Geometric Reference Systems in Geodesy

by

Christopher Jekeli

Division of Geodetic Science
School of Earth Sciences

Ohio State University

August 2016

Foreword

These notes are the result of combining two graduate courses, *Geometric Geodesy* and *Geodetic Astronomy*, taught for many years at Ohio State University to students pursuing the Master of Science degree in Geodesy. Since the early 1990s, parts of these two courses have become irrelevant, anachronistic, and in need of revision. The resulting course, now called *Geometric Reference Systems*, combines the geometrical aspects of terrestrial and celestial reference systems with an emphasis on modern realizations of these geodetic coordinate systems. The adjective, *geometric*, implies that no attempt is made to introduce the Earth's gravity field, which historically formed an integral part of geodetic control. Of course, the gravity field still holds a prominent place in geodesy and it is covered in other courses. But with the advent of the Global Positioning System (GPS), it arguably has a more specialized role to play in establishing and realizing our reference systems, restricted essentially to traditional vertical control. For this reason, the vertical datum is covered only briefly, since a thorough understanding (especially with respect to transformations between vertical datums) can only be achieved with a solid background in geopotential modeling.

These notes are fashioned after corresponding texts of the previous courses, notably R.H. Rapp's lecture notes, P.K. Seidelmann's supplement to the Astronomical Almanac, and the International Earth Rotation and Reference Systems Service (IERS) Technical Notes on reference system conventions. The present exposition is largely self-contained, however, and the reader need only refer to these and other texts in a few instances to obtain an extended discussion. The new reference system conventions recently (2003, 2010) adopted by the International Astronomical Union (IAU) and the IERS have been added in a way that emphasizes and illustrates the evolution of reference systems that new satellite and space observations have wrought. The current (2016) edition of these notes replaces the previous (2006, 2012) editions with several revisions that correct errors or better elaborate some concepts and that bring the entire content up to date, although the general topic is in a permanent state of evolution as new techniques and observational accuracies are achieved. In particular, the upcoming (already implemented in some cases) new paradigms in geodetic control in the U.S. and elsewhere will modernize and bring improved consistency to this important aspect of infrastructure for society and geophysical science.

Problems are included to help the reader get involved in the derivations of the mathematics of reference systems and to illustrate, in some cases, the numerical aspects of the topics.

<this page is left intentionally blank>

Table of Contents

Chapter 1

Introduction

Geodesy is the science of the measurement and mapping of the Earth's surface, and being essentially an application of mathematics it makes use of coordinates and associated reference systems. The object of this book is to study the various local, regional, and global reference systems that are in use to describe coordinates of points on the Earth's surface or in near space and to relate them to each other as well as to some "absolute" frame, visually, a celestial frame. As the title of the book implies, we deal mostly with the geometry of these systems, although the physics of the Earth plays a very important part. However, the relevant geophysics and geodynamics are discussed more comprehensively in other courses on physical geodesy and geodynamics. Also, the mapping of points and their coordinates onto the plane, that is, the topic of map projections, is not treated in this text. The purpose is mainly to explore the geometric definition of reference systems and their practical realizations.

To establish coordinates of points requires that we set up a coordinate system with origin, orientation, and scale defined in such a way that these are accessible to all users. Only until recently, the most accessible reference for coordinates from a global perspective was the celestial sphere of stars that were used primarily for charting and navigation, but also served as a fundamental system to which other terrestrial coordinate systems could be oriented. Still today, the celestial reference system is used for that purpose and may be thought of as the ultimate in reference systems. At the next level, one defines coordinate systems attached to the Earth with various origins (and perhaps different orientations and scale). We thus have two fundamental tasks before us:

1) to establish an external ("inertial") coordinate system of our local universe that we assume remains fixed in the sense of no rotation; and

2) to establish a coordinate system attached to our rotating and orbiting Earth, and in so doing to find the relationship between these two systems.

In fact, we will develop the terrestrial coordinate system before discussing the celestial system, since the latter is almost trivial by comparison and the important aspects concern the transformation between the systems.

1.1 Preliminary Mathematical Relations

Although the conventional and well known Cartesian coordinates, x, y, z, are certainly the simplest from a mathematical perspective, the Earth is nearly spherical and for global applications, some type of curvilinear coordinates may be preferable. Indeed, spherical coordinates and spherical trigonometry are essential tools for the mathematical manipulations of coordinates of objects on the celestial sphere. Similarly, for global terrestrial coordinates, the early map makers used spherical coordinates, although, today, these are rarely used for geodetic terrestrial systems except with justified approximations. It is useful, nevertheless, to review the *polar spherical coordinates*, according to Figure 1.1, where θ is the co-latitude (angle from the pole), λ is the longitude (angle from the *x*-axis), and r is radial distance of a point. Sometimes the latitude, ϕ, is used instead of the co-latitude – but we reserve ϕ for the "geodetic latitude" (Figure 2.5) and use $\psi = \pi/2 - \theta$ instead to mean "geocentric" latitude.

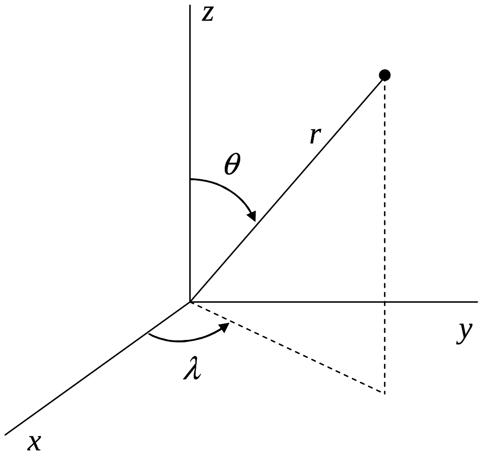


Figure 1.1: Spherical polar coordinates.

On a unit sphere, the "length" (in radians) of a great circle arc is equal to the angle subtended at the center (see Figure 1.2). For a spherical triangle, we have the following useful identities (Figure 1.2):

law of sines: $$\frac{\sin a}{\sin\alpha} = \frac{\sin b}{\sin\beta} = \frac{\sin c}{\sin\gamma}; \tag{1.1}$$

law of cosines: $$\cos c = \cos a \cos b + \sin a \sin b \cos\gamma . \tag{1.2}$$

If a set of coordinate axes is rotated about any axis through the origin, the Cartesian coordinates of a given point change as reckoned in the rotated set. The coordinates change according to an orthogonal transformation, known as a rotation, defined by a 3×3 matrix, e.g., $R(\alpha)$:

$$\begin{pmatrix} x \\ y \\ z \end{pmatrix}_{\text{new}} = R(\alpha)\begin{pmatrix} x \\ y \\ z \end{pmatrix}_{\text{old}} , \tag{1.3}$$

where α is the angle of rotation (positive if counterclockwise as viewed along the axis toward the origin).

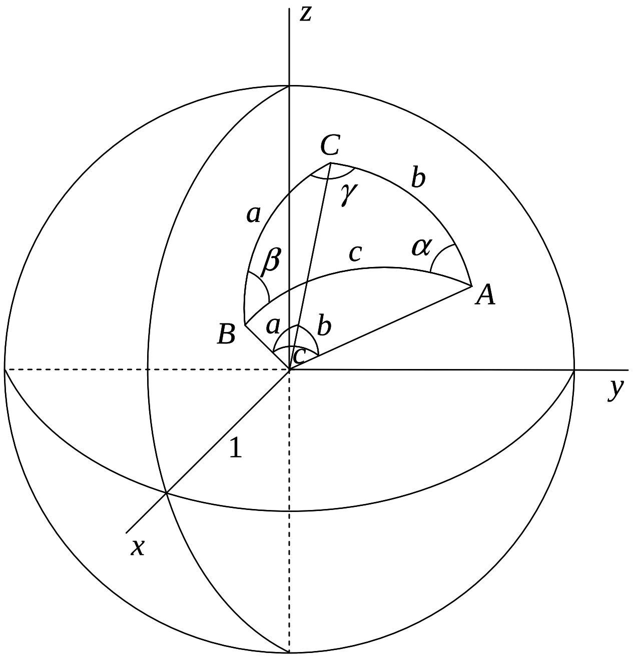


Figure 1.2: Spherical triangle on a unit sphere.

Specifically (see Figure 1.3), a rotation about the x-axis (1-axis) by the angle, α, is represented by

$$R_1(\alpha)=\begin{pmatrix}1 & 0 & 0\\ 0 & \cos\alpha & \sin\alpha\\ 0 & -\sin\alpha & \cos\alpha\end{pmatrix}; \tag{1.4}$$

a rotation about the y-axis (2-axis) by the angle, β, is represented by

$$R_2(\beta)=\begin{pmatrix}\cos\beta & 0 & -\sin\beta\\ 0 & 1 & 0\\ \sin\beta & 0 & \cos\beta\end{pmatrix}; \tag{1.5}$$

and a rotation about the z-axis (3-axis) by the angle, γ, is represented by

$$R_3(\gamma)=\begin{pmatrix}\cos\gamma & \sin\gamma & 0\\ -\sin\gamma & \cos\gamma & 0\\ 0 & 0 & 1\end{pmatrix}; \tag{1.6}$$

where, for example, $\left(R_1(\alpha)\right)^{-1}=R_1(-\alpha)$, and the property of orthogonality yields

$$R_j^{-1}=R_j^{\mathrm{T}}, \quad j=1,2,3. \tag{1.7}$$

The rotations may be applied in sequence and the total rotation thus achieved will always result in an orthogonal transformation. However, the rotations are not commutative; in general, $R(\alpha)R(\beta)\neq R(\beta)R(\alpha)$.

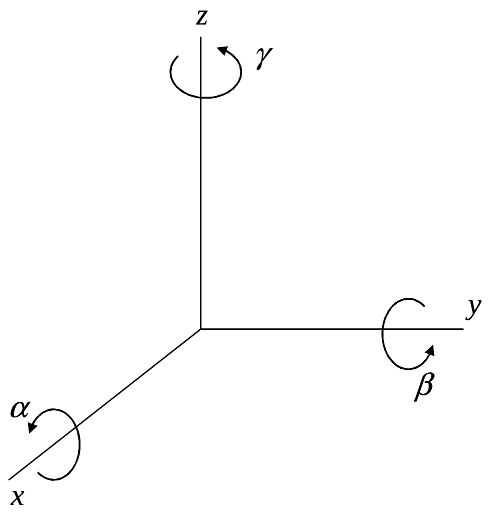


Figure 1.3: Rotations about coordinate axes.

1.2 Reference Systems and Frames

It is important to understand the difference between a reference system for coordinates and a reference frame since these concepts apply throughout the discussion of coordinate systems in geodesy. According to the International Earth Rotation and Reference Systems Service (IERS, see Section 3.3):

> A *Reference System* is a set of prescriptions and conventions together with the modeling required to define at any time a triad of coordinate axes.
>
> A *Reference Frame* realizes the system by means of coordinates of definite points that are accessible directly by occupation or by observation.

