direction of the vector so that the force is directed from the Earth's center towards the moon's center);
M, m = the mass of the Earth and the Moon, respectively; and
$\mathbf{F}_E$ = the gravitational force vector produced by the Moon on the Earth.

The gravitational force of the Earth exerted on the Moon can be found simply by defining $\mathbf{r}$ to have the opposite direction, so the magnitudes of the two forces are equal. The gravitational attraction of the Earth on the Moon causes the Moon to orbit the Earth rather than to move off into space. However, Equation 4.1 also means that the *Earth is orbiting the Moon*, but this motion is much less obvious due to the difference in masses of the two bodies. If we take 5.9742×10^{27} g to be the Earth's mass, 7.38×10^{25} g to be the Moon's mass, and 3.84×10^8 m to be their mean separation, then the barycenter of the Earth-Moon system can be found to be at a point on a line connecting their two centers approximately 4.69×10^6 m from the Earth's center. This point is inside the Earth, being about 73.5 percent of the length of the GRS 80 semimajor axis.

It is critical to understand the nature of the motion of the Earth's orbiting the Moon. The diurnal rotation of the Earth, the source of days and nights, is a rotation around its axis, which is nominally the North Pole. Points on a rigid rotating body that are on different radii move in different directions and at different instantaneous linear velocities (see Figure IV1a). However, a rigid body can rotate around only one axis at any moment in time. Therefore, the Earth does *not* rotate about the Earth-Moon barycenter. To understand this orbital motion, envision someone waxing a tabletop with a cloth by rubbing it in a circular motion, such that their fingers remain parallel to some wall in the room. If the circular motion of the cloth has a fairly small radius, then the point around which the cloth is moving is always beneath the cloth, just as the motion of the Earth around the barycenter has its center at a point within the Earth. Now, it is apparent that every point on the cloth is actually moving with the same velocity (same direction and speed). Similarly, the orbital motion of the Earth around the Moon gives rise to a *constant* acceleration that is always directed opposite to the line connecting the Earth's center to the Moon's. In particular, *everywhere and everything on and in the Earth is accelerating away from the Moon as if the Earth were moving in a straight line along the instantaneous axis between them*; see Figure IV.1b. This acceleration gives rise to a component of observable gravity that is at most 3.4 percent of the total acceleration (Vaníček & Krakiwsky 1986, p.125).

The moon's gravitational attraction gives rise to a force at any particular place on the Earth that is directed (approximately[1]) along the line from the point of interest to the Moon's center. In contrast, the orbital acceleration experienced at that place is always parallel to the line connecting the Earth-Moon centers, so these forces are not generally parallel to each other. Furthermore, places on the side of the Earth opposite the Moon experience a smaller attraction than places on the same side as the Moon due to being closer to the Moon, giving rise to the asymmetry evident

[1] The moon is too close to the Earth for this to be exact. The actual direction of the vector would be determined by triple integrating over the Moon's mass, and approximately end up pointing at the Moon's center of mass, approximate because the Moon is not a perfectly homogeneous sphere.

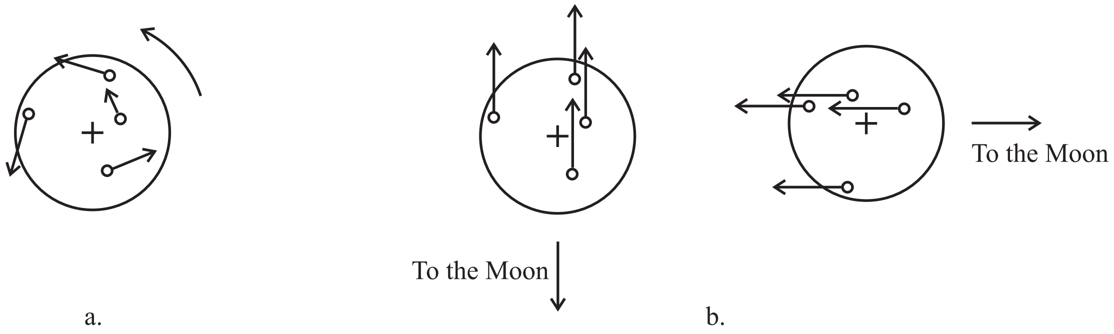

Figure 4.1: Panel (a) presents the instantaneous velocity vectors of four places on the Earth; the acceleration vectors (not shown) would be perpendicular to the velocity vectors directed radially toward the rotation axis. The magnitude and direction of these velocities are functions of the distances and directions to the rotation axis, shown as a plus sign. Panel (b) presents the acceleration vectors of the same places at two different times of the month, showing how the acceleration magnitude is constant and its direction is always away from the moon.

in Figure IV.2. Each of the vectors in Figure IV.2 indicates the force vector of the place located at the tail of the vector resulting from the combination of the orbital acceleration and the Moon's attraction at that place.

Figure IV.3 shows the details of the vector addition of three points of interest from Figure IV.2. Orange vectors are the Moon's attraction; their non-parallelism with the orbital acceleration vectors, shown in blue, is greatly exaggerated. The vector result of the addition of these two vectors is shown in black. Figure IV.3a represents the situation at point a, which is located at the top of the circle in Figure IV.2. The Moon's attraction is the most non-parallel with the orbital acceleration at this place and its antipodal counterpart. Given the roughly equal magnitude of the orbital acceleration and Moon attraction forces, their component in the direction of the Moon largely cancels at a, leaving a small result oriented sharply toward the Earth's middle. Figure IV.3b represents the situation at b which is located at the point furthest from the Moon. The Moon's attraction is parallel but opposite in direction with the orbital acceleration at this place. The orbital acceleration is moderately stronger than the Moon's attraction here, creating the force primarily responsible for the lower high tide of the day. Figure IV.3c represents the situation at c which is located at the point closest to the Moon. The Moon's attraction is considerably stronger here than the orbital acceleration, creating the force that is primarily responsible for the higher tide of the day (see Vaníček & Krakiwsky (1986, p.124) and Bearman (1999, pp. 52-61)).

The magnitude and direction of the Moon's attraction is periodic due to the nature of its orbit around the Earth. The situation is complicated but made tractable by accounting for individual **tidal constituents**. It is possible to decompose the Moon's attraction into individual constituents, a constituent being a sinusoid with a particular amplitude, frequency and phase that arises due to a particular phenomenon. As discussed by Boon (2004), some of the prominent tidal constituents are caused by

- the inclination of the Moon's orbital plane with respect to the ecliptic giving rise to the lunar declination (tropic-equatorial) cycle,
- the Sun's attraction giving rise to the spring-neap cycle,
- the eccentricity of the Moon's orbit giving rise to the perigean-apogean cycle, and
- the precession of the lunar nodes giving rise to the metonic cycle.

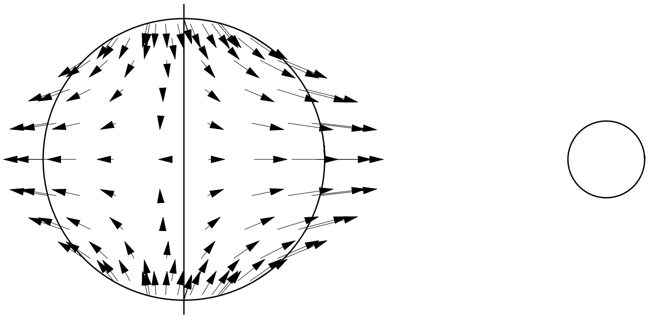

Figure 4.2: Arrows indicate force vectors that are the combination of the moon's attraction and the Earth's orbital acceleration around the Earth-moon barycenter. This force is identically zero at the Earth center of gravity. The two forces generally act in opposite directions. Points closer to the moon experience more of the moon's attraction whereas points furthest from the moon primarily experience less of the moon's attraction; c.f. (Bearman 1999, pp.54-56) and (Vanicek and Krakiwsky 1996, p.124).

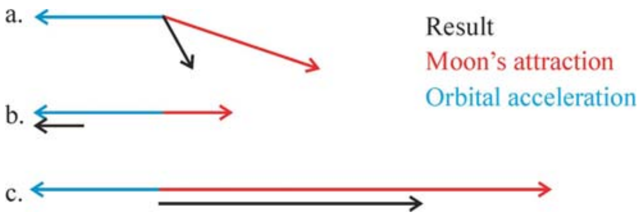

Figure 4.3: Details of the force combinations at three places of interest; c.f. Fig. IV.2.

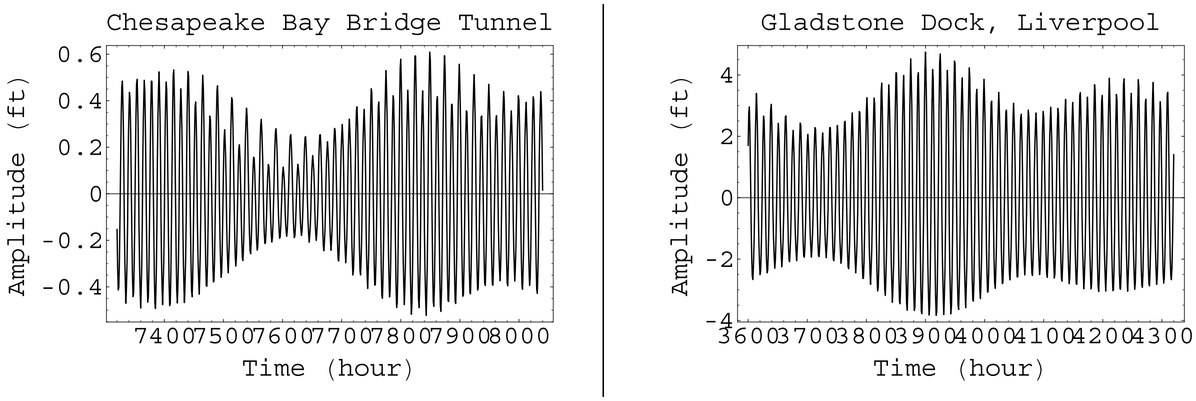

Figure 4.4: Two simulations of tide cycles illustrating the variety of possible affects.

Simple ocean tide models include as few as six constituents; complicated models can incorporate more than 100 (Wilhelm & Wenzel 1997). These models produce tidal predictions such as those shown in Figure IV.4. The predictions in Figure IV.4 use constituents from Boon (2004, pp.97-102) and clearly show higher high water, lower high water, higher lower water, and lower low water, as well as many longer-period variations.

Up to this point we have been concerned with gravity *force*. We now consider how tides affect gravity *potential* because, after all, the geoid (an equipotential surface) is a principle datum of interest, hence we must examine how these temporal changes come into play. The gravitational potential field created by the Moon at some point of interest can be expressed as an infinite series of which only the second term is important for tides. This second term W_2 takes the form of the expression (Vaníček 1980, p.5, Equation (12)):

$$W_2 \approx D \left[\overbrace{\cos^2\phi \cos^2\delta \cos 2t}^{\text{sectorial}} + \overbrace{\sin 2\phi \sin 2\delta \cos t}^{\text{tesseral}} + \overbrace{3(\sin^2\phi - 1/3)(\sin^2\delta - 1/3)}^{\text{zonal}} \right], \tag{4.2}$$

where:
D = Doodson's constant (Doodson 1922);
ϕ = geocentric latitude;
δ = the declination of the Moon; and
t = the Moon's hour angle (see any standard work on celestial mechanics for exact definitions of δ and t).

Doodson's constant is given by Vaníček (1980, p.4, Equation (7)) as:

$$D = \frac{3}{4} G m \frac{R^2}{r_m^3}, \tag{IV.3}$$

where:
G = the universal gravitation constant;
R = the mean (equivoluminous) radius of the Earth; and
r_m = the mean distance to the Moon.

D has a value of approximately 2.6277×10^7 cm mgal. Equation 4.2 consists of three terms within the brackets. The first term contains **sectorial** constituents; the second term contains

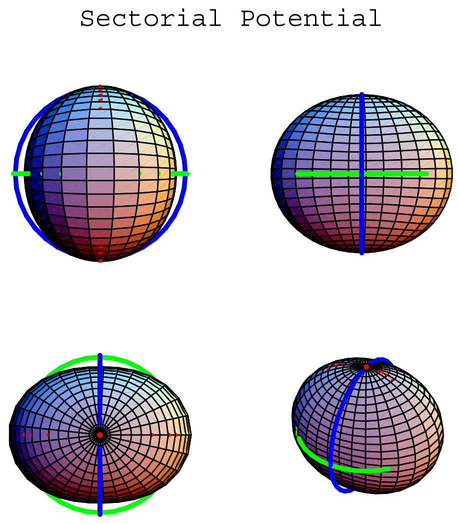

Figure 4.5: The sectorial constituent of tidal potential. The green line indicates the Equator. The red and blue lines indicate the Prime Meridian/International Date line and the 90/270 degree meridians at some arbitrary moment in time. In particular, these circles give the viewer a sense of where the potential is outside or inside the geoid. The oceans will try to conform to the shape of this potential field and, thus, the sectorial constituent gives rise to the two high/low tides each day.

tesseral constituents, and the third term contains **zonal** constituents. These three components are shown in Figures IV.5-7 and their combination in Figure IV.8. Sectorial constituents vary in longitude (time), much like the sectors of an orange, and give rise to the two daily tides. Tesseral constituents possess both latitude and longitude components and give rise to patterns resembling the tessellation of a checker board. The zonal constituents do *not* vary in time and give rise to so-called “permanent” tides.

4.2.5 Body Tides

The first clear evidence of body tides came from the measurement of ocean tides, which showed that they were consistently about two-thirds as high as Newton’s physics predicted. It was eventually shown that the missing one-third was due to deformation of the Earth itself, moving with the oceans (Melchior 1974). The tides of the solid Earth behave in the same manner as the ocean tides, but in a simpler manner because the Earth deforms like an elastic solid at the frequencies of tides, rather than with all the freedom of a liquid, like the oceans.

It is remarkable that the effect of the Moon’s potential field upon the Earth can be described with such high accuracy by such a simple equation as Equation 4.2; compare this with the effort necessary to determine the geoid! The simplicity of Equation (IV.2) is because (1) the Moon is far enough away to be treated as a point mass, and (2) the motion of the Moon is very accurately described by celestial mechanics. Therefore, no gravity observations are needed to determine the potential from the Moon; it all falls out of the mathematics.

The parameters that describe the response of the Earth’s shape and gravitational potential field to tidal forces are called **Love** and **Shida** numbers, which are empirically derived. They are used in equations similar to Equation 4.2 and sufficiently capture the deformation of the Earth so that tidal affects may be removed from geoid models, gravity observations, GPS observations,

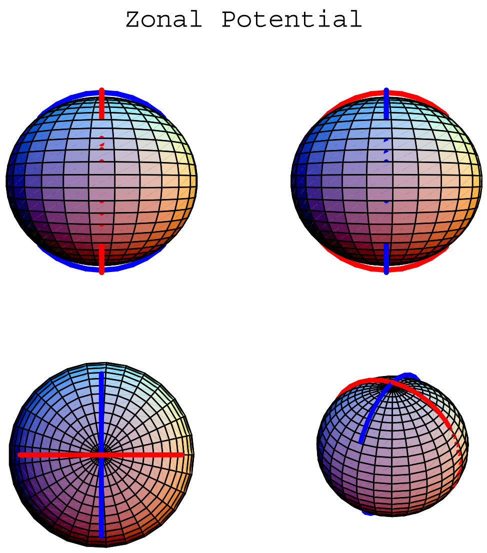

Figure 4.6: The zonal constituent of tidal potential. The red and blue lines are as in Fig. IV.5; the equatorial green line is entirely inside the potential surface. The zonal constituent to tidal potential gives rise to latitudinal tides because it is a function of latitude.

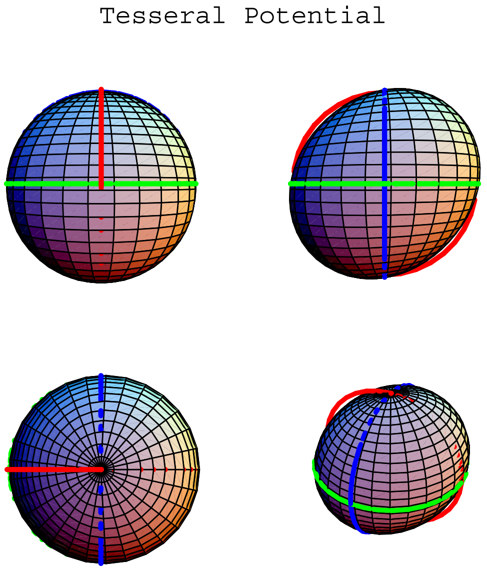

Figure 4.7: The tesseral constituent of tidal potential. The green, red and blue lines are as in Fig. IV.5. The tesseral constituent to tidal potential gives rise to both longitudinal and latitudinal tides, producing a somewhat distorted looking result, which is highly exaggerated in the Figure for clarity. The tesseral constituent accounts for the moon's orbital plane being inclined by about 5 degrees from the plane of the ecliptic.

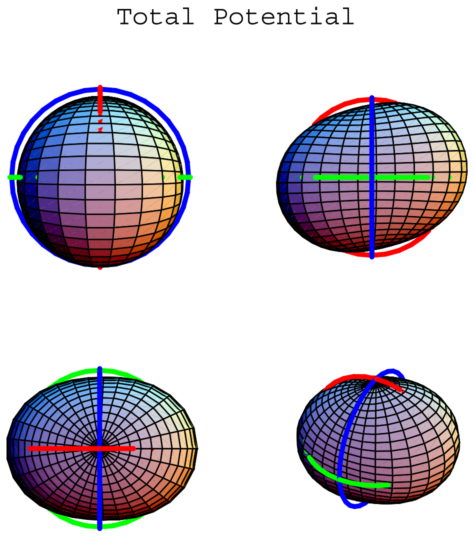

Figure 4.8: The total tidal potential is the combination of the sectorial, zonal, and tesseral constituents. The green, red and blue lines are as in Fig. IV.5. The complicated result provides some insight into why tides have such a wide variety of behaviors.

and other geodetic quantities (Vaníček 1980). However, it should be noted that the permanent tides (those portions of the tidal equations which describe the non-time-varying, or "permanent" deformations) are not *completely* determinable empirically. There are two components of this permanent tide: first, the permanent deformation of Earth's geopotential field due to the existence of the permanent (non-zero time-averaged) Sun and Moon and second, the permanent deformation of Earth's geopotential field due to the existence of the permanent deformation of Earth's crust (which, in turn, is due to the existence of the permanent Sun and Moon).

The first part (called the "direct" component of the permanent Earth tide) is computable empirically, as it deals solely with the Sun's and Moon's mass affecting the Earth's geopotential field. The second part is not computable empirically. This is because the permanent deformation of the Earth's crust can not be directly observed. The Earth's crust perpetually ("permanently") exhibits a deformation due to the permanent existence of the Sun and Moon. Because we can not observe how the crust would react without a permanent Sun and Moon, we can not determine empirically how much permanent deformation actually exists (that is, we can not determine a "zero degree Love number" for the Earth), and thus can not compute what the effect of this permanent crustal deformation is on the Earth's geopotential.

4.2.6 Ocean Tides

Ocean tides affect the geoid by redistributing the mass of the oceans, which has the following effects. First, the redistribution of the water in the oceans creates a discernible change in the geoid. Second, the weight of the water deforms the Earth below it, in addition to the tidal potential also deforming the Earth (Vaníček 1980, pp. 9-12). The deformation of the Earth due to tidal loading can also be modeled by certain Love numbers that parameterize Equation 4.2. The liquid nature of the oceans allows dramatically more complexity in their response to gravitational attraction and, consequently, its modeling is likewise more complex.

4.3 Global Navigation Satellite System (GNSS) Heighting

Global navigation satellite systems, such as the European Union's Galileo system, the Russian *Global'naya Navigatsionnaya Sputnikovaya Sistema* (GLONASS), and the U.S. Global Positioning System (GPS) offer, in conjunction with a highly accurate model of the gravimetric geoid, the potential of determining orthometric heights with centimeter accuracy without conventional leveling. The prospect of establishing vertical control in remote locations without running levels to established distant bench marks holds great promises of time savings, and therefore, cost savings. These savings are the reward for surveyors who practice GPS heighting and were a primary motivation for this series. According to Zilkoski, Carlson & Smith (2000), "GPS-derived orthometric heights can now provide a viable alternative to classical geodetic leveling techniques for many applications."