A simple example of a reference system is the set of three mutually orthogonal axes that are aligned with the Earth's spin axis, a prime (Greenwich) meridian, and a third direction orthogonal to these two. That is, a system defines how the axes are to be established (e.g., mutual orthogonality and right-handedness), what theories or models are to be used (e.g., what we mean by a spin axis), and what conventions are to be used (e.g., how the *x*-axis is to be chosen – where the Greenwich meridian is). A simple example of a frame is a set of points globally distributed whose coordinates are given numbers that are mutually consistent in the reference system. That is, a frame is the physical realization of the system defined by actual coordinate values of actual points in space that are accessible to anyone. A frame cannot exist without a system, and a system is of no practical value without a frame. The explicit difference between frame and system was articulated fairly recently in geodesy (see, e.g., Moritz and Mueller, 1987, Ch.9), but the concepts have been embodied in the terminology of a *geodetic datum* that can be traced to the eighteenth century and earlier (Torge, 1991; Rapp, 1992). Indeed, the definition of a datum today refers specifically to the conventions that establish how the system is attached to the Earth – its origin, its orientation, and its scale. In this sense the definition of a datum has not changed. The meaning of a datum within the context of frames and systems is explored in more detail in Chapter 3.

1.3 The Earth's Shape

The *Figure of the Earth* is defined as the physical (and mathematical, to the extent it can be formulated) surface of the Earth. It is *realized* by a set of (control) points whose coordinates are determined in some well defined coordinate system. The realization of the system applies traditionally to land areas, but is extended today to include the ocean surface and the ocean floor with appropriate methods for their realizations.

The first approximation to the figure of the Earth is a sphere; and the coordinates to be used would naturally be the spherical coordinates, as defined above (Figure 1.1). Even in antiquity it was recognized that the Earth must be (more or less) spherical in shape. The first actual numerical determination of the size of the Earth is credited to the Greek scholar Eratosthenes (276 – 195 B.C.) who noted that at a particular time of year when the sun is directly overhead in Syene (today's Aswan) it makes an angle, according to his measurement, of $7°12'$ in Alexandria[1]. Further measuring the arc length between the two cities, he used simple geometry (Figure 1.4),

$$R = \frac{s}{\psi}, \tag{1.8}$$

to arrive at a radius of $R = 6267\text{ km}$, which differs from the actual mean Earth radius by only 104 km (1.6%) (scholars think that it may be a lucky result, considering the various assumptions that were made[2]). See also Rapp (1991), who quotes an error of 1% and Torge (2001), who gives an error of 7%. These errors depend entirely on the accepted, perhaps controversial, conversion of the distance unit used by Eratosthenes, the 'Egyptian stadium,' to the current unit, the meter. Further details are given by Fischer (1975).

[1] which, however, is slightly ($3°$) west of Aswan in longitude.

[2] see, e.g., http://en.wikipedia.org/wiki/History_of_geodesy

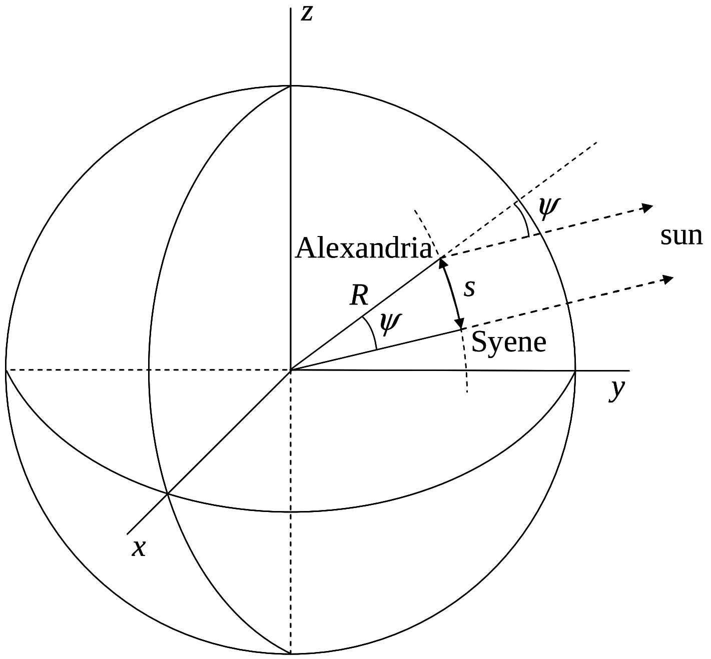


Figure 1.4: Eratosthenes' determination of Earth's radius.

A few other determinations were made, but not until the middle of the Renaissance in Europe (16th century) did the question seriously arise regarding improvements in determining Earth's size. Using very similar, but more elaborate procedures, several astronomers and scientists made various determinations with not always better results. Finally by the time of Isaac Newton (1643 – 1727) the question of the departure from the spherical shape was debated. Various arc measurements in the 17th and 18th centuries, as well as Newton's (and others') arguments based on physical principles, gave convincing proof that the Earth is *ellipsoidal* in shape, flattened at the poles, with approximate rotational symmetry about the polar axis. An engaging account of the controversy raging through the scientific community in the 18th century of whether the Earth is flattened or elongated at the poles is given by Whitaker (2004). The debate was finally resolved conclusively with measurements of arcs near the equator and at higher latitude, where a pole-flattened Earth implies that one degree of arc is subtended by a shorter distance near the equator.

The next best approximation to the figure of the Earth, after the ellipsoid, is known as the *geoid*, the equipotential surface of the Earth's gravity field that closely approximates mean sea level. An *equipotential* surface is a surface on which the gravity potential is a constant value. While the mean Earth sphere deviates radially by up to 14 km (at the poles) from a mean Earth ellipsoid (a surface generated by rotating an ellipse about its minor axis; see Chapter 2), the difference between the mean Earth ellipsoid and the geoid amounts to no more than 110 m, and in a root-mean-square sense by only 30 m. Thus, at least over the oceans (over 70% of Earth's surface), the ellipsoid is an extremely good approximation (5 parts per million) to the figure of the Earth. Although this is not sufficient accuracy for geodesists, it serves as a good starting point for many applications; the ellipsoid is also the mapping surface for most national and international control surveys. Therefore, a study of the geometry of the ellipsoid is given in some detail in the next chapter.

1.4 Problems

1. Write both the forward and the reverse relationships between Cartesian coordinates, (x, y, z), and spherical polar coordinates, (r, θ, λ).

2. Write the law of cosines for the spherical triangle, analogous to (1.2), when the left side is $\cos b$. Also, write the law of cosines for the triangle angles, instead of the triangle sides (consult a book on spherical trigonometry).

3. Show that for rotations about the x-, y-, and z-axes, by small angles, α, β, and γ, the following approximation holds:

$$R_3(\gamma) R_2(\beta) R_1(\alpha) = \begin{pmatrix} 1 & \gamma & -\beta \\ -\gamma & 1 & \alpha \\ \beta & -\alpha & 1 \end{pmatrix}; \tag{1.9}$$

and, that this is independent of the order of the rotation matrices.

4. Determine the magnitude of the angles that is allowed so that the approximation (1.9) does not cause errors greater than 1 mm when applied to terrestrial coordinates (use the mean Earth radius, $R = 6371$ km).

Chapter 2

Coordinate Systems in Geodesy

Coordinates in geodesy traditionally have conformed to the Earth's shape, being spherical or a type of ellipsoidal coordinates for regional and global applications, and Cartesian for local applications where planar geometry suffices. Nowadays, with satellites providing essential reference systems for coordinates, the Cartesian type is as important and useful for global geospatial referencing. Because the latitude/longitude concept will always have the most direct appeal for terrestrial applications (surveying, near-surface navigation, positioning and mapping), it is important to study in detail the coordinates associated with an ellipsoid. In addition, since astronomic observations have a profound historical significance in defining and realizing our reference systems and should be in the knowledge bank of any geodesist, both natural (astronomic) and celestial coordinates are covered. Local coordinates are based on the local vertical and deserve special attention, not only with respect to the *definition* of the vertical, but in regard to their connection to global coordinates. In all cases the coordinate systems are orthogonal, meaning that surfaces of constant coordinates intersect always at right angles. Some Cartesian coordinate systems, however, are left-handed, rather than the usual right-handed, and this will require extra (but not burdensome) care.

2.1 The Ellipsoid and Geodetic Coordinates

We treat the ellipsoid of revolution, its geometry, associated coordinates of points on or above (below) it, and geodetic problems of positioning and establishing networks in an elementary way. The motivation is to give the reader a more practical appreciation and utilitarian approach rather than a purely mathematical treatise of ellipsoidal geometry (especially differential geometry), as well as a window into past geodetic practices. The reader may argue that even the present text is rather mathematical, which, however, cannot be avoided (and no apologies are made), and, that forays into historical methods have little bearing on modern geodesy, but they offer a deeper appreciation for the marvels of satellite-based geodetic control.

2.1.1 Basic Ellipsoidal Geometry

It is assumed that the reader is familiar at least with the basic shape of an ellipse (Figure 2.1). The *ellipsoid* for geodetic applications is formed by rotating an ellipse about its *minor* axis, which for present visualization is aligned with the Earth's spin axis. This creates a surface of revolution that is symmetric with respect to the polar axis and the equator. Because of this symmetry, one often depicts the ellipsoid simply as an ellipse. The basic geometric construction of an ellipse is as follows: for any two points, F_1 and F_2, called *focal points*, the ellipse is the locus (path) of points, P, such that the sum of the distances $\overline{PF_1} + \overline{PF_2}$ is a constant.

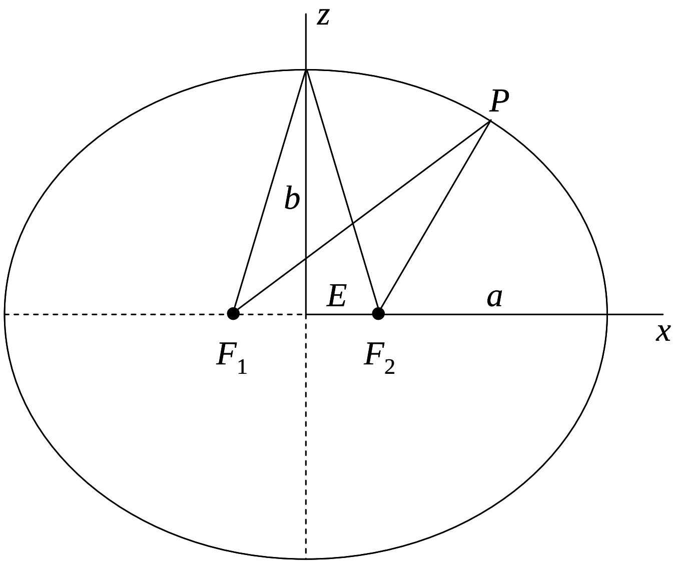


Figure 2.1: The ellipsoid represented as an ellipse.