Deriving orthometric heights from ellipsoid heights is mathematically very simple. As explained in the previous papers, a geoid height is the geometrical separation (distance) from some reference ellipsoid to the geoid, an ellipsoid height is the geometrical separation from some reference ellipsoid to a point of interest, and an orthometric height is the length of the plumb line from the geoid to a point of interest. Were plumb lines straight lines and if they were normal to the reference ellipsoid, these three definitions would immediately lead to an exact relationship:

$$H = h - N, \tag{4.3}$$

where
H = orthometric height;
N = geoid height; and
h = ellipsoid height.
However, plumb lines are curved and not normal to reference ellipsoids, in general. Therefore, we cannot be correct in using an equality relationship and must instead write:

$$H \approx h - N. \tag{4.4}$$

Although Equation 4.4 is not exact, it is close enough for most practical purposes (Hein 1985, Zilkoski & Hothem 1989, Zilkoski 1990, Henning, Carlson & Zilkoski 1998, Vaníček, Huang, Novak, Pagiatakis, Veronneau, Martinec & Featherstone 1999). For example, an extreme case of a two-arc-minute deflection of the vertical would introduce less than two millimeters of error in the orthometric height (Tenzer et al. 2005, p.89), based on Equation 4.4.

Much of the information from this series is contained within Equation Much of the information from this series is contained within Equation 4.4 (Hwang & Hsiao 2003, Kao et al. 2000, Sun 2002). For example, the choice of the reference ellipsoid is important. Local geodetic reference ellipsoids are generally not geocentric, so their normal directions could differ significantly from those of ellipsoids that are geocentric insofar as was possible at the time of their creation. It is important not to mix heighting systems. The GPS surveyor must therefore use a reference ellipsoid of a datum that matches the reference ellipsoid of the gravimetric geoid model. In the U.S., NGS recommends using GEOID03 which is modeled relative to the NAD 83 datum (which uses the GRS 80 ellipsoid). Therefore, for example, GPS heighting should not be done with GEOID03 and the WGS 84 datum. Also, because Equation 4.4 is an approximation rather than an equality (due to the non-parallelism of the equipotential surfaces), dynamic/orthometric corrections will have to be applied to (the purely geometric) spirit leveling measurements (Strang van Hees 1992, Hwang & Hsiao 2003, Kao et al. 2000, Sun 2002). For example, the choice of the reference ellipsoid is important. Local geodetic reference ellipsoids are generally not geocentric, so their normal directions could differ significantly from those of ellipsoids that are geocentric insofar as was possible at the time of their creation. It is important not to mix heighting systems. The GPS surveyor must therefore use a reference ellipsoid of a datum that matches the reference ellipsoid of the gravimetric geoid model. In the U.S., NGS recommends using GEOID03 which is modeled relative to the NAD 83 datum (which uses the GRS 80 ellipsoid). Therefore, for example, GPS heighting should not be done with GEOID03 and the WGS 84 datum. Also, because Equation 4.4 is an approximation rather than an equality (due to the non-parallelism of the equipotential surfaces), dynamic/orthometric corrections will have to be applied to (the purely geometric) spirit leveling measurements (Strang van Hees 1992).

In theory, GPS heighting is simple: determine an ellipsoid height with a GPS receiver and subtract the geoid height, which is provided by a gravimetric geoid model, to obtain the approximate orthometric height. In practice, things are more complicated. This fourth paper now presents a survey of GPS heighting error sources and best practice guidelines put forth by NGS and other authors in the peer-reviewed literature. Although this paper depends in large part on previous work by Zilkoski and others at the NGS (Zilkoski, D'Onofrio & Frakes 1997), it is not our intention to restate that material verbatim (Zilkoski et al. 2000). Instead, this final paper will provide commentary on the guidelines and explanations why some of the recommendations were made. We will emphasize the key issues necessary for achieving the accuracies in those guidelines and provide examples from the literature that illustrate them, when possible. More detailed and comprehensive treatments include (Leick 1995, Hofmann-Wellenhof, Lichtenegger & Collins 1997, Seeber 2003, Hofmann-Wellenhof & Moritz 2005).

4.4 Error Sources

Effective GPS heighting depends upon having an understanding of the measurement error budget and acting in such a manner as to eliminate or mitigate those errors. Error sources have been grouped in three main categories: satellite position and clock errors, signal propagation errors, and receiver errors (Seeber 2003). We will discuss these error sources and explain what, if anything, can or should be done about them according to best practices reported in the current literature. Although it is beyond the scope of this paper to review GNSS as a whole, the reader is referred to the large existing literature on the topic, such as (Leick 1995, Hofmann-Wellenhof et al. 1997, Seeber 2003, Van Sickle 1996) and collections of articles published by the U.S. Institute of Navigation (ION). However, before discussing these error sources, we present issues that arise due to the Earth itself.

4.4.1 Geophysics

There are several issues pertaining to the Earth itself that factor into GNSS heighting. Most of these pertain to the dynamic shape of the Earth but one arises simply because the Earth is opaque at the radio frequencies broadcast by GNSS satellites.

No Satellites Below

We begin by explaining why it is that GNSS positioning cannot be expected to be as accurate for vertical coordinates as for horizontal ones. Currently operational GNSS satellites, abbreviated as SV for "space vehicle," are stationed in orbital planes inclined from the equator by 55 degrees (for GPS) or 64.8 degrees (for GLONASS). Consequently, any place on Earth is always surrounded by SVs, above and below. However, the Earth completely blocks signals from SVs below the horizon from reaching a receiver; the radio signals cannot penetrate solid rock. Therefore, receivers on the ground cannot detect signals from SVs below the horizon. As a result, while it is possible to be surrounded on all azimuth points by SVs, one cannot be surrounded on all zenith angles (essentially none greater than 90 degrees). Consequently, the local vertical is not controlled as well as the local horizontal. As stated by Brunner & Walsh (1993), "We note that even without any tropospheric propagation errors, an inherent geometrical weakness exists in the GPS baseline results that usually makes the determination of height differences worse by a factor of about 3 compared with the horizontal baseline components." Therefore, we cannot expect the best GNSS heighting to be as accurate as the best GNSS horizontal positioning.

Earth Tides, Ocean and Atmospheric Loading

GNSS post-processing software often includes tide corrections which remove these effects, creating a **tide-free** system. See the opening discussion for more elaboration.

Crustal Motion

Plate tectonics constantly move the Earth's crust both horizontally and vertically. Horizontal motions can be accounted for by modeling the position and velocity of fiducial stations and then interpolating to places of interest. The NGS Horizontal Time-Dependent Positioning (HTDP) software (Snay 1999, Snay 2003) allows U.S. users to reconcile control coordinates published in the past with current position measurements that have moved due to plate motion, including earthquakes. Of particular note to heighting, a vertical equivalent, VTDP, has been created for the lower Mississippi valley and the northern Gulf Coast (Shinkle & Dokka 2004). Vertical crustal motion includes both tectonic crustal motion and anthropogenic factors, such as liquid extraction resulting in ground subsidence (Gabrysch & Coplin 1990), which complicates matters considerably.

4.4.2 Satellite Position and Clock Errors

We now begin a discussion of the GNSS error budget. Because GNSS positioning is accomplished by a process similar to trilateration there are two key pieces of information upon which GNSS positioning depends: signal propagation time and the location of the SVs. Signal propagation time is used to infer the range from the SVs to a receiver antenna's phase center, and SV locations are used as the coordinates of the known points in the trilateration scheme. However, the signal propagation time is biased due to an immeasurable time offset between GPS time and a receiver's internal clock; this results in a **pseudo-range** rather than the actual range. The implications of this will be discussed below. Any errors in locating a SV and any inconsistencies in the clocks on board the SVs that govern its operation result directly in positioning errors.

Orbit Errors and Ephemerides

Knowing the position of the satellites at any given moment in time is a cornerstone of how GNSS positioning works. The satellites themselves should be perceived as being moving monuments because pseudo-range positioning (positioning using pseudo-ranges) is based on trilateration: given three (or more) known locations and a distance from those locations to the point of interest, determine the coordinates of the point of interest.[2] Therefore, since the satellites are in motion, it does not suffice to publish a single set of coordinates for them. Instead, ephemerides are created for each SV so that the processing software can determine SV positions at the moment of transmission, which form the basis for the trilateration.

In broad strokes, GNSS ephemerides come in two types: broadcast and precise. Broadcast ephemerides, as the name implies, are broadcast by the SVs and read by GNSS receivers as they

[2] In fact, three known locations and distances do not uniquely determine a three-dimensional position; the problem is reduced to a selection between two solutions. One of these solutions will either be deep inside the Earth or in outer space and can be discarded by inspection for terrestrial GNSS positioning. See Awange & Grafarend (2005) for novel solutions of this problem based on Groebner bases.

operate. Broadcast ephemerides are essentially highly educated, physics-based guesses about the future locations of the SVs based on their past locations and velocities. Precise ephemerides are produced by observing the SVs and deducing their positions after-the-fact. Needless to say, broadcast ephemerides are not as accurate as precise ephemerides. The accuracy of the broadcast ephemerides is currently around one to three meters (Seeber 2003).

The International GNSS Service (IGS) provides three types of precise ephemerides, which differ by how much time elapses before they are available (IGS 2005). The most accurate are the "final ephemerides" which are updated weekly with a latency of about 13 days and have accuracy reported to be better than 5 cm (Seeber 2003). The "rapid ephemerides" are updated daily with a 17-hour latency and an accuracy around 5 cm. Ultra-rapid ephemerides are updated four times daily with a latency of either 3 hours (observed half) or none (predicted half) with an accuracy around 25 cm. It can be shown that the error introduced into computed positions varies by baseline length as a function of ephemeris accuracy: the longer the baseline, the more accurate the ephemeris needs to be (Eckl, Snay, Soler, Cline & Mader 2002) and (Seeber 2003, p.305). For high-accuracy GPS heighting, final precise ephemerides are required by NGS guidelines (Zilkoski et al. 1997).

Satellite Clock Errors

Although GNSS satellites have onboard atomic time standards that are highly accurate and precise, they are not perfect. Like all clocks, atomic clocks drift and experience unpredictable jumps, albeit very small ones (Diddams, Bergquist, Jefferts, & Oates 2004, Flowers 2004). GPS time is a weighted average of the clocks in the controlling station on Earth and the GPS satellite clocks. Each SV clock is monitored for its offset from GPS time, and this time bias estimate is included with the ephemerides, both broadcast and precise, to be accounted for in the positioning software.

4.4.3 GPS Signal Propagation Delay Errors

GNSS ranges are inferred by measuring a (biased) elapsed time from the satellite to the receiver; it is biased due to an immeasurable time offset between GPS time and a receiver's internal clock. This elapsed time interval is scaled to be a distance by multiplying by the speed of light. Although the speed of light is constant in a vacuum, electromagnetic waves propagating through media can be delayed and refracted. GNSS signals propagate through the Earth's atmosphere and are affected by the ionosphere and the troposphere. Both of these atmospheric layers delay the signals, thus introducing timing/ranging errors.

Ionosphere Delays

The ionosphere is a high-altitude (roughly 50 km to 1000 km above the Earth's surface) part of the atmosphere that is composed of charged particles that have been ionized by solar radiation. The ionosphere refracts radio signals in a manner similar to how water in a glass refracts light, such that a pencil appears to have a sharp bend in it. It happens that the ionosphere refracts radio-frequency electromagnetic waves of different frequencies differently. Consequently, it delays the two GPS broadcast frequencies, L1 and L2, differently. This difference can be detected by dual-frequency receivers and subsequently virtually eliminated by post-processing. For more details consult, for example, (Brunner & Walsh 1993, Hofmann-Wellenhof et al. 1997, Leick 1995, Seeber 2003). Single-frequency receivers cannot detect the ionosphere delays, but differencing processing on short baselines can cancel out most of the error, leaving errors on the order of 1 to 2 ppm of the interstation distance (Seeber 2003). The NGS guidelines require dual-frequency receivers for baselines greater than 10 km, and they are the preferred type of GPS receiver for all observations (Zilkoski et al. 1997).

According to Jakowski, Standov & Klaehn (2005, p.3071), “The space weather is defined as the set of all conditions -on the Sun, and in the solar wind, magnetosphere, ionosphere and the thermosphere-that can influence the performance and reliability of space-borne and ground-based technological systems and can endanger human life.” Space weather can significantly influence the propagation of the SV transmissions through the ionosphere, resulting in a degradation of positioning quality (*ibid*). Dual-frequency receivers are not able to eliminate the problems caused by severe space weather, hence observations should not be performed during severe ionospheric storms. The National Oceanic and Atmospheric Administration (NOAA) includes space weather reporting from its Space Environment Center, which is part of the National Weather Service (http://www.sec.noaa.gov/).

Troposphere Delays

The troposphere is that part of the atmosphere in which weather (in the ordinary sense) occurs. Atmospheric density gradients of the troposphere, like the ionosphere, refract GNSS radio waves. However, the tropospheric delays do not depend upon the frequency of the electromagnetic waves. No hardware exists today that can directly measure the delay created by the troposphere, so its affect must be accounted for by modeling the troposphere or by treating it as an unknown nuisance variable determined using least squares techniques.

The errors associated with the troposphere are considered the most problematic member of the GNSS heighting error budget. According to Seeber (2003), “[tropospheric delay] is one of the reasons why the height component is much worse than the horizontal components in precise GPS positioning.” According to Brunner & Walsh (1993), “Tropospheric delay errors mainly affect the accuracy of height differences. Today this must be considered the main limitation of the attainable accuracy using GPS, which seems to be around 2.5 centimeters for height differences of baselines longer than about 50 kilometers.”

Marshall, Schenewerk, Snay & Gutman (2001) performed a detailed study of the affect of tropospheric modeling successfulness at addressing the tropospheric delay on baselines from 62 km to 304 km in length. Based on their experiments conducted using the NGS Continuously Operating Reference Stations (CORS), they show significant reductions in height standard deviations by increasing session duration from one to four hours, and that the choice of the tropospheric model has a strong influence on the precision and accuracy of the resulting heights. Some of these models depend upon measured tropospheric parameters such as atmospheric pressure, atmospheric temperature, and relative humidity, i.e., the quantities that determine the static density of the atmosphere and its density gradient. Others rely on standard models of the atmosphere and are parameterized by latitude and day of the year. Another approach is to treat the tropospheric delay as another unknown parameter and estimate it using statistics from the GPS observables. Marshall et al. (2001) concluded that, “Session lengths shorter than two hours contain insufficient GPS data to estimate both heights and nuisance parameters, and hence more accurate weather information is needed to obtain more precise heights for these shorter sessions.” The models showed a large amount of variability among each other and all of them displayed significant individual variability-more than 5 cm. This fact would appear to contradict NGS claims that following their guidelines should result in 2 cm - 5 cm ellipsoid height accuracy. The difference is the length of the baselines. Marshall’s study had baselines not shorter than 60 km, but NGS requires lines no longer than 10 km. This is an important difference because the unmodeled tropospheric delay error is spatially auto correlated, meaning that the closer two stations are, the more likely they are to “see” the same tropospheric delay. If the delays were exactly the same, they would be canceled by post-processing differencing. To what degree they are not the same, they do not cancel.

According to the current literature, measuring weather parameters is not very helpful. Marshall et al. (2001) found that, “For session lengths greater than two hours, we conclude that sufficiently precise NAD [neutral atmospheric delay] modeling for geodetic activities may be achieved by coupling nuisance parameter estimation with the relatively crude seasonal model.” This means that weather measurements were not needed to achieve sufficiently precise error models. Brunner & Walsh (1993) note that:

> In general, the tropospheric delay models using meteorological ground observations have produced rather poor, and in most cases worse, results compared with the results from the default model values that replaced the actual observed meteorological values. We would like to comment on this surprising finding. Taking accurate meteorological observations is a somewhat difficult task, and frequently large observation errors can occur. In addition, the closeness of the ground and very local micrometeorological conditions severely affect meteorological observations.

These comments appear to support the conclusions found by Marshall et al. (2001). Recently, Ray, Hilla, Dillinger & Mader (2005) noted succinctly: “To the central question, whether measured surface met data can be used to improve geodetic performance, we find no such utility.” Nevertheless, NGS guidelines require meteorological data to be collected (Zilkoski et al. 1997). It has been shown (Marshall et al. 2001) that “Weather fronts may cause the GPS signal delay to vary by greater than 3 centimeters over a 1-hour period, potentially leading to ellipsoidal height errors exceeding 9 cm.” Surface met data are not collected for modeling purposes. Rather, they are useful for *a posteriori* error detection as they help to spot the passing of a weather front through the surveying network, something that could possibly go unnoticed by the ground crews.

The affect of the troposphere increases with zenith angle. For this reason (among others) NGS recommends a 15 degree minimum elevation mask (Zilkoski et al. 1997).

Multipath

One of the two GNSS observables is carrier phase: “carrier” refers to the unmodulated radio signal broadcast by the SVs and “phase” refers to the total number of cycles of the carrier waves from its transmission to its reception, including a partial wavelength at the end. In relative positioning, baselines between phase centers are deduced by differencing phase observations from multiple SVs; see Hofmann-Wellenhof et al. (1997) among many others for more details. Multipath is the situation where GNSS radio signals arrive at the receiver via more than one path. This happens by the signal reflecting from some surface such as a chain link fence, a building, a car, or the ground. According to Seeber (2003), “Multipath influences on carrier phase observations produce a phase shift that introduces a significant periodic bias of several centimeters into the range observation Their propagation into height errors may reach ±15 cm (Georgiadou & Kleusberg 1988)”. Multipath also affects pseudo-range derived positions, introducing errors potentially on the order of meters.

Multipath can be reduced by antenna design, principally choke rings and ground planes, and by elevation masks. Multipath is more likely to occur at low elevation angles so, again, NGS recommends a 15 degree minimum elevation mask (Zilkoski et al. 1997). Ground planes are known to reduce multipath, especially spurious signals arriving at the receiver from below, perhaps being reflected off the ground. Likewise, choke ring antennas mitigate multipath by attenuating reflected signals. Therefore, NGS requires ground planes for GPS antennas and recommends choke rings (Zilkoski et al. 1997). There are also software techniques for multipath reduction (e.g., Seeber, Menge, Volksen, Wubbena & Schmitz (1997) that are available in some processing packages and, sometimes, in the receiver itself (Townsend & Fenton 1994).

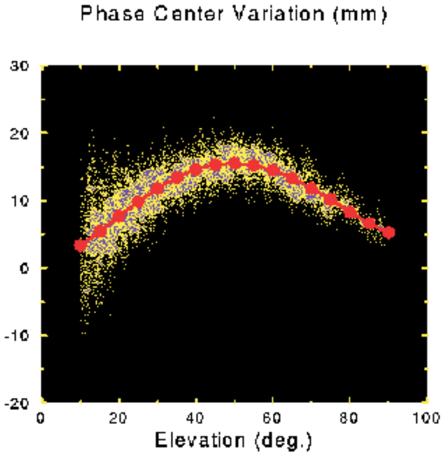

Figure 4.9: This image depicts the location of a GPS receiver's phase center as a function of the elevation angle of the incoming GPS radio signal.

4.4.4 Receiver Errors and Interference

No instrument is perfect, and GNSS receivers are no exception. The receivers themselves cannot determine positions exactly, but we know the error sources associated with the receiver hardware. Also, since the presence of electromagnetic noise in the environment has the potential to interfere with the GNSS radio signals, electromagnetic noise requires some attention, too.

Antenna Phase Center Variation

The electrical phase center of a GNSS receiver antenna is a point in space where the antenna detects the radio signal broadcast from the satellites; it is the point whose coordinates are being determined. That is to say, unless the position is reduced to the antenna reference point (ARP) or a surveying marker, the latitude, longitude, and ellipsoid height reported by the GNSS post-processing package are those of the phase center. Interestingly, the phase center is not on the physical surface of the antenna; indeed, it is not on or in the hardware at all. It is above the antenna and, furthermore, it is not a single location (see Figure IV.9). Although most modern antennas are azimuthally symmetric electrically, local environmental conditions can produce dependences on azimuth. Therefore, phase centers can change with the zenith and azimuth angle of the incoming signal. Additionally, the phase center for L1 is typically different than that for L2 (Mader 1999). Because the phase center is the position being determined by the receiver, as the satellites move, the phase centers move, which is an effect called phase center variation (PCV). As a phase centers moves, its coordinates change. If left uncorrected, phase center variations can introduce as much as a decimeter of error into the vertical coordinate. The NGS antenna calibration program has produced models of phase center variation that are available for downloading at http://www.ngs.noaa.gov/ANTCAL/. These models can be entered into the post-processing software, which will adjust for the effect.