Introducing a coordinate system (x, z) with origin halfway on the line, $\overline{F_1F_2}$, and z-axis perpendicular to $\overline{F_1F_2}$, we see that if P is on the x-axis, this constant is equal to twice the distance from P to the origin; this is the length of the *semi-major axis*; call it a:

$$\overline{PF_1} + \overline{PF_2} = 2a \,. \qquad (2.1)$$

Moving the point, P, to the z-axis, and letting the distance from the origin point to either focal point (F_1 or F_2) be E, we also find by the theorem of Pythagoras that

$$E = \sqrt{a^2 - b^2}\,, \tag{2.2}$$

where b is the length of the *semi-minor* axis. E is called the *linear eccentricity* of the ellipse (and of the ellipsoid). From these geometrical considerations it is easy to prove (left to the reader), that the equation of the ellipse is given by

$$\frac{x^2}{a^2} + \frac{z^2}{b^2} = 1\,. \tag{2.3}$$

An alternative geometric construction of the ellipse is shown in Figure 2.2, where points on the ellipse are the intersections of the projections, perpendicular to the axes, of points, A and B, sharing the same radius to concentric circles with radii, a and b, respectively. That is, a point, P, on the ellipse is the intersection of lines $\overline{AD}$ and $\overline{BF}$. The proof is as follows. Let x, z, s be distances as shown in Figure 2.2. Now

$$\Delta OCB \sim \Delta ODA \quad \Rightarrow \quad \frac{z}{b} = \frac{s}{a} \quad \Rightarrow \quad \frac{z^2}{b^2} = \frac{s^2}{a^2}\,;$$

but $x^2 + s^2 = a^2$; hence $0 = \dfrac{z^2}{b^2} - \dfrac{a^2 - x^2}{a^2} = \dfrac{x^2}{a^2} + \dfrac{z^2}{b^2} - 1$. Hence, P represented by (x, z) is on the ellipse. QED

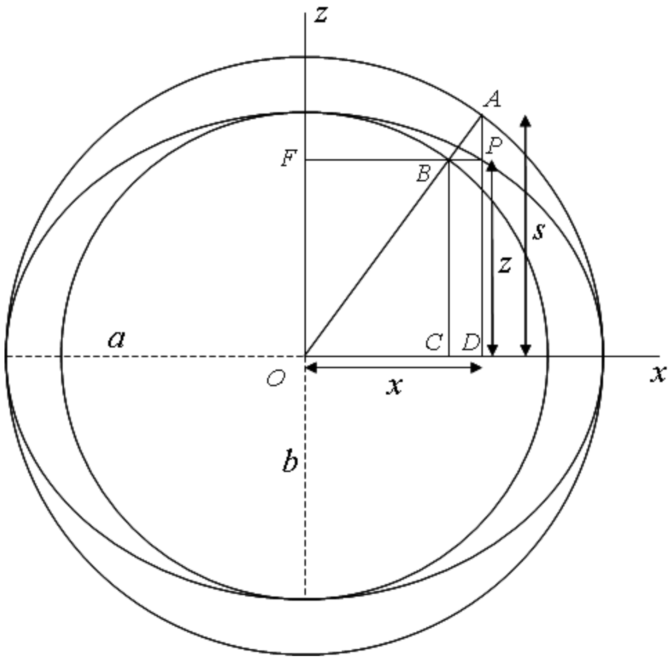


Figure 2.2: Ellipse construction.

The ellipse, and hence the ellipsoid, is defined by two essential parameters: a shape parameter and a size (or scale) parameter (unlike the circle or sphere that requires only one parameter, the radius, which specifies its size). In addition to the semi-major axis, a, that usually serves as the size parameter, any one of a number of shape parameters could be used. We have already encountered one of these, the linear eccentricity, E. The following are also used; in particular, the *flattening*:

$$f = \frac{a-b}{a}; \tag{2.4}$$

the *first eccentricity*:

$$e = \frac{\sqrt{a^2-b^2}}{a}; \tag{2.5}$$

and, the *second eccentricity*:

$$e' = \frac{\sqrt{a^2-b^2}}{b}. \tag{2.6}$$

Note that the shape parameters (2.4), (2.5), and (2.6) are unit-less, while the linear eccentricity, (2.2) has units of distance. There are useful relationships among these parameters (which are left to the reader to derive):

$$e^2 = 2f - f^2, \tag{2.7}$$

$$E = ae, \tag{2.8}$$

$$e^2 = \frac{e'^2}{1+e'^2}, \quad e'^2 = \frac{e^2}{1-e^2}, \quad \left(1-e^2\right)\left(1+e'^2\right) = 1, \tag{2.9}$$

$$e'^2 = \frac{2f - f^2}{\left(1-f\right)^2}. \tag{2.10}$$

When specifying a particular ellipsoid, one generally denotes it by the pair of parameters, (a, f). Many different ellipsoids have been defined in the past. The current internationally adopted mean Earth ellipsoid is part of the Geodetic Reference System of 1980 (GRS80) and has parameter values given by

$$\begin{aligned} a &= 6378137 \text{ m} \\ f &= 1/298.257222101 \end{aligned} \tag{2.11}$$

Table 2.1 from (Rapp, 1991, p.169) lists ellipsoids defined in modern geodetic history. The parameter estimates of the best-fitting ellipsoid (in the mean tide system) were published by Groten (2004) as

$$\begin{aligned} a &= 6378136.72 \pm 0.1 \text{ m} \\ 1/f &= 298.25231 \pm 0.00001 \end{aligned} \tag{2.12}$$

Note that these values do not define an adopted ellipsoid; they include standard deviations and merely give the best determinable values based on current technology. On the other hand, certain specialized observing systems, like the TOPEX satellite altimetry system, have adopted ellipsoids that differ from the standard ones like GRS80 or WGS84 (Table 2.1). It is, therefore, extremely important that the user of any system of coordinates or measurements understands what ellipsoid is implied. It is noted that the IERS (Petit and Luzum 2010) recommends the use of the GRS80 ellipsoid.

Table 2.1: Terrestrial Ellipsoids, from (Rapp 1991, Table 10.1).

Ellipsoid Name (year computed)	semi-major axis, a [m]	inverse flattening, $1/f$
Airy (1830)	6377563.396	299.324964
Everest (1830)	6377276.345	300.8017
Bessel (1841)	6377397.155	299.152813
Clarke (1866)	6378206.4	294.978698
Clarke (1880)	6378249.145	293.465
Modified Clarke (1880)	6378249.145	293.4663
International (1924)	6378388.	297.
Krassovski (1940)	6378245.	298.3
Mercury (1960)	6378166.	298.3
Geodetic Reference System (1967), GRS67	6378160.	298.2471674273
Modified Mercury (1968)	6378150.	298.3
Australian National	6378160.	298.25
South American (1969)	6378160.	298.25
World Geodetic System (1966), WGS66	6378145.	298.25
World Geodetic System (1972), WGS72	6378135.	298.26
Geodetic Reference System (1980), GRS80	6378137.	298.257222101
World Geodetic System (1984), WGS84	6378137.	298.257223563
TOPEX/Poseidon (1992) (IERS recomm.)*	6378136.3	298.257

* McCarthy (1992)

2.1.1.1 Problems

1. From the geometrical construction described prior to equation (2.3), derive the equation for an ellipse, (2.3). [Hint: For a point on the ellipse, show that

$$\sqrt{\left(x+E\right)^2+z^2}+\sqrt{\left(x-E\right)^2+z^2}=2a\ .$$

Square both side and show that

$$2a^2-x^2-E^2-z^2=\sqrt{\left(x+E\right)^2+z^2}\sqrt{\left(x-E\right)^2+z^2}\ .$$

Finally, square both sides again and reduce the result to find (2.3).]
What would the equation be if the center of the ellipse were not at the origin of the coordinate system?

2. Derive equations (2.7) through (2.10).

3. Consider the determination of the parameters of an ellipsoid, including the coordinates of its center, with respect to the Earth. For example, suppose it is desired to find the ellipsoid that best fits through a given number of points at mean sea level. Assume that the orientation of the ellipsoid is fixed so that its axes are parallel to the global, geocentric coordinate frame attached to the Earth.

a) What is the minimum number of points with known $\left(x,y,z\right)$ coordinates that are needed to determine the ellipsoid and its center coordinates? Justify your answer.

b) Describe cases where the geometry of a given set of points would not allow (a robust) determination of 1) the flattening, 2) the size of the ellipsoid.

c) What distribution of points would give the strongest solution? Provide a sufficient discussion to support your answer.

d) Set up the linearized observation equations and the normal equations for a least-squares adjustment of the ellipsoidal parameters (including its center coordinates).

2.1.2 Ellipsoidal Coordinates

In order to define practical coordinates of points in relation to the ellipsoid, we consider the ellipsoid with conventional (x, y, z) axes whose origin is at the center of the ellipsoid. For any particular point, P, in space, it is necessary first to define the *meridian plane* for that point. It is the plane that contains the point, as well as the minor axis of the ellipsoid. The *longitude* of P is then given by the angle in the equatorial plane from the x-axis to the meridian plane. This is the same as the spherical longitude (due to the rotational symmetry); see Figure 1.1. For the latitude, we have a choice. The *geocentric latitude* of P is the angle, ψ, at the origin and in the meridian plane from the equator to the *radial* line through P (Figure 2.3). Note, however, that the geocentric latitude is independent of any defined ellipsoid and, as already noted in Section 1.1, it is identical to the complement of the polar angle defined for the spherical coordinates.

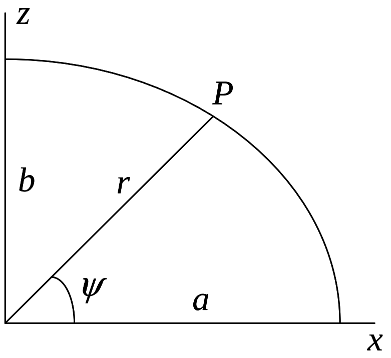


Figure 2.3: Geocentric latitude.

Next, consider the ellipsoid through P that is confocal (sharing the same focal points) with the ellipsoid, (a, f); that is, it has the same linear eccentricity, E. Its semi-minor axis is u (Figure 2.4), which can also be considered a *coordinate* of P. The *reduced latitude*, β, of P is defined as the angle at the origin and in the meridian plane from the equator to the radial line that intersects the projection of P, along the perpendicular to the equator, at the sphere of radius, $v = \sqrt{E^2 + u^2}$.

Finally, we introduce the most common latitude used in geodesy, appropriately called the *geodetic latitude*. This is the angle, ϕ, in the meridian plane from the equator to the line through P that is also perpendicular to the basic ellipsoid (a, f); see Figure 2.5. The perpendicular to the ellipsoid is also called the *normal* to the ellipsoid. Both the reduced latitude and the geodetic latitude depend on the underlying ellipsoid, (a, f).