National Geodetic Survey publishes several coordinates for its CORS base stations. Coordinates are currently given in the ITRF00 (epoch 1997.0) and NAD 83 (CORS96) datums for both the ARP and the L1 phase center. Coordinates for the ARP and the phase center are different by several centimeters, typically. For example, the NAD 83 ellipsoid height for the DE6429 NRME COOP CORS L1 phase center is 163.027 m, whereas the ellipsoid height of the ARP is 162.951 m, a difference of 7.6 cm. Surveyors clearly need to be very careful in choosing their control coordinates and know what their post-processing software does with those coordinates. Some packages may assume that the vertical coordinate refers to some particular place, typically the ARP or the phase center; others allow the user to specify to which place the control coordinates refer. Surveyors should take care to pick coordinates that match the expectations of their software or they will introduce systematic vertical errors by accounting for the phase center-ARP separation incorrectly. Furthermore, some packages have antenna geometry databases to allow the software to compute the distance from the ARP to the phase center. Surveyors should check the values in such a database to verify they are correct by comparing with designs provided by manufacturers or by information on the aforementioned NGS antenna calibration website.

Also, GNSS observation files often allow for marker offsets. Some CORS base station RINEX observation files have offsets that reflect the phase center-ARP separation, typically a negative number a few centimeters in magnitude. Surveyors will need to zero these offsets if their processing software assumes the control coordinates refer to the ARP and computes the offsets automatically via the antenna geometry database. If they are not zeroed, the software will account for the distance from the phase center to the ARP twice, introducing a several-centimeter blunder into the vertical control coordinate. Such a blunder can be extremely difficult to find if the processing package does not give a complete account (report) of how the vertical coordinate was determined. The NGS processing software, PAGES, does report all the offsets that go into determining the spatial location of the phase center, so the surveyor knows whether all the control coordinates and offsets are consistent.

Additionally, as CORS stations are increasingly being used in local surveys, it is likely that a mixture of antenna types will occur in a single survey. Any azimuthal PCV inconsistencies among the antennas will not cancel in the differencing processing unless the same inconsistency occurs for all antennas. Therefore, it is important to orient all antennas in the survey to the North so that any residual azimuthal effects are canceled. CORS antennas are already oriented to the North, which means that surveyors need only be concerned about their own antennas.

Electromagnetic interference and signal attenuation

The radio signals currently broadcast by the GPS satellites are relatively low power, around 50 watts. Although GNSS signals occupy a protected frequency band, nearby sources of broadband electromagnetic noise can overwhelm them (Johannessen 1997, Butsch 2002), thus causing decreased signal to noise ratios, increased difficulty or prevention of GNSS signal acquisition, and loss of signal tracking (Seeber 2003, p.320). Power transmission lines, television and radio stations, and radar installations are possible examples of such noise sources. To help address this problem, the GPS modernization program includes a third, higher-power frequency (L5) which is expected to reduce this problem (Hatch, Jung, Enge & Pervan 2000). Unfortunately, new receivers will probably have to be purchased when enough satellites have been placed in orbit to make using L5 practical and to take advantage of its potential. In the mean time, surveyors should occupy sites that are not directly below electromagnetic noise sources, if possible. Overhead vegetation that comes between the receiver and the SVs can also attenuate or block the SV transmissions, causing the same problems as with decreased signal to noise ratios (Spilker 1996, Meyer, Bean, Ferguson & Naismith 2002).

4.4.5 Error Summary

Table tab:IV.1 provides a summary of error sources and recommended remedies.

Table 4.1: A summary of GNSS error sources and their recommended remedy.

Error	Remedy
Orbit errors and clock errors	Use final precise ephemerides; Double differencing of phase observations eliminates orbit and clock errors
Ionospheric delay	Use dual-frequency receivers; Can be reduced on short baselines by differencing phase observables
Tropospheric delay	Modeled or determined in post processing; Longer observation times yield better results; Can be reduced on short baselines by differencing phase observables
Multipath	Avoid multipath-prone locations; Use a ground plane or choke ring antenna
Phase center variation	Use antenna calibration models; Orient antennas to North; Check antenna offsets and antenna geometry databases to ensure consistency with control coordinates
Electromagnetic noise	L5 receivers; Avoid problem sites if possible

4.5 NGS Guidelines for GPS Ellipsoid and Orthometric Heighting

NGS has guidelines and suggested practices that, if followed exactly, are intended to achieve ellipsoid / orthometric height network accuracies of 5 cm (95 percent confidence level) and ellipsoid / orthometric height local accuracies of 2 cm and 5 cm (95 percent) (Zilkoski et al. 1997, Zilkoski et al. 2000). The **local accuracy** of a control point is defined as:

> . . . a value expressed in cm that represents the uncertainty in the coordinates of the control point relative to the coordinates of the other directly connected, adjacent control points at the 95 percent confidence level. The reported local accuracy is an approximate average of the individual local accuracy values between this control point and other observed control points used to establish the coordinates of the control point (Zilkoski et al. 1997).

The network accuracy of a control point is defined as:

> ". . . a value expressed in cm that represents the uncertainty in the coordinates of the control point with respect to the geodetic datum at the 95 percent confidence level. For NSRS network accuracy classification, the datum is considered to be best supported by NGS. By this definition, the local and network accuracy values at CORS sites are considered to be infinitesimal, i.e., to approach zero. (*ibid*)

This section presents an overview of these guidelines and of currently available U.S. geoid models and how local geoid modeling is used in practice.

4.5.1 Three Rules, Four Requirements, Five Procedures

The National Geodetic Survey created a series of rules, requirements and procedures to derive orthometric heights using GPS (Zilkoski et al. 1997, Zilkoski et al. 2000). We now review this material.

Three Rules

Rule 1. Follow NGS' guidelines to establish GPS-derived ellipsoid heights (Zilkoski et al. 1997) when performing a GPS survey;

Rule 2. Use NGS's latest National Geoid Model, i.e., GEOID03 (Roman et al. 2004), when computing GPS-derived orthometric heights; and

Rule 3. Use the latest National Vertical Datum, i.e., NAVD 88 (Zilkoski et al. 1992), height values to control the project's adjusted heights.

We note that GEOID03 is a **hybrid geoid model** for the conterminous U.S. and, as such, has been custom-crafted to fit properly with the NAVD 88 level surface (Milbert 1991, Milbert & Smith 1996*b*, Roman et al. 2004, Smith & Milbert 1999, Smith & Roman 2000, Smith & Roman 2001, Smith 1998). Inferior results would likely result from using a geoid model that had not been so fitted. There are many studies on how to apply local geoid models for surveying purposes; for example see (Amod & Merry 2002, Corchete, Chourak & Khattach 2005, Featherstone & Olliver 2001, Forsberg, Strykowski, Iliffe, Ziebart, Cross, Tscherning, Cruddace, Finch, Bray & Stewart 2002, Fotopoulos 2005, Luo & Chen 2002, Pellinen 1962, Soycan & Soycan 2003, Tranes, Meyer & Massalski 2007). Some of these are studies were across very limited areas (Soycan & Soycan 2003, Tranes et al. 2007) in which the geoid could be adequately modeled with simple polynomial models. The others are local improvements over global models for regions as large as Iberia (Corchete et al. 2005), Hong Kong (Luo & Chen 2002, Luo, Ning, Chen & Yang 2005, Zhan-ji & Yong-qi 2001), the Caribbean Sea (Smith & Small 1999), Taiwan (Hwang & Hsiao 2003), and the British Isles (Featherstone & Olliver 2001, Forsberg et al. 2002, Iliffe, Griffiths & Message 2000, Iliffe, Ziebart, Cross, Forsberg, Strykowski & Tscherning 2003), where a simple polynomial model will not suffice. These approaches depend upon absolute and relative gravity measurements.

Although it can be shown that completely rigorous orthometric heighting also depends on such data (Tenzer et al. 2005), collecting them is impractical for most surveyors. Fortunately, U.S. surveyors need not resort to such efforts because GEOID03 has been shown to be accurate at the 2 cm (95 percent confidence) level on average for the continental U.S (Roman et al. 2004). Although newer versions are planned to be released in the future, GEOID03 is sufficient for GPS orthometric heighting at the 2 cm and 5 cm accuracy levels as put forth by NGS, thus typically eliminating the need for U.S. surveyors to create their own gravimetric geoid models.

Four Requirements (Control)

Requirement 1. GPS-occupy stations with valid NAVD 88 orthometric heights; stations should be evenly distributed throughout (the) project.

Requirement 2. For project areas less than 20 km on a side, surround project with valid NAVD 88 bench marks, i.e., minimum number of stations is four; one in each corner of the project.

Requirement 3. For project areas greater than 20 km on a side, keep distance between valid GPS-occupied NAVD 88 bench marks to less than 20 km.

Requirement 4. For projects located in mountainous regions, occupy valid bench marks at the base and summit of mountains, even if distance is less than 20 km.

NGS guidelines repeatedly stress the need to tie to *valid* NAVD 88 bench marks, although (unfortunately) the criteria for validity are not discussed. Obviously, bench marks without NAVD 88 heights are not valid. This disqualifies NGVD 29 heights or bench marks tied to tide gauges. A valid bench mark is one that has been tied into NAVD 88 and has not been disturbed either by natural and human forces in such a way as to render its published NAVD 88 height inconsistent with the remainder of the network. Caution should be used in areas of ground subsidence or uplift, such as along the U.S. Gulf Coast or in California, for example.

GNSS heighting can take advantage of four-dimensional markers, where they exist. The National Geodetic Survey has conducted "GPS-on-bench-mark" field surveys as part of its height modernization program, thereby establishing many four-dimensional markers: geodetic latitude, longitude, ellipsoid height, and Helmert orthometric height. For example, according to the data sheet for Y88 (PID LX3030) in Connecticut, Y88 is vertical First-Order, Class II; Horizontal Order A and ellipsoid Fourth Order, Class I. Four-dimensional bench marks are very useful for GNSS adjustment software packages because they eliminate the need to estimate any of the four coordinates (usually either ellipsoid or orthometric height) with a model. Occupying bench marks at the bases and summits of mountains helps overcome error sources in geoid models typically caused by a lack of gravity measurements at such places (Featherstone & Alexander 1996, Allister & Featherstone 2001, Dennis & Featherstone 2002, Featherstone & Kirby 2000, Goos, Featherstone, Kirby & Holmes 2003, Kirby & Featherstone 2001, Zhang & Featherstone 2004).

Five Procedures

Procedure 1. Perform a 3-D minimum constraint least squares adjustment of the GPS survey project, i.e., constrain one latitude, one longitude, and one orthometric height value.

Procedure 2. Using the results from the adjustment in procedure 1 above, detect and remove all data outliers. Repeat procedures 1 and 2 until all outliers have been removed.

Procedure 3. Compute differences between the set of GPS-derived orthometric heights from the minimum constraint adjustment (using the latest national geoid model, i.e., GEOID03) from procedure 2 above and published NAVD 88 bench marks.

Procedure 4. Using the result from procedure 3 above, determine which bench marks have valid NAVD 88 height values. This is the most important step in the process. Determining which bench marks have valid heights is critical to computing accurate GPS-derived orthometric heights.

Procedure 5. Using the results from procedure 4 above, perform a constrained adjustment fixing one latitude and one longitude value and all valid NAVD 88 height values.

Correctness is ascertained by repeatability in GPS heighting.

4.6 Discussion and Summary

GNSS surveying is becoming more commonly used for vertical control. GNSS heighting can be attractive from a cost perspective because it offers the possibility of reducing or eliminating the need for leveling runs and trig-heighting, which are very costly. Although GNSS heighting is not a panacea, the prospect of establishing high-quality vertical control in a remote site without running levels to distant bench marks is very attractive.

Unfortunately, traditional training in leveling does not adequately prepare a surveyor to perform GNSS heighting because the two techniques are nearly completely different. For example, different instruments are used for each technique; the concept of a leveling route does not exist in GNSS heighting; they have different error budgets; they reference different vertical datums; and they are even based on different conceptualizations of height itself.

This series presented concepts such as reference ellipsoids, vertical datums, mean sea level, level surfaces and the geoid, gravity and potential, and orthometric vs. geometric vs. ellipsoid heights. From these concepts come applications such as why some reference ellipsoids are suitable as vertical datums while others are not; what is a GNSS receiver really doing when used for heights and how to integrate its measurements with those of a spirit level, and what is an orthometric correction. Finally, this last paper presented practical aspects of GNSS heighting based on suggested practices given by the National Geodetic Survey in light of its height modernization program. This paper considered network design and control, observation strategies, the role and application of geoid models, and the integration of leveled heights with GNSS-determined heights.

Although there are many issues affecting GNSS-determined orthometric heights, we believe the key points are these. GNSS heighting depends on using consistent control, control from a single, modern datum such as NAVD 88. For example, mixing heights in NGVD 29 and those referenced to a mean sea level station with NAVD 88 heights would violate this rule. Since orthometric heights are derived from ellipsoid heights by subtracting the geoid height from them, the geoid model must be referred to the same heighting system as the control. Currently, in the United States, GEOID03 is the correct model to use, although surveys over very small areas can also benefit from polynomial-based geoid models derived from GPS-on-bench-mark observations.

The primary error factor is the difficulty to measure and model wet zenith delay. In arid regions the wet zenith delay is very small, and short occupations (even as short as 30 minutes) have been used successfully. In humid regions, this is seldom true. It has also been shown by several investigators that collecting meteorological measurements for the purpose of tropospheric delay modeling is ineffectual. However, these measurements should be collected for use as evidence regarding which baselines need to be re-observed. It has been shown by Marshall et al. (2001) that "Weather fronts may cause the GPS signal delay to vary by greater than 3 centimeters over a 1-hour period, potentially leading to ellipsoidal height errors exceeding 9 cm." Therefore, weather observations are useful not so much for tropospheric modeling as they are for detecting that a weather front may have passed through unnoticed. The key for reducing tropospheric delay errors to acceptable levels is to keep baselines very short, less than 10 km in length. By doing so the delay at both ends is nearly the same, and it is subsequently removed by post-processing differencing. The accuracy of GNSS heighting on long baselines is currently limited by wet zenith delay errors.

The importance of antenna modeling cannot be overstated, as well. Ellipsoid height errors as much as 10 cm for certain antennas can be introduced simply by failing to include phase center variation correction models in the processing. It is critical to check the database of the post-processing software to ensure that the antenna geometry is entered correctly and that a PCV model is used. Similarly, when using RINEX observations, make sure that the offsets that may come with those data have correct signs for the conventions of your software and that they ultimately refer to your control coordinates, which can be either ARP or phase center. Any mistakes here will introduce a several-centimeter bias in all baselines with an endpoint at the receiver.

Bibliography

Allister, N. & Featherstone, W. (2001), 'Estimation of Helmert orthometric heights using digital barcode leveling, observed gravity and topographic mass-density data over part of Darling Scarp, Western Australia', *Geomatics Research Australia* **75**, 25–52.

Amod, A. & Merry, C. (2002), 'The use of the two-dimensional spherical FFT for quasi-geoid modeling in South Africa', *Survey Review* **36**(285), 508–520.

Awange, J. L. & Grafarend, E. (2005), *Solving algebraic computational problems in geodesy and geoinformatics*, Springer, New York, New York. 333p.

Bearman, G. (1999), *Waves, tides and shallow-water processes*, 2 edn, Butterworth-Heinemann, Boston, Massachusetts. 224p.

Berry, R. (1976), 'History of geodetic leveling in the united states', *Surveying and Mapping* **36**(2), 137–153.

Blakely, R. J. (1995), *Potential Theory in Gravity and Magnetic Applications*, Cambridge University Press, Cambridge. 441p.

Bomford, G. (1980), *Geodesy*, 4 edn, Clarendon Press, Oxford. 561p.

Boon, J. (2004), *Secrets of the tide: Tide and tidal current analysis and applications, storm surges and sea level trends*, Horwood Publishing, Chichester, U.K. 212p.

Broecker, W. (1983), 'The ocean', *Scientific American* **249**, 79–89.

Brunner, F. K. (2002), The role of local quasi-dynamic heights in engineering geodesy, *in* 'INGEO2002, 2nd Conference of Engineering Surveying', Bratislava, pp. 21–24.

Brunner, F. & Walsh, W. (1993), 'Effect of the troposphere on gps measurements', *GPS World* **4**(1), 42–51.

Bugayevskiy, L. M. & Snyder, J. P. (1995), *Map projections: A reference manual*, Taylor & Francis, Philadelphia, PA. 328p.

Buis, A. (2005), *NASA details earthquake effects on the Earth*, Jet Propulsion Laboratory. http://www.jpl.nasa.gov/news/news.cfm?release=2005-009.

Burša, M. (1969), 'Potential of the geoidal surface, the scale factor for lengths and earth's figure parameters from satellite observations', *Studia Geophysica et Geodaetica* **13**, 337–358.

Burša, M. (1994), 'Testing geopotential models', *Earth, Moon and Planets* **64**(3), 293–299.

Burša, M. (1995), *Report of Special Commission SC3, Fundamental constants*, The 21st General Assembly of the International Association of Geodesy, Boulder, Colorado. 2-14 July.

Burša, M., Kouba, J., Kumar, M., Müller, A., Radej, K., True, S., V., V. & Vojtíšková, M. (1999), 'Geoidal geopotential and world height system', *Studia Geophysica et Geodaetica* **43**(4), 327–337.

Burša, M., Radej, K., Sima, Z., True, S. & Vatrt, V. (1997), 'Determination of the geopotential scale factor from topex/poseidon satellite altimetry', *Studia Geophysica et Geodaetica* **41**(3), 203–216.

Butsch, F. (2002), 'Radiofrequency interference and GPS: A growing concern', *GPS World* **13**(10), 40–49.

Cazenave, A. & Nerem, R. (2002), 'Redistributing earth's mass', *Science* **297**(5582), 783–784.

Chao, B. & Gross, R. S. (1987), 'Changes in the earth's rotation and low degree gravitational field induced by earthquakes', *Geophysics Journal of the Royal Astronomical Society* **91**, 569–596.

Chelton, D., Schlax, M., Freilich, M. & Milliff, R. (2004), 'Satellite measurements reveal persistent small-scale features in ocean winds', *Science* **303**(5660), 978–982.

Corchete, V., Chourak, M. & Khattach, D. (2005), 'The high-resolution gravimetric geoid of Iberia: IGG2005', *Geophysical Journal International* **162**(3), 655–684.

Crandall, C. L. (1914), *Geodesy and Least Squares*, John Wiley & Sons, New York. 329p.

Davis, H. & Snider, A. D. (1979), *Introduction to vector analysis*, Allyn and Bacon, Inc., Boston, Massachusetts. 340p.

D.E., S. & Manshina, L. (1971), 'The elasticity theory of dislocation in real earth models and changes in the rotation of the earth', *Geophysics Journal of the Royal Astronomical Society* **23**, 329–354.

Dennis, M. L. & Featherstone, W. (2002), Evaluation of orthometric and related height systems using a simulated mountain gravity field, *in* '3rd Meeting of the International Gravity and Geoid Commission: Gravity and Geoid 2002 - GG2002', Section VI, Thessaloniki, Greece.

Diddams, S., Bergquist, J., Jefferts, S., & Oates, C. (2004), 'Standards of time and frequency at the outset of the 21st century', *Science* **306**(5700), 1318–1324.

DMA (1995), World Geodetic System 1984, its definition and relationship with local geodetic systems, US DMA Technical Report 8350.2, U.S. Defense Mapping Agency, Denver Federal Center: USGS Information Services. 170p.

Doodson, A. (1922), The harmonic development of the tide-generating potential, *in* 'Proceedings of the Royal Society of London', Series A 100, pp. 305–329.

Dracup, J. F. (1995), Geodetic surveys in the United States: The beginning and the next one hundred years 1807 - 1940, *in* 'ACSM/ASPRS Annual Convention & Exposition Technical Papers', Vol. 1, Bethesda, Maryland, p. 24.

Eckl, M., Snay, R., Soler, T., Cline, M. & Mader, G. (2002), 'Accuracy of GPS-derived relative positions as a function of interstation distance and observing-session duration', *Journal of Geodesy* **75**(12), 633–640.

Faller, J. E. & Vitouchkine, A. L. (2003), 'Prospects for a truly portable absolute gravimeter', *Journal of Geodynamics* **35**(4-5), 567–572.

Featherstone, W. & Alexander, K. (1996), ‘An analysis of GPS height determination in Western Australia’, *The Australian Surveyor* **41**(1), 29–34.