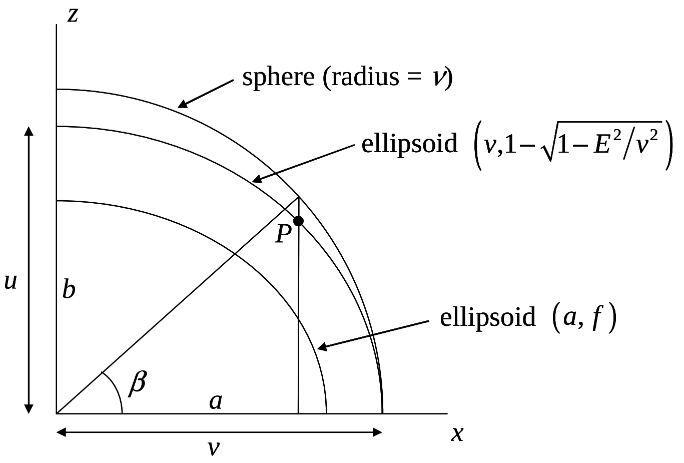


Figure 2.4: Reduced latitude, β, of P. Ellipsoid (a, f) and the ellipsoid through P have the same linear eccentricity, E.

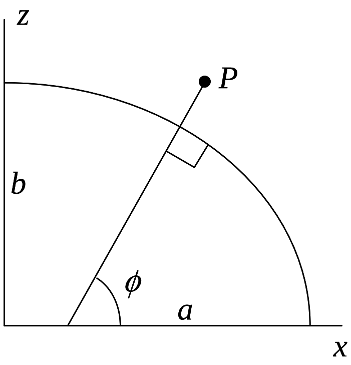


Figure 2.5: Geodetic latitude.

The relationships between these various latitudes may be determined by formulating the coordinates x, z of P in terms of each type of latitude. It turns out that these relationships are straightforward only when P is on the ellipsoid; but for later purposes, they are derived for arbitrary points. For the geocentric latitude, ψ, simple trigonometry gives (Figure 2.3)

$$x = r\cos\psi, \quad z = r\sin\psi. \tag{2.13}$$

Also, for the reduced latitude, simple trigonometric formulas applied in Figure 2.4 as in Figure 2.2 yield

$$x = v\cos\beta, \quad z = u\sin\beta. \tag{2.14}$$

For the geodetic latitude, consider first the point, P, on the ellipsoid, (a, f). From Figure 2.6, the geometric interpretation of the derivative, or slope, of the ellipse gives

$$\tan\left(90° - \phi\right) = \frac{dz}{-dx} . \tag{2.15}$$

The right side is determined from equation (2.3),

$$z^2 = b^2\left(1 - \frac{x^2}{a^2}\right) \quad \Rightarrow \quad 2z\,dz = -2\frac{b^2}{a^2}x\,dx \quad \Rightarrow \quad \frac{dz}{-dx} = \frac{b^2}{a^2}\frac{x}{z} ; \tag{2.16}$$

and, when substituted into equation (2.15), this yields

$$b^4 x^2 \sin^2\phi = a^4 z^2 \cos^2\phi . \tag{2.17}$$

Also from equation (2.3), there is

$$b^2 x^2 + a^2 z^2 = a^2 b^2 . \tag{2.18}$$

Now, multiply equation (2.18) by $-b^2 \sin^2\phi$ and add it to equation (2.17), thus obtaining

$$z^2\left(a^2\cos^2\phi + b^2\sin^2\phi\right) = b^4 \sin^2\phi , \tag{2.19}$$

which reduces to

$$z = \frac{a\left(1 - e^2\right)\sin\phi}{\sqrt{1 - e^2\sin^2\phi}} . \tag{2.20}$$

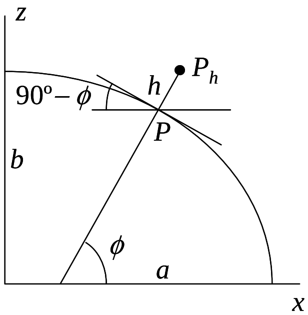


Figure 2.6: Slope of ellipsoid.

With a similar procedure, multiplying equation (2.18) by $a^2\cos^2\phi$, adding it to equation (2.17), and simplifying, one obtains (*the reader should verify this*)

$$x = \frac{a\cos\phi}{\sqrt{1-e^2\sin^2\phi}}. \tag{2.21}$$

To find the (x,z) coordinates of a point above (or below) the ellipsoid, we need to introduce a height coordinate, in this case the *ellipsoidal height,* h, above the ellipsoid (it is negative, if P is below the ellipsoid); h is reckoned along the perpendicular (the normal) to the ellipsoid (Figure 2.6). It is a simple matter now to express (x,z) in terms of geodetic latitude and ellipsoidal height:

$$x = \frac{a\cos\phi}{\sqrt{1-e^2\sin^2\phi}} + h\cos\phi, \qquad z = \frac{a\left(1-e^2\right)\sin\phi}{\sqrt{1-e^2\sin^2\phi}} + h\sin\phi. \tag{2.22}$$

It is easy to find the relationship between the different latitudes, *if the point is on the ellipsoid* ($h=0$). Combining equations (2.13), (2.14), both specialized to the basic ellipsoid ($u=b$), with equations (2.20) and (2.21), one obtains the following relationships among these three latitudes, using the ratio z/x,

$$\tan\psi = \frac{b}{a}\tan\beta = \frac{b^2}{a^2}\tan\phi, \tag{2.23}$$

which also shows that

$$\psi \le \beta \le \phi. \tag{2.24}$$

Again, it is noted that the relationship (2.23) holds only for points on the ellipsoid. For arbitrary points in space the problem is not straightforward and is connected with the problem of finding the geodetic latitude from given rectangular (Cartesian) coordinates of the point (see Section 2.1.5).

The ellipsoidal height, geodetic latitude, and longitude, (h,ϕ,λ), constitute the *geodetic coordinates* of a point with respect to a given ellipsoid, (a,f). These are *orthogonal* coordinates, in the sense that surfaces of constant h, ϕ, and λ are orthogonal to each other. The triple of *ellipsoidal coordinates,* (u,β,λ), is also orthogonal; and, in fact, these coordinates are more useful than geodetic coordinates for some mathematical developments since the coordinate surfaces are relatively simple constructs (Problem 2.1.2.1-4). The surface of constant h, on the other hand, is not a simple shape (it is not an ellipsoid). However, the geodetic coordinates, ϕ, h, are certainly more intuitive than β, u. They are the coordinates of choice for many standard geodetic applications.

2.1.2.1 Problems

1. Derive the following expressions for the differences between the geodetic latitude and the geocentric, respectively, the reduced latitudes of points on the ellipsoid:

$$\tan\left(\phi-\psi\right)=\frac{e^2\sin 2\phi}{2\left(1-e^2\sin^2\phi\right)}, \tag{2.25}$$

$$\tan\left(\phi-\beta\right)=\frac{n\sin 2\phi}{1+n\cos 2\phi}, \tag{2.26}$$

where $n=\left(a-b\right)/\left(a+b\right)$. (Hint: see Rapp 1991, p.26.)

2. Calculate and plot the differences (2.25) and (2.26) for all latitudes, $0\le\phi\le 90°$ using the GRS80 ellipsoid parameter values.

3. Show that the difference $\left(\phi-\beta\right)$ is maximum when $\phi=\frac{1}{2}\cos^{-1}\left(-n\right)$.

4. Mathematically and geometrically describe the surfaces of constant u, β, and, λ, respectively. As the linear eccentricity approaches zero, what do these ellipsoidal coordinates and surfaces degenerate into?

2.1.3 Elementary Differential Geodesy

This section derives the differential elements on the surface of the ellipsoid and, in the process, describes the curvature of the ellipsoid. The differential elements are used in developing the geometry of geodesics on the ellipsoid and in solving the principal problems in geometric geodesy, namely, determining the endpoint coordinates of geodesics, which are the elements (sides) of triangulation networks.

2.1.3.1 Radii of Curvature

Consider a curve on a surface, for example a meridian arc or a parallel circle on the ellipsoid, or any other arbitrary curve. The meridian arc and the parallel circle are examples of *plane curves*, curves that are contained in a plane that intersects the surface. The amount by which the tangent to the curve changes in direction as one moves along the curve indicates the *curvature* of the curve. Curvature may be defined geometrically as follows:

> The *curvature*, χ, of a plane curve is the absolute rate of change of the slope angle of the tangent line to the curve with respect to arc length along the curve.

If α is the slope angle and s is arc length, then mathematically,

$$\chi = \left| \frac{d\alpha}{ds} \right| . \tag{2.27}$$

With regard to Figure 2.7a, let $\boldsymbol{\lambda}$ be the unit tangent vector at a point on the curve; the direction of $\boldsymbol{\lambda}$ identifies the slope of the curve at that point. Consider also the plane that locally contains the infinitesimally close neighboring tangent vectors; that is, it contains the direction in which $\boldsymbol{\lambda}$ changes due to the curvature of the curve. For plane curves, this is the plane that contains the curve. The unit vector that is in this plane and perpendicular to $\boldsymbol{\lambda}$, called $\boldsymbol{\mu}$, identifies the direction of the *principal normal* to the curve. Note that the curvature, as given in equation (2.27), has units of inverse-distance. The reciprocal of the curvature is called the *radius of curvature*, ρ:

$$\rho = \frac{1}{\chi} . \tag{2.28}$$

The radius of curvature is a distance along the principal normal to the curve. In the special case that the curvature is a constant, the radius of curvature is also a constant and the curve is the arc of a circle. One may think of the radius of curvature at a point of an arbitrary curve as being the radius of the circle tangent to the curve at that point and having the same curvature.

A curve on the surface may also have curvature such that it cannot be embedded in a plane. A corkscrew is such a curve. Geodesics on the ellipsoid are geodetic examples of such curves. In this case, the curve has double curvature, or *torsion*. We consider only plane curves for the moment.

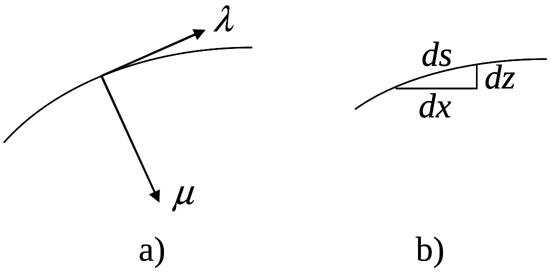


Figure 2.7: Curvature of plane curves.

Let $z = z(x)$ describe the plane curve in terms of space coordinates (x, z). In terms of arc length, s, one can write $x = x(s)$ and $z = z(s)$; and, a differential arc length, ds, is then given by

$$ds = \sqrt{dx^2 + dz^2} \ . \tag{2.29}$$

This can be re-written as

$$ds = \sqrt{1 + \left(\frac{dz}{dx}\right)^2} \ dx \ . \tag{2.30}$$

Now, the tangent of the slope angle of the curve is exactly the derivative of the curve, dz/dx; hence

$$\alpha = \tan^{-1}\left(\frac{dz}{dx}\right). \tag{2.31}$$

Using equations (2.27) and (2.30), one obtains for the curvature,

$$\begin{aligned}\chi &= \left|\frac{d\alpha}{ds}\right| = \left|\frac{d\alpha}{dx}\right|\left|\frac{dx}{ds}\right| \\ &= \frac{1}{1+\left(\frac{dz}{dx}\right)^2}\left|\frac{d^2z}{dx^2}\right|\frac{1}{\sqrt{1+\left(\frac{dz}{dx}\right)^2}}\end{aligned} \tag{2.32}$$

so that, finally,

$$\chi = \frac{\left|\frac{d^2z}{dx^2}\right|}{\left(1+\left(\frac{dz}{dx}\right)^2\right)^{3/2}} . \tag{2.33}$$

For the meridian ellipse, one has from equations (2.15) and (2.16),

$$\frac{dz}{dx} = -\frac{b^2}{a^2}\frac{x}{z} = -\frac{\cos\phi}{\sin\phi} ; \tag{2.34}$$

and, the second derivative is obtained as follows (the details are left to the reader),

$$\frac{d^2z}{dx^2} = -\frac{b^2}{a^2}\frac{1}{z}\left(1+\frac{a^2}{b^2}\left(\frac{dz}{dx}\right)^2\right). \tag{2.35}$$

Making use of equations (2.19), (2.34), and (2.35), the curvature, equation (2.33), becomes

$$\begin{aligned}\chi &= \frac{\frac{b^2}{a^2}\frac{\sqrt{a^2\cos^2\phi+b^2\sin^2\phi}}{b^2\sin\phi}\frac{a^2\cos^2\phi+b^2\sin^2\phi}{b^2\sin^2\phi}}{\left(1+\frac{\cos^2\phi}{\sin^2\phi}\right)^{3/2}} \\ &= \frac{a}{b^2}\left(1-e^2\sin^2\phi\right)^{3/2}\end{aligned} \tag{2.36}$$

This is the curvature of the meridian ellipse; its reciprocal is the radius of curvature, denoted conventionally as M ,

$$M = \frac{a\left(1-e^2\right)}{\left(1-e^2\sin^2\phi\right)^{3/2}}, \tag{2.37}$$

where equation (2.5) is used. Note that M is a function of geodetic latitude (but not longitude, because of the rotational symmetry of the ellipsoid). Using the expression (2.27), the curvature of the meridian curve on the ellipsoid is

$$\frac{1}{M} = \left|\frac{d\phi}{ds}\right|, \tag{2.38}$$

since the slope angle of the ellipse is $90° - \phi$ (see Figure 2.6); and, hence, since $M > 0$ (always)

$$ds_{\text{meridian}} = M d\phi, \tag{2.39}$$

which is the differential element of arc along the meridian. The absolute value is removed with the convention that if $d\phi > 0$, then $ds > 0$; and, if $d\phi < 0$, then $ds < 0$.

The radius of curvature, M, is the principal normal to the meridian curve; and, therefore, it lies along the normal (perpendicular) to the ellipsoid (see Figure 2.8). At the pole ($\phi = 90°$) and at the equator ($\phi = 0°$) it assumes the following values, from equation (2.37):

$$\begin{aligned} M_{\text{equator}} &= a\left(1-e^2\right) < a \\ M_{\text{pole}} &= \frac{a}{\sqrt{1-e^2}} > a \end{aligned} \tag{2.40}$$

showing that M increases monotonically from equator to either pole, where it is maximum. Thus, also the curvature of the meridian decreases (becomes less curved) as one moves from the equator to the pole, which agrees with the fact that the ellipsoid is flattened at the poles. The length segment, M, does not intersect the polar axis, except at $\phi = 90°$. It happens that the "lower" endpoint of the radius falls on a curve as indicated in Figure 2.8. The values Δ_1 and Δ_2 are computed as follows

$$\begin{aligned} \Delta_1 &= a - M_{\text{equator}} = a - a\left(1-e^2\right) = ae^2 \\ \Delta_2 &= M_{\text{pole}} - b = \frac{a}{\frac{b}{a}} - b = be'^2 \end{aligned} \tag{2.41}$$

Using values for the ellipsoid of the Geodetic Reference System 1980, equation (2.11), these are

$$\begin{aligned} \Delta_1 &= 42697.67 \text{ m} \\ \Delta_2 &= 42841.31 \text{ m} \end{aligned} \tag{2.42}$$

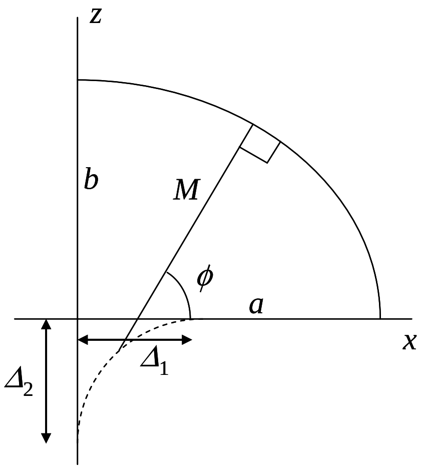


Figure 2.8: Meridian radius of curvature.

So far only the meridian curve has been considered. At a point on the ellipsoid, the curvature of any other curve through that point is generally different. In particular, imagine the class of curves that are generated as follows. At a point on the ellipsoid, let ξ be the unit vector defining the direction of the normal to the surface. By the symmetry of the ellipsoid, ξ lies in the meridian plane. Now consider any plane that contains ξ; it intersects the ellipsoid in a curve known as a *normal section* ("normal" because the plane contains the normal to the ellipsoid at a point) (see Figure 2.9). The meridian curve is a special case of a normal section; but the parallel circle is not a normal section; even though it is a plane curve, the plane that contains it does not contain the normal, ξ. A normal section on a sphere is a great circle; and the shortest spherical path between two points on the sphere is a great circle arc. However, as shown below, the shortest ellipsoidal path between two points on the ellipsoid is (usually) *not* a normal section.

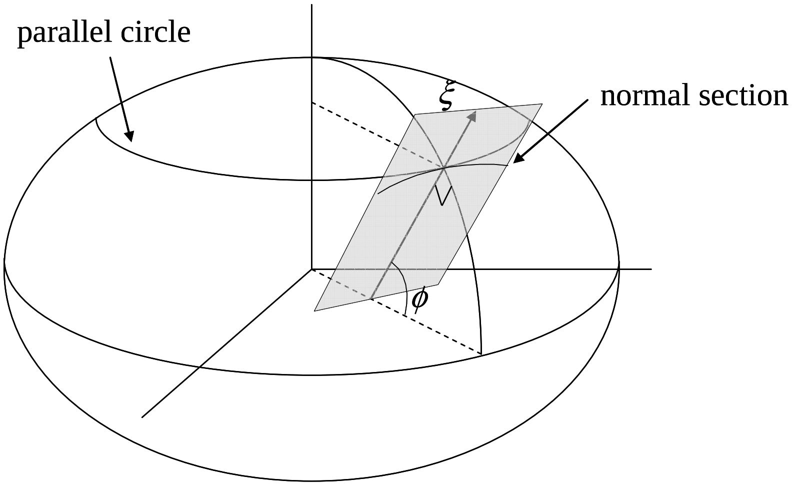


Figure 2.9: Normal section (shown for the prime vertical).

The normal section drawn in Figure 2.9, is another special case; it is the *prime vertical normal section* – it is perpendicular to the meridian. Note that while the prime vertical normal section and the parallel circle have the same tangent where they meet, they have different principal normals. The principal normal of the parallel circle (its radius of curvature) is parallel to the equator, while the principal normal of the prime vertical normal section (or any normal section) is the normal to the ellipsoid – but at this point only!

In differential geometry, there is the following theorem due to *Meusnier* (e.g., McConnell, 1957)

Theorem: *For all surface curves, C, with the same tangent vector at a point, each having curvature, χ_C, at that point, and the principal normal of each making an angle, θ_C, with the normal to the surface, there is*

$$\chi_C \cos\theta_C = \text{constant} . \tag{2.43}$$

$\chi_C \cos\theta_C$ is called the *normal curvature* of the curve, C, at the point of tangency.

Applying this theorem to the ellipsoid, consider the set of all curves that share the same tangent at a point as the prime vertical normal section. For the prime vertical normal section, one clearly has, $\theta_C = 0°$, since its principal normal is also the normal to the ellipsoid at that point. Hence, the constant in equation (2.43) for this set of curves is

$$\text{constant} = \chi_{\text{prime vertical normal section}} . \tag{2.44}$$

The constant is the curvature of that normal section at the point. Defining

$$\chi_{\text{prime vertical normal section}} = \frac{1}{N}, \tag{2.45}$$

N is the radius of curvature of the prime vertical normal section at the point on the ellipsoid where this radius is normal to the ellipsoid. The parallel circle through that point has the same tangent as the prime vertical normal section, and its radius of curvature is $p = 1 / \chi_{\text{parallel circle}}$. The angle of its principal normal, *p*, with respect to the ellipsoid normal is the geodetic latitude, ϕ (Figure 2.6). Hence, from equations (2.43) - (2.45):

$$\frac{1}{p}\cos\phi = \frac{1}{N}, \tag{2.46}$$

which implies that

$$p = N\cos\phi, \tag{2.47}$$

and that N is the length of the normal to the ellipsoid from the point on the ellipsoid to its minor axis (see Figure 2.10).

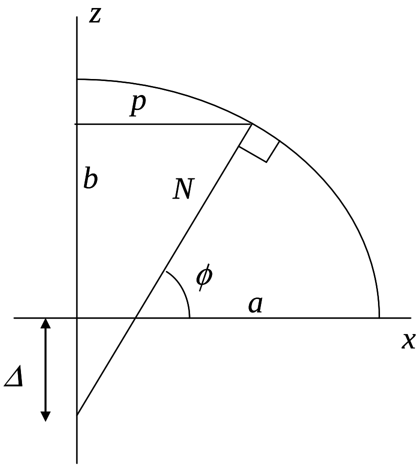


Figure 2.10: Prime vertical radius of curvature.