Featherstone, W. & Kirby, J. (2000), ‘The reduction of aliasing in gravity anomalies and geoid heights using digital terrain data’, *Geophysical Journal International* **141**(1), 204–212.

Featherstone, W. & Olliver, J. (2001), ‘A review of geoid models over the British Isles: Progress and proposals’, *Survey Review* **36**(280), 78–100.

Fischer, I. (2004), ‘Geodesy? What’s that? Ch. 2’, *ACSM Bulletin* **208**, 43–52.

Flowers, J. (2004), ‘The route to atomic and quantum standards’, *Science* **306**(5700), 1324–1330.

Forsberg, R. (1984), A study of terrain reductions, density anomalies and geophysical inversion methods in gravity field modeling, Technical Report Rep. 355, Department of Geodetic Science and Surveying, The Ohio State University, Columbus, Ohio.

Forsberg, R., Strykowski, G., Iliffe, J., Ziebart, M., Cross, C., Tscherning, C., Cruddace, P., Finch, O., Bray, C. & Stewart, K. (2002), OSGM02: A new geoid model of the British Isles, *in* ‘Gravity and Geoid 2002, 3rd Meeting of the International Gravity and Geoid Commission’, Thessaloniki, Greece, pp. 132–137.

Fotopoulos, G. (2005), ‘Calibration of geoid error models via a combined adjustment of ellipsoidal, orthometric and gravimetric geoid height data’, *Journal of Geodesy* **79**(1-3), 111–123.

Gabrysch, R. & Coplin, L. (1990), Land-surface subsidence resulting from ground-water withdrawals in the Houston-Galveston region, Texas, through 1987, Technical Report Report of Investigations 90-01, Harris-Galveston Coastal Subsidence District. 53p.

Georgiadou, Y. & Kleusberg, A. (1988), ‘On carrier signal multipath effects in relative GPS positioning’, *Manuscripta Geodaetica* **13**, 172–179.

Goos, J., Featherstone, W., Kirby, J. & Holmes, S. A. (2003), ‘Experiments with two different approaches to gridding terrestrial gravity anomalies and their effect on regional geoid computation’, *Survey Review* **37**(288), 92–112.

Gore, J. H. (1889), *Elements of Geodesy*, 2 edn, John Wiley & Sons, New York. 282p.

Grafarend, E. & Ardalan, A. (1997), ‘W0: an estimate in the Finnish Height Datum N60, epoch 1993.4, from twenty-five GPS points of the Baltic Sea Level Project’, *Journal of Geodesy* **71**(11), 673–679.

Groten, E. (2004), ‘Fundamental parameters and current (2004) best estimates of the parameters of common relevance to astronomy, geodesy, and geodynamics’, *Journal of Geodesy* **77**(10-11), 724–797.

Hall, S. (1992), *Mapping the next millennium*, Random House, New York, New York. 477p.

Hatch, R., Jung, J., Enge, P. & Pervan, B. (2000), ‘Civilian GPS: The benefits of three frequencies’, *GPS Solutions* **3**(4), 1–9.

Hein, G. (1985), Orthometric height determination using GPS observations and the integrated geodesy adjustment model, Technical Report NOAA Technical Report No 110 NGS 32, National Oceanic and Atmospheric Administration, Rockville, Maryland. 16p.

Heiskanen, W. A. & Moritz, H. (1967), *Physical Geodesy*, W. H. Freeman and Company, San Fransisco. 364p.

Helmert, F. (1890), 'Die schwerkraft im hochgebirge, insbesondere den tyroler alpen', *Veroff. Konigl. Preuss., Geod. Inst.* **1**.

Henning, W., Carlson, E. & Zilkoski, D. (1998), 'Baltimore county, maryland, navd 88 gps-derived orthometric height project', *Surveying and Land Information Systems* **58**(2), 97–113.

Hicks, S. (1985), 'Tidal datums and their uses-a summary', *Shore & Beach (Journal of the American Shore and Beach Preservation Association)* **53**(1), 27–32.

Hipkin, R. (2002), Defining the geoid by $W = W_0 = U_0$: Theory and practice of a modern height system, *in* 'Gravity and Geoid 2002, 3rd Meeting of the International Gravity and Geoid Commission', Thessaloniki, Greece, pp. 367–377.

Hofmann-Wellenhof, B. H., Lichtenegger, H. & Collins, J. (1997), *Global Position System: Theory and Practice*, 4 edn, Springer–Verlag Wien New York, New York.

Hofmann-Wellenhof, B. H. & Moritz, H. (2005), *Physical Geodesy*, SpringerWienNewYork, New York, New York. 403p.

Hwang, C. (2002), 'Adjustment of relative gravity measurements using weighted and datum-free constraints', *Computers & Geosciences* **28**(9), 1005–1015.

Hwang, C. & Hsiao, Y. (2003), 'Orthometric corrections from leveling, gravity, density and elevation data: a case study in Taiwan', *Journal of Geodesy* **77**(5-6), 279–291.

IGLD85 (1995), Establishment of international great lakes datum (1985), Technical report, Coordinating Committee on Great Lakes Basic Hydraulic and Hydrologic Data, Chicago, Illinois. 48 pp.

Ihde, J. & Augath, W. (2000), The vertical reference system for europe, *in* J. Torres & H. Hornik, eds, 'Report on the Symposium of the IAG Subcommission for Europe (EUREF)', Veröffentlichungen der Bayerischen Kommission für die Internationale Erdmessung, Astronomisch-Geodätische Arbeiten, June 22-24, Tromsö, pp. 99–101.

Iliffe, J., Griffiths, W. & Message, E. (2000), 'Localized geoid determination for engineering control surveys', *Survey Review* **35**(275), 320–328.

Iliffe, J., Ziebart, M., Cross, P., Forsberg, R., Strykowski, G. & Tscherning, C. C. (2003), 'OSGM02: A new model for converting GPS-derived heights to local height datums in Great Britain and Ireland', *Survey Review* **37**(290), 276–293.

Ingle, J. C. J. (2000), Deep-sea and global ocean circulation, *in* W. G. Ernst, ed., 'Earth systems: Processes and issues', Cambridge University Press, Cambridge, U.K., pp. 169–181.

Jakowski, N., Standov, S. & Klaehn, D. (2005), 'Operational space weather service for GNSS precise positioning', *Annales Geophysicae* **23**, 3071–3079.

Johannessen, R. (1997), 'Interference: Sources and symptoms', *GPS World* **8**(11), 45–48.

Kao, S., Hsu, R. & Ning, F. (2000), 'Results of field test for computing orthometric correction based on measured gravity', *Geomatics Research Australia* **72**, 43–60.

Keay, J. (2000), *The Great Arc: the dramatic tale of how India was mapped and Everest was named*, Harper Collins College Publishers, New York. 182p.

Kellogg, O. D. (1953), *Foundations of Potential Theory*, Dover Publications, Inc., New York.

Kirby, J. & Featherstone, W. (2001), 'Anomalously large gradients in the "Geodata 9 Second" digital elevation model of Australia, and their effects on gravimetric terrain corrections', *Cartography* **30**(1), 1–10.

Kumar, M. (2005), 'When ellipsoidal heights will do the job, why look elsewhere?', *Surveying and Land Information Science* **65**(2), 91–94.

Leick, A. (1995), *GPS Satellite Surveying*, 2 edn, John Wiley & Sons, New York.

Luo, Z. & Chen, Y. (2002), 'Evaluation of geo-potential models EGM96, WDM94, GPM98CR in Hong Kong and Shenzhen', *Journal of Geospatial Engineering* **4**(1), 21–30.

Luo, Z., Ning, J., Chen, Y. & Yang, Z. (2005), 'High precision geoid models HKGEOID-2000 for Hong Kong and SZGEOID-2000 for Shenzhen, China', *Marine Geodesy* **28**(2), 191–200.

Mader, G. (1999), 'GPS antenna calibration at the National Geodetic Survey', *GPS Solutions* **3**(1), 1521–1886.

Marmer, H. (1951), Tidal datum planes, Technical Report Special Publication No. 135, NOAA National Ocean Service, U.S. Coast and Geodetic Survey.

Marsden, J. & Tromba, A. J. (1988), *Vector calculus*, 3rd edn, W.H. Freeman and Company, New York, New York. 655p.

Marshall, J., Schenewerk, M., Snay, R. & Gutman, S. (2001), 'The effect of the MAPS weather model on GPS-determined ellipsoidal heights', *GPS Solutions* **5**(1), 1–14.

McCullough, D. (1978), *Path between the seas: The creation of the Panama Canal, 1870-1914*, Simon & Schuster, New York, New York. 704p.

Melchior, P. (1974), 'Earth tides', *Geophysical Surveys* **13**, 275–303.

Meyer, T. H. (2002), 'Grid, ground, and globe: Distances in the GPS era', *Surveying and Land Information Science* **62**(3), 179–202.

Meyer, T. H., Bean, J. E., Ferguson, C. R. & Naismith, J. M. (2002), 'The effect of broadleaf canopies on survey-grade horizontal gps/glonass measurements', *Surveying and Land Information Science* **62**(4), 215–224.

Meyer, T. H., Roman, D. R. & Zilkoski, D. B. (2005*a*), 'What does *height* really mean? part I: Introduction', *Surveying and Land Information Science* **64**(4), 223–234.

Meyer, T. H., Roman, D. R. & Zilkoski, D. B. (2005*b*), 'What does *height* really mean? part II: Physics and gravity', *Surveying and Land Information Science* **65**(1), 5–15.

Milbert, D. G. (1991), Computing GPS-derived orthometric heights with the GEOID90 geoid height model, *in* 'Technical Papers of the 1991 ACSM-ASPRS Fall Convention', Atlanta, Georgia, pp. A46–55.

Milbert, D. G. & Smith, D. A. (1996*a*), Converting GPS height into NAVD 88 elevation with the GEOID96 geoid height model, *in* 'Proceedings of GIS/LIS '96 Annual Conference and Exposition', Denver, Colorado, pp. 681–692.

Milbert, D. G. & Smith, D. A. (1996*b*), Converting GPS height into NAVD 88 elevation with the GEOID96 geoid height model, *in* 'Proceedings of GIS/LIS '96 Annual Conference and Exposition', Denver, Colorado, pp. 681–692.