The x-coordinate of a point on the ellipsoid whose y-coordinate is zero is given by equation (2.21); but this is also p. Hence, from equation (2.47)

$$N = \frac{a}{\sqrt{1 - e^2 \sin^2\phi}}. \tag{2.48}$$

From Figure 2.10 and equation (2.20), the point of intersection of N with the minor axis is the distance from the ellipsoid center given by

$$\mathit{\Delta} = N \sin\phi - z = Ne^2 \sin\phi . \tag{2.49}$$

At the equator ($\phi = 0°$) and at the poles ($\phi = \pm 90°$), the prime vertical radius of curvature assumes the following constants, according to equation (2.48):

$$\begin{aligned} N_{\text{equator}} &= a \\ N_{\text{pole}} &= \frac{a}{\sqrt{1-e^2}} > a \end{aligned} \tag{2.50}$$

and we see that N increase monotonically from the equator to either pole, where it is maximum. Note that at the pole,

$$N_{\text{pole}} = M_{\text{pole}} , \tag{2.51}$$

since all normal sections at the pole are meridians. Again, the increase in N toward the poles implies a decrease in curvature (due to the flattening of the ellipsoid). Finally, $N_{\text{equator}} = a$ agrees with the fact that the equator, being the prime vertical normal section for points on the equator, is a circle with radius, a.

Making use of the basic definition of curvature as the absolute change in slope angle with respect to arc length of the curve, equation (2.27), we find for the parallel circle,

$$\frac{1}{p} = \left| \frac{d\lambda}{ds} \right| ; \tag{2.52}$$

and, therefore, again removing the absolute value with the convention that if $d\lambda < 0$ ($d\lambda > 0$), then also $ds < 0$ ($ds > 0$), there is,

$$ds_{\text{parallel circle}} = N \cos\phi \, d\lambda = ds_{\text{prime vertical normal section}} , \tag{2.53}$$

where the second equality holds only where the parallel circle and the prime vertical normal section are tangent.

From equations (2.37) and (2.48), it is easily verified that, always,

$$M \le N . \tag{2.54}$$

Also, at any point M and N are, respectively, the minimum and maximum radii of curvature for all normal sections through that point. M and N are known as the *principal radii of curvature* at a point of the ellipsoid. For any arbitrary curve, the differential element of arc, using equations (2.39) and (2.53), is given by

$$ds = \sqrt{M^2 d\phi^2 + N^2 \cos^2\phi \, d\lambda^2} \; . \qquad (2.55)$$

To determine the curvature of an arbitrary normal section firstly requires a definition of the *direction* of the normal section. The *normal section azimuth,* α, is the angle measured in the plane tangent to the ellipsoid at a point, clockwise about the normal to that point, from the (northward) meridian plane to the plane of the normal section. *Euler's formula* gives the curvature of the normal section having normal section azimuth, α, in terms of the principal radii of curvature:

$$\chi_\alpha = \frac{1}{R_\alpha} = \frac{\sin^2\alpha}{N} + \frac{\cos^2\alpha}{M} \, . \qquad (2.56)$$

The radius of curvature, R_α, of the normal section in azimuth, α, can be used to define a *mean* local radius of the ellipsoid. This is appropriate (although rarely used) if locally one wishes to approximate the ellipsoid by a sphere – this radius is the radius of the locally approximating sphere. For example, one type of mean local radius is the *Gaussian mean radius*, which is the average of the radii of curvature of all normal sections at a point:

$$\begin{aligned} R_G &= \frac{1}{2\pi}\int_0^{2\pi} R_\alpha d\alpha = \frac{1}{2\pi}\int_0^{2\pi} \frac{d\alpha}{\frac{\sin^2\alpha}{N} + \frac{\cos^2\alpha}{M}} \\ &= \sqrt{MN} = \frac{a\left(1-f\right)}{1-e^2\sin^2\phi} \end{aligned} \qquad (2.57)$$

as shown in (Rapp, 1991, p.44; see also Problem 2.1.3.4.-1.). Note that the Gaussian mean radius is a function of latitude. Another approximating radius is the *mean radius of curvature*, defined from the average of the principal curvatures:

$$R_m = \frac{1}{\frac{1}{2}\left(\frac{1}{N} + \frac{1}{M}\right)} \, . \qquad (2.58)$$

For the sake of completeness, other mean radii are defined here that approximate the ellipsoid globally rather than locally. One is the average of the semi-axes of the ellipsoid,

$$R = \frac{1}{3}(a + a + b)\,; \qquad (2.59)$$

another is the radius of the sphere whose surface area equals that of the ellipsoid,

$$R_A = \sqrt{\frac{\Sigma}{4\pi}}\,, \qquad (2.60)$$

where Σ is the area of the ellipsoid, given by (Rapp, 1991, p.42; see also Problem 2.1.3.4.-4.)

$$\Sigma = 2\pi b^2 \left(\frac{1}{1-e^2} + \frac{1}{2e} \ln \frac{1-e}{1+e} \right); \qquad (2.61)$$

and, a third is the radius of the sphere whose volume equals that of the ellipsoid,

$$R_V = \left(\frac{3}{4\pi} V \right)^{1/3}, \qquad (2.62)$$

where V is the volume of the ellipsoid, given by

$$V = \frac{4}{3}\pi a^2 b\,. \qquad (2.63)$$

Using the values of GRS80, all of these approximations imply

$$R = 6371 \text{ km}\,, \qquad (2.64)$$

as the *mean Earth radius*, to the nearest km.

2.1.3.2 Normal Section Azimuth

Consider again a normal section defined at a point, A, and passing through a target point, B; see Figure 2.11. We note that the points n_A and n_B, being the intersections with the minor axis of the normals through A and B, respectively, do not coincide (unless, $\phi_A = \phi_B$). Therefore, the normal plane at A that also contains the point B, while it contains the normal at A, does not contain the normal at B. And, vice versa! Therefore, unless $\phi_A = \phi_B$, the normal section at A through B is not the same as the normal section at B through A. In addition, the normal section at A through a different target point, B', along the normal at B, but at height $h_{B'}$, is different than the normal section through B (Figure 2.12). Note that in Figure 2.12, ABn_A and $AB'n_A$ define two different planes containing the normal at A.

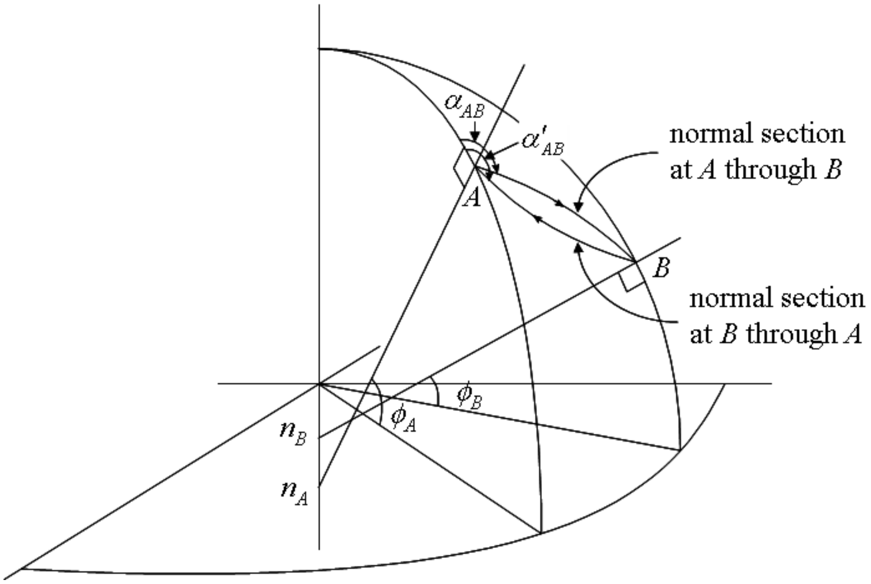
Both of these geometries (Figures 2.11 and 2.12) affect the definition of the azimuth at A of the (projection of the) target point, B. If α_{AB} is the normal section azimuth of B at A, and α'_{AB} is the azimuth, at A, of the "reverse" normal section coming from B through A, then the difference between these azimuths is given by Rapp (1991, p.59),

$$\alpha_{AB} - \alpha'_{AB} \simeq \frac{e^2}{2}\sin\alpha_{AB}\left(\frac{s}{N_A}\right)^2\cos^2\phi_A\left(\cos\alpha_{AB} - \frac{1}{2}\frac{s}{N_A}\tan\phi_A\right), \tag{2.65}$$

where s is the length of the normal section. This is an approximation where higher powers of s/N_A are neglected. Furthermore, if $\alpha_{AB'}$ is the normal section azimuth of B' at A, where B' is at a height, $h_{B'}$, along the ellipsoid normal at *B*, then Rapp (1991, p.63) gives the difference,

$$\alpha_{AB} - \alpha_{AB'} \simeq \frac{h_{B'}}{N_A}e'^2\cos^2\phi_A\sin\alpha_{AB}\left(\cos\alpha_{AB} - \frac{1}{2}\frac{s}{N_A}\tan\phi_A\right). \tag{2.66}$$

The latter difference is independent of the height of the point *A* (the reader should understand why!).



α_{AB}
α'_{AB}
normal section at *A* through *B*
A
B
normal section at *B* through *A*
ϕ_B
ϕ_A
n_B
n_A

Figure 2.11: Normal sections at A and B.

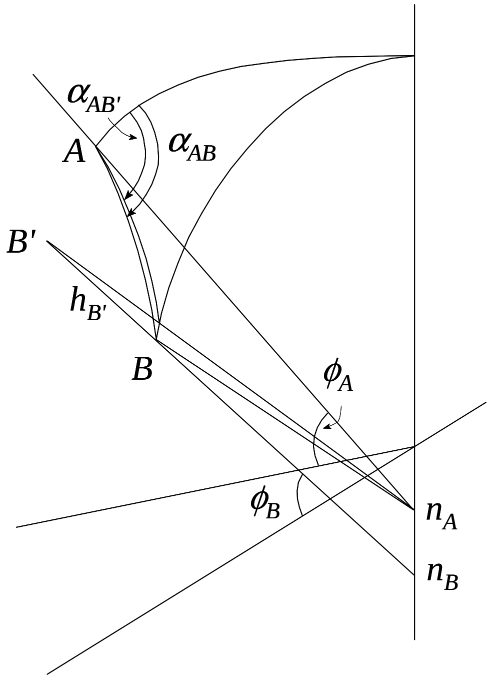


Figure 2.12: Normal sections for target points at different heights.

2.1.3.3 Geodesics

Consider the following problem: given two points on the surface of the ellipsoid, find the curve on the ellipsoid connecting these two points and having the shortest length. This curve is known as the *geodesic* (curve). Geodesics on a sphere are great circle arcs and these are plane curves; but, as already mentioned, on the ellipsoid, geodesics have double curvature – they are not plane curves and their geometry is more complicated. We will find the conditions that must be satisfied by geodetic coordinates of points on a geodesic. The problem can be solved using the *calculus of variations*, as follows.

Let ds be the differential element of arc of an *arbitrary* curve on the ellipsoid. In terms of differential latitude and longitude, this element is given by equation (2.55), repeated here for convenience,

$$ds = \sqrt{M^2 d\phi^2 + N^2 \cos^2 \phi \, d\lambda^2} \; . \tag{2.67}$$

If α is the azimuth of the curve at a point then the element of arc at that point may also be decomposed according to the latitudinal and longitudinal elements using equations (2.39) and (2.53):

$$\begin{aligned} ds\cos\alpha &= Md\phi \\ ds\sin\alpha &= N\cos\phi\, d\lambda \end{aligned} \tag{2.68}$$

Let I denote the length of a curve between two points, P and Q, on the ellipsoid. The geodesic between these two points is the path, s, that satisfies the condition:

$$I = \int_P^Q ds \rightarrow \min . \tag{2.69}$$

The problem of finding the equation of the curve under the condition (2.69) can be solved by the method of the calculus of variations. This method has many applications in mathematical physics and general procedures may be formulated. In particular, consider the more general problem of minimizing the integral of some function, $F\left(x, y(x), y'(x)\right)$, where y' is the derivative of y with respect to x:

$$I = \int F dx \rightarrow \min . \tag{2.70}$$

It can be shown (Arfken 1970) that the integral in equation (2.70) is minimized *if and only if* the following differential equation holds

$$\frac{d}{dx}\frac{\partial F}{\partial y'} - \frac{\partial F}{\partial y} = 0 . \tag{2.71}$$

This is *Euler's equation*. Note that both total and partial derivatives are used in equation (2.71). It is an equation for $y(x)$. A solution to this equation (in essence, by integration) provides the necessary and sufficient conditions on $y(x)$ that minimize the integral (2.70).

In our case, by comparing equations (2.69) and (2.70), we have

$$Fdx = ds ; \tag{2.72}$$

and, the points on an arbitrary curve on the ellipsoid are identified by

$$\phi = \phi(\lambda) . \tag{2.73}$$

That is, λ is chosen to be the independent variable of the functional description of the curve (i.e., $y \equiv \phi$ and $x \equiv \lambda$ in the more general formulation above). From equation (2.67),

$$ds = \sqrt{M^2 d\phi^2 + N^2 \cos^2\phi\, d\lambda^2} = \sqrt{M^2\left(\frac{d\phi}{d\lambda}\right)^2 + \left(N\cos\phi\right)^2}\; d\lambda\,; \tag{2.74}$$

so that

$$F = \sqrt{M^2\left(\frac{d\phi}{d\lambda}\right)^2 + \left(N\cos\phi\right)^2} = F\left(\phi, \phi'\right), \tag{2.75}$$

where $\phi' = d\phi/d\lambda$.

Immediately, it is seen that in this case F does not depend on λ explicitly, i.e.,

$$\frac{\partial F}{\partial \lambda} = 0\,. \tag{2.76}$$

Now let F be that function that minimizes the path length; that is, F must satisfy Euler's equation. A first integral of Euler's equation (2.71) can be obtained from equation (2.76); it will be shown that it is given by

$$F - \phi'\frac{\partial F}{\partial \phi'} = \text{constant}\,. \tag{2.77}$$

To prove this, we work backwards. That is, starting with equation (2.77), we obtain something we know to be true, and in the end we argue that our steps of reasoning can be reversed to get equation (2.77). Thus, differentiate equation (2.77) with respect to λ:

$$\frac{d}{d\lambda}\left(F - \phi'\frac{\partial F}{\partial \phi'}\right) = 0\,. \tag{2.78}$$

Explicitly, the derivative is

$$\frac{dF}{d\lambda} - \phi''\frac{\partial F}{\partial \phi'} - \phi'\frac{d}{d\lambda}\frac{\partial F}{\partial \phi'} = 0\,. \tag{2.79}$$

Now, by the chain rule applied to $F\left(\lambda, \phi(\lambda), \phi'(\lambda)\right)$, we get

$$\begin{aligned}\frac{dF}{d\lambda} &= \frac{\partial F}{\partial \lambda} + \frac{\partial F}{\partial \phi}\phi' + \frac{\partial F}{\partial \phi'}\phi'' \\ &= \frac{\partial F}{\partial \phi}\phi' + \frac{\partial F}{\partial \phi'}\phi''\end{aligned} \tag{2.80}$$

because of equation (2.76). Substituting equation (2.80) into equation (2.79) yields

$$\phi'\left(\frac{\partial F}{\partial \phi} - \frac{d}{d\lambda}\frac{\partial F}{\partial \phi'}\right) = 0\,. \tag{2.81}$$

Since, in general, $\phi' \neq 0$, we must have

$$\frac{\partial F}{\partial \phi} - \frac{d}{d\lambda}\frac{\partial F}{\partial \phi'} = 0\,. \tag{2.82}$$

But this is Euler's equation, assumed to hold for our particular F. That is, the F defined by equation (2.77) also satisfies Euler's equation. The process can be reversed to get equation (2.77) from equation (2.82); therefore, equations (2.77) and (2.82) are equivalent in this case and equation (2.77) is a first integral of Euler's equation (it has now been reduced to a *first-order* differential equation).

From equation (2.75), it follows that

$$\frac{\partial F}{\partial \phi'} = \frac{M^2\phi'}{\sqrt{M^2\phi'^2 + \left(N\cos\phi\right)^2}}\,. \tag{2.83}$$

Substituting this into equation (2.77) yields

$$\begin{aligned}F - \phi'\frac{\partial F}{\partial \phi'} &= \sqrt{M^2\phi'^2 + \left(N\cos\phi\right)^2} - \frac{M^2\phi'^2}{\sqrt{M^2\phi'^2 + \left(N\cos\phi\right)^2}} \\ &= \frac{\left(N\cos\phi\right)^2}{\sqrt{M^2\phi'^2 + \left(N\cos\phi\right)^2}} = \text{constant}\end{aligned} \tag{2.84}$$

The last equation is the condition on $\phi(\lambda)$ that must be satisfied for points having coordinates (ϕ, λ) that are on the geodesic.

The derivative, ϕ', can be obtained by dividing the two equations (2.68):

$$\frac{d\phi}{d\lambda} = \frac{N\cos\phi}{M}\cot\alpha \,. \tag{2.85}$$

Substituting this derivative which holds for an arbitrary curve into the condition (2.84) which holds only for geodesics, we get

$$\frac{\left(N\cos\phi\right)^2}{\sqrt{M^2\left(\frac{N\cos\phi}{M}\cot\alpha\right)^2 + \left(N\cos\phi\right)^2}} = \frac{N\cos\phi}{\sqrt{\cot^2\alpha+1}} = \text{constant} \,. \tag{2.86}$$

The last equality can be simplified to

$$N\cos\phi\sin\alpha = \text{constant} \,. \tag{2.87}$$

This is the famous equation known as *Clairaut's equation*. All points on a geodesic must satisfy this equation. That is, if C is a geodesic curve on the ellipsoid, where ϕ is the geodetic latitude of an arbitrary point on C, and α is the azimuth of the geodesic at that point (i.e., the angle with respect to the meridian of the tangent to the geodesic at that point), then ϕ and α are related according to equation (2.87). Note that Clairaut's equation by itself is only a necessary condition, not a sufficient condition, for a curve to be a geodesic; that is, if points on a curve satisfy equation (2.87), then this is no guarantee that the curve is a geodesic (e.g., consider an arbitrary parallel circle). However, Clairaut's equation combined with the condition, $\phi' \neq 0$, is sufficient to ensure that the curve is a geodesic. This can be proved by reversing the arguments of equations (2.77) – (2.87) (see Problem 2.1.3.4-8).

From equations (2.47) and (2.14), specialized to $u = b$, we find

$$\begin{aligned} p &= N\cos\phi \\ &= a\cos\beta \end{aligned} \tag{2.88}$$

which gives another form of Clairaut's equation:

$$\cos\beta\sin\alpha = \text{constant} \,. \tag{2.89}$$

Therefore, for points on a geodesic, the product of the cosine of the reduced latitude and the sine of the azimuth is always the same value. The same equation holds for great circles on the sphere, where, of course, the reduced latitude becomes the geocentric latitude.

Substituting equation (2.88) into equation (2.87) gives

$$p\sin\alpha = \text{constant} \,. \tag{2.90}$$

Taking differentials leads to

$$\sin\alpha\, dp + p\cos\alpha\, d\alpha = 0\,. \tag{2.91}$$

With equations (2.88) and (2.68), equation (2.91) may be expressed as

$$d\alpha = -\frac{dp}{\cos\alpha\, ds} d\lambda\,. \tag{2.92}$$

Again, using equation (2.68), this is the same as

$$d\alpha = -\frac{dp}{M\, d\phi} d\lambda\,. \tag{2.93}$$

It can be shown, from equations (2.37) and (2.48), that

$$\frac{dp}{d\phi} = \frac{d}{d\phi}\left(N\cos\phi\right) = -M\sin\phi\,. \tag{2.94}$$

Putting this into equation (2.93) yields another famous equation, *Bessel's equation*:

$$d\alpha = \sin\phi\, d\lambda\,. \tag{2.95}$$

This also holds only for points on the geodesic; it is both a necessary and a sufficient condition for a curve to be a geodesic under the same restriction as before. That is, the arguments leading to equation (2.95) can be reversed to show that the consequence of equation (2.95) is equation (2.87), provided $\phi' \neq 0$ (or, $\cos\alpha \neq 0$), thus proving sufficiency.

Geodesics on the ellipsoid have a rich geometry that we cannot begin to explore in this text. The interested reader is referred to Rapp (1992) and Thomas (1970). However, it is worth mentioning, without proof, some of the interesting facts of geodesics on the ellipsoid.

1) Any meridian is a geodesic.

2) The equator is a geodesic up to a point; that is, the shortest distance between two points on the equator is along the equator, but not always. Clearly, for two diametrically opposite points on the equator, the shortest distance is along the meridian (because of the flattening of the ellipsoid). So, starting from a given point on the equator, the equator serves as the geodesic to another point on the equator, if that point is not too close to the "antipode," depending on the flattening. At some critical point near the antipode (and for points beyond that), the geodesic from the starting point jumps off the equator and runs along the ellipsoid with varying latitude, until for the antipode, itself, the meridian is the geodesic.

3) Except for the equator, no other parallel circle is a geodesic (see Problem 2.1.3.4-7.).

4) In general, a geodesic on the ellipsoid is not a plane curve; that is, it is not generated by the intersection of a plane with the ellipsoid. The geodesic has double curvature, or torsion.

5) It can be shown that the principal normal of the geodesic curve is also the normal to the ellipsoid at *each* point of the geodesic (for the normal section, the principal normal coincides with the normal to the ellipsoid only at the point where the normal is in the plane of the normal section).

6) Following a continuous geodesic curve on the ellipsoid, it reaches maximum and minimum latitudes, $\phi_{\max} = -\phi_{\min}$, like a great circle on a sphere, but it does not repeat itself on circumscribing the ellipsoid (like the great circle does), which is a consequence of its not being a plane curve; the meridians and equator are the only exceptions to this.