Molodensky, M. (1945), Fundamental problems of geodetic gravimetry (in russian), Technical Report TRUDY Ts 42, Geodezizdat Novosibirskiy Institut Inzhenerov Geodezii, Aerofotos yemki i Kartografii (NIIGAiK), Moscow.

Moritz, H. (1980), *Advanced physical geodesy*, Abacus Press, Tunbridge Wells, U.K. 500p.

Moritz, H. (2000), 'Geodetic reference system 1980', *Journal of Geodesy* **74**(1), 128–162.

National Geodetic Survey (1986), Geodetic glossary, Technical report, National Geodetic Survey, Rockville, MD. 274p.

National Geodetic Survey (1998), 'National height modernization study report to congress'.

Nesvorny, D. & Sima, Z. (1994), 'Refinement of the geopotential scale factor r0 on the satellite altimetry basis', *Earth, Moon and Planets* **65**(1), 79–88.

NGS (2003). http://www.ngs.noaa.gov/GEOID/GEOID03/images/geoid03.b.jpg.

NOAA (2007). http://oceanservice.noaa.gov/education/kits/tides/media/supp_tide10b.html.

Pellinen, L. (1962), 'Accounting for topography in the calculation of quasigeoidal heights and plumb-line deflections from gravity anomalies', *Bulletin Geodesique* **63**, 57–65.

Qihe, Y., Snyder, J. P. & Tobler, W. R. (2000), *Map projection transformation principles and applications*, Taylor & Francis, Philadelphia, PA. 367p.

Ramsey, A. (1981), *Newtonian Attraction*, Cambridge University Press, Cambridge.

Ray, J. andMorrison, M., Hilla, S., Dillinger, W. & Mader, G. (2005), 'Geodetic sensitivity to surface meteorological data: 24-hr and 6-hr observing sessions', *GPS Solutions* **9**(1), 12–20.

Rebhan, H., Aguirre, M. & Johannessen, J. (2000), 'The Gravity Field and Steady-State Ocean Circulation Explorer Mission - GOCE', *ESA Earth Observation Quarterly* **66**, 6–11.

Roman, D., Wang, Y., Henning, W., & Hamilton, J. (2004), Assessment of the new National Geoid Height Model, GEOID03, *in* 'Proceedings of 2004 Conference of the American Congress on Surveying and Mapping', Nashville, Tennessee. April 19-23.

Schey, H. M. (1992), *Div, Grad, Curl and All That: An Informal Text on Vector Calculus*, W. W. Norton & Company, New York.

Schwarz, C., ed. (1989), *North American Datum of 1983*, number NOS 2 *in* 'NOAA Professional Paper', National Geodetic Information Center, NOAA, Rockville, Maryland. 256p.

Seeber, G. (2003), *Satellite Geodesy*, 2 edn, Walter de Gruyter, New York. 589p.

Seeber, G., Menge, F., Volksen, C., Wubbena, G. & Schmitz, M. (1997), Precise GPS positioning improvements by reducing antenna and site dependent effects, *in* 'IAG Symposium No. 115', Rio de Janeiro, Brazil, pp. 237–244.

Shalowitz, A. (1938), 'The geographic datums of the coast and geodetic survey', *US Coast and Geodetic Survey Field Engineers Bulletin* **12**, 10–31.

Shinkle, K. & Dokka, R. (2004), Rates of vertical displacement at benchmarks in the lower Mississippi valley and the northern Gulf Coast, Technical Report NOAA Technical Report NOS/NGS 50, U.S. Department of Commerce. 147p.

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* **8**, 17–28.

Smith, D. A. & Milbert, D. G. (1999), 'The GEOID96 high-resolution geoid height model for the United States', *Journal of Geodesy* **73**(5), 219–236.

Smith, D. A. & Roman, D. R. (2000), NAVD 88 Helmert orthometric heights from NAD 83 GPS heights and the GEOID99 high resolution geoid height model, *in* '2000 Conference of the American Congress on Surveying and Mapping', Little Rock, Arkansas.

Smith, D. A. & Roman, D. R. (2001), 'GEOID99 and G99SSS:1-arc-minute geoid models for the United States', *Journal of Geodesy* **75**(9), 469–490.

Smith, D. & Small, H. (1999), 'The CARIB97 high-resolution geoid height model for the Caribbean Sea', *Journal of Geodesy* **73**(1), 1–9.

Snay, R. (1999), 'Using the HTDP software to transform spatial coordinates across time and between reference frames', *Surveying and Land Information Systems* **59**(1), 15–25.

Snay, R. (2003), 'Horizontal time-dependent positioning', *Professional Surveyor* **23**(11), 30–32.

Snyder, J. P. (1987), Map projections – a working manual, Professional Paper 1395, U.S. Geological Survey, Washington, DC. 383p.

Soldati, G. & Spada, G. (1999), 'Large earthquakes and earth rotation: The role of mantle relaxation', *Geophysical Research Letters* **26**, 911–914.

Somiglinana, C. (1929), 'Teoria generale del campo gravitazionale dell' ellissoide di rotazione', *Mem. Soc. Astron. Ital.* **IV**.

Soycan, M. & Soycan, A. (2003), 'Surface modeling for GPS-leveling geoid determination', *Newton's Bulletin* **1**, 41–52.

Speed, Jr., M. F., Newton, H. J. & Smith, W. B. (1996*a*), On the utility of mean higher high water and mean lower low water datums for texas inland coastal water, *in* 'Oceans 96 MTS/IEEE', Fort Lauderdale, Florida.

Speed, Jr., M. F., Newton, H. J. & Smith, W. B. (1996*b*), Unfortunate law-unfortunate technology: The legal and statistical difficulties of determining coastal boundaries in areas with non-traditional tides, *in* 'The Third International Conference on Forensic Statistics', The University of Edinburgh, Scotland.

Spilker, J. (1996), Foliage attenuation for land mobile users, *in* 'Global Positioning System: Theory and Applications', Vol. 1, American Institute of Aeronautics and Astronautics, Inc. (Progress in Astronautics and Aeronautics. Vol. 163), Washington, DC, pp. 569–583.

Srinivasan, M. (2004), 'Ocean surface topography from space'. http://topex-www.jpl.nasa.gov/education/factor-height.html.

Strang van Hees, G. (1992), ‘Practical formulas for the computation of the orthometric and dynamic correction’, *Zeitschrift fur Vermessungswesen* **117**.

Sun, W. (2002), ‘A formula for gravimetric terrain corrections using powers of topographic height’, *Journal of Geodesy* **76**(8), 399–406.

Survey, N. G. (1981), *Geodetic Leveling*, National Oceanic and Atmospheric Administration, Rockville, MD.

Tapley, B., Bettadpur, S., Watkins, M. & Reigber, C. H. (2004), ‘The gravity recovery and climate experiment: Mission overview and early results’, *Geophysical Research Letters* **31**, L09607.

Tenzer, R., Vaníček, P., Santos, M., Featherstone, W. & Kuhn, M. (2005), ‘The rigorous determination of orthometric heights’, *Journal of Geodesy* **79**(1-3), 82–92.

Torge, W. (1997), *Geodesy, 2nd ed.*, Walter de Gruyter, New York.

Townsend, B. R. & Fenton, P. (1994), A practical approach to the reduction of pseudorange multipath errors in a L1 GPS receiver, *in* ‘ION GPS-94’, Salt Lake City, Utah, pp. 1–6.

Tranes, M. D., Meyer, T. H. & Massalski, D. (2007), ‘Comparisons of GPS-derived orthometric heights using local geometric geoid models’, *Journal of Surveying Engineering* **133**(1), 6–13.

Van Sickle, J. (1996), *GPS for Land Surveyors*, Ann Arbor Press, Inc., Chelsea, MI.

Vaníček, P. (1980), Tidal corrections to geodetic quantities, Technical Report NOAA Technical Report NOS 83 NGS 14, National Oceanic and Atmospheric Administration, Rockville, Maryland. 30p.

Vaníček, P., Huang, J., Novak, P., Pagiatakis, S., Veronneau, M., Martinec, Z. & Featherstone, W. (1999), ‘Determination of the boundary values for the Stokes-Helmert problem’, *Journal of Geodesy* **73**(4), 180–192.

Vaníček, P. & Krakiwsky, E. (1986), *Geodesy: The Concepts*, 2 edn, Elsevier Scientific Publishing Company, Amsterdam.

Volgyesi, L. (2006), ‘Physical backgrounds of earth’s rotation, revision of the terminology’, *Acta Geodaetica et Geophysica Hungarica* **41**(1), 31–44.

Whitehead, J. (1989), ‘Giant ocean cataracts’, *Scientific American* **260**, 50–57.

Wilhelm, H. & Wenzel, H.-G. (1997), *Tidal phenomena*, Springer, New York, New York. 398p.

Zhan-ji, Y. & Yong-qi, C. (2001), ‘Determination of the Hong Kong gravimetric geoid’, *Survey Review* **36**(279), 23–34.

Zhang, K. & Featherstone, W. E. (2004), ‘Investigation of the roughness of the Australian gravity field using statistical, graphical, fractal and Fourier power spectrum techniques’, *Survey Review* **37**(293), 520–530.

Zilkoski, D. (1990), Establishing vertical control using gps satellite surveys, *in* ‘Proceedings of the 19th International Federation of Surveying Congress (FIG), Commission 5’, Helsinki, Finland, pp. 282–294.

Zilkoski, D. (2001), Vertical datums, *in* D. Maune, ed., 'Digital elevation model technologies and applications: The DEM users manual', American Society for Photogrammetry and Remote Sensing, Bethesda, Maryland, pp. 35–60.

Zilkoski, D., Carlson, E. & Smith, C. (2000), A guide for establishing gps-derived orthometric heights (standards: 2 cm and 5 cm). draft v1.1., Technical report, National Geodetic Survey, Silver Spring, Maryland. 30p.

Zilkoski, D., D'Onofrio, J. & Frakes, S. J. (1997), Guidelines for establishing GPS-derived ellipsoidal heights (Standards: 2cm and 5cm)-Version 4.3, Technical Report NOAA Technical Memorandum NOS NGS-58, National Geodetic Survey, Silver Spring, Maryland.

Zilkoski, D. & Hothem, L. (1989), 'GPS satellite surveys and vertical control', *Journal of Surveying Engineering* **115**(2), 262–281.

Zilkoski, D., Richards, J. & Young, G. (1992), 'Results of the general adjustment of the North American Vertical Datum of 1988', *Surveying and Land Information Systems* **52**(3), 133–149.