7) Rapp (1991, p.84) gives the following approximate formula for the difference between the normal section azimuth and the geodesic azimuth, $\tilde{\alpha}_{AB}$ (see Figure 2.13):

$$\begin{aligned} \alpha_{AB} - \tilde{\alpha}_{AB} &\simeq \frac{e'^2}{6} \sin\alpha_{AB} \left(\frac{s}{N_A}\right)^2 \cos^2\phi_A \left(\cos\alpha_{AB} - \frac{1}{4}\frac{s}{N_A}\tan\phi_A\right) \\ &\simeq \frac{1}{3}\left(\alpha_{AB} - \alpha'_{AB}\right) \end{aligned} \tag{2.96}$$

where the second approximation neglects the second term within the parentheses.

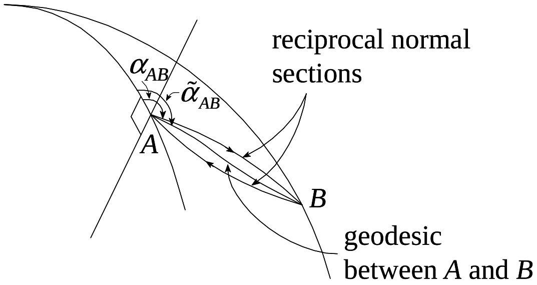


Figure 2.13: Normal sections versus geodesic on the ellipsoid.

2.1.3.4 Problems

1. Split the integral in equation (2.57) into four integrals, one over each quadrant, and consult a Table of Integrals to prove the result.

2. Show that the length of a parallel circle arc between longitudes λ_1 and λ_2 is given by

$$L = \left(\lambda_2 - \lambda_1\right) N \cos\phi . \qquad (2.97)$$

3. Find an expression for the length of a meridian arc between geodetic latitudes ϕ_1 and ϕ_2. Can the integral be solved analytically?

4. Show that the area of the ellipsoid surface between longitudes λ_1 and λ_2 and geodetic latitudes ϕ_1 and ϕ_2 is given by

$$\Sigma\left(\phi_1, \phi_2, \lambda_1, \lambda_2\right) = b^2 \left(\lambda_2 - \lambda_1\right) \int\limits_{\phi_1}^{\phi_2} \frac{\cos\phi}{\left(1 - e^2 \sin^2\phi\right)} d\phi . \qquad (2.98)$$

Then consult a Table of Integrals to show that this reduces to

$$\Sigma\left(\phi_1, \phi_2, \lambda_1, \lambda_2\right) = \frac{b^2}{2} \left(\lambda_2 - \lambda_1\right) \left(\frac{\sin\phi}{1 - e^2 \sin^2\phi} + \frac{1}{2e} \ln \frac{1 + e\sin\phi}{1 - e\sin\phi} \right)_{\phi_1}^{\phi_2} , \qquad (2.99)$$

(where e is the first eccentricity, not the exponential). Finally, prove equation (2.61).

5. Consider two points, A and B, that are on the same parallel circle.

a) What should be the differences, $\alpha_{AB} - \alpha'_{AB}$ and $\alpha_{AB} - \alpha_{AB'}$, given by equations (2.65) and (2.66), and why?

b) Show that in spherical approximation the parenthetical term in equations (2.65) and (2.66) is approximately zero if the two points, *A* and *B*, are on the same parallel, and if the distance *s* is not large (hint: use the law of cosines on spherical triangle *ABO*, where *O* is the north pole, to show that approximately

$$\sin\phi_A \simeq \sin\phi_A \cos\frac{s}{N_A} + \cos\phi_A \sin\frac{s}{N_A} \cos\alpha_{AB} .$$

Then solve for $\cos\alpha_{AB}$ and use small-angle approximations to second order for the sine and cosine).

6. Suppose that a geodesic curve on the ellipsoid attains a maximum geodetic latitude, $\phi_{\max}$. Show that the azimuth of the geodesic as it crosses the equator is given by

$$\alpha_{\text{equator}} = \sin^{-1}\left(\frac{\cos\phi_{\max}}{\sqrt{1-e^2\sin^2\phi_{\max}}}\right). \tag{2.100}$$

7. Using Bessel’s equation show that a parallel circle arc (except the equator) can not be a geodesic.

8. Prove that if $\phi' \neq 0$ then equation (2.87) is a sufficient condition for a curve to be a geodesic, i.e., equations (2.77) and hence (2.69) are satisfied. That is, if all points on a curve satisfy equation (2.87) , the curve must be a geodesic.

2.1.4 Direct / Inverse Problems

There are two essential problems in the computation of coordinates, directions, and distances on a particular given ellipsoid (see Figure 2.14):

The Direct Problem: Given the geodetic coordinates of a point on the ellipsoid, the azimuth to a second point, and the geodesic distance between the points, find the geodetic coordinates of the second point, as well as the back-azimuth (azimuth of the first point at the second point), where all azimuths are *geodesic* azimuths. That is,

$$\text{given: } \phi_1, \lambda_1, \alpha_1, s_{12} \text{; find: } \phi_2, \lambda_2, \bar{\alpha}_2 .$$

The Inverse Problem: Given the geodetic coordinates of two points on the ellipsoid, find the geodesic forward- and back-azimuths (geodesic azimuths), as well as the geodesic distance between the points. That is,

$$\text{given: } \phi_1, \lambda_1, \phi_2, \lambda_2 \text{; find: } \alpha_1, \bar{\alpha}_2, s_{12} .$$

The solutions to these problems form the basis for relating traditional geodetic observations of angles and distances to the establishment of a horizontal control network of point coordinates for a region. That is, they provide for the solution of general ellipsoidal triangles (Ehlert 1993), analogous to the relatively simple solutions of spherical triangles, which constitute the elements of a triangulation network on the mapping surface, the ellipsoid. There are many solutions that hold for short lines (generally less than 100 – 200 km) and are based on some kind of approximation; in fact, one solution to the problem is developed by approximating the ellipsoid locally by a sphere. None of these developments is simpler in essence than the exact (iterative, or series) solution which holds for any length of line. The latter solutions are fully developed in (Rapp, 1992). However, we will consider only one of the approximate solutions in order to obtain some tools for simple applications. In fact, today with GPS virtually replacing the traditional distance and angle measurements for geodetic control, the direct problem is essentially irrelevant in geodesy. The indirect problem is still quite useful as applied to long-range surface navigation and guidance (e.g., for oceanic commercial navigation).

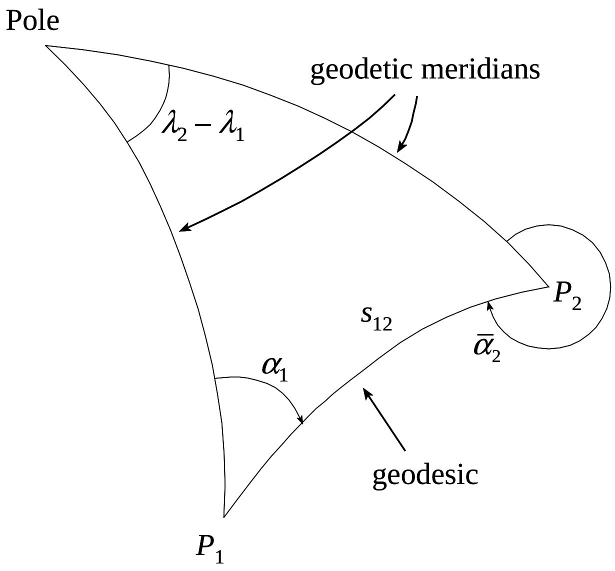


Figure 2.14: Ellipsoidal geometry for direct and inverse geodetic problems.

One set of solutions of these problems is the Legendre-series solution, first developed by Legendre and published in the *Mémoires* of the Paris Academy in 1806 (Jordan 1962). The geodesic may be parameterized by the arc length, s, so that both coordinates and the forward geodesic azimuth for points along the geodesic are functions of s,

$$\phi = \phi(s), \quad \lambda = \lambda(s), \quad \alpha = \alpha(s). \tag{2.101}$$

Let $\bar{\alpha}$ denote the back-azimuth, so that $\bar{\alpha} = \alpha + \pi$. Then, a Taylor series expansion formally yields:

$$\phi_2 = \phi_1 + \left.\frac{d\phi}{ds}\right|_1 s_{12} + \frac{1}{2!}\left.\frac{d^2\phi}{ds^2}\right|_1 s_{12}^2 + \cdots; \tag{2.102}$$

$$\lambda_2 = \lambda_1 + \left.\frac{d\lambda}{ds}\right|_1 s_{12} + \frac{1}{2!}\left.\frac{d^2\lambda}{ds^2}\right|_1 s_{12}^2 + \cdots; \tag{2.103}$$

$$\bar{\alpha}_2 = \alpha_1 + \pi + \left.\frac{d\alpha}{ds}\right|_1 s_{12} + \frac{1}{2!}\left.\frac{d^2\alpha}{ds^2}\right|_1 s_{12}^2 + \cdots; \tag{2.104}$$

where s_{12} is the geodesic distance from $P_1 = (\phi_1, \lambda_1)$ to $P_2 = (\phi_2, \lambda_2)$. The derivatives in each case are obtained from the differential elements of a geodesic and evaluated at the starting point,

P_1. The convergence of the series is not guaranteed for all s_{12}, but it is expected for $s_{12} \ll R$ (mean radius of the Earth), although the convergence may be slow.

We recall the equations (2.68):

$$\begin{aligned} ds\cos\alpha &= Md\phi \\ ds\sin\alpha &= N\cos\phi\, d\lambda \end{aligned} \tag{2.105}$$

which hold for any curve on the ellipsoid; and Bessel's equation (2.95):

$$d\alpha = \sin\phi\, d\lambda\,, \tag{2.106}$$

which holds only for geodesics. Thus, from equation (2.105),

$$\left.\frac{d\phi}{ds}\right|_1 = \frac{\cos\alpha_1}{M_1}\,, \tag{2.107}$$

and

$$\left.\frac{d\lambda}{ds}\right|_1 = \frac{\sin\alpha_1}{N_1\cos\phi_1}\,. \tag{2.108}$$

Substituting $d\lambda$, given by equation (2.105), into equation (2.106), we find

$$\left.\frac{d\alpha}{ds}\right|_1 = \frac{\sin\alpha_1}{N_1}\tan\phi_1\,. \tag{2.109}$$

This completes the determination of the first derivatives.

For the second derivatives, it can be shown that (derivations are left to the reader):

$$\frac{dM}{d\phi} = \frac{3MN^2e^2\sin\phi\cos\phi}{a^2}\,; \tag{2.110}$$

$$\frac{dN}{d\phi} = Me'^2\sin\phi\cos\phi\,; \tag{2.111}$$

$$\frac{d}{d\phi}\left(N\cos\phi\right) = -M\sin\phi\,. \tag{2.112}$$

Using the chain rule of standard calculus, we have

$$\frac{d^2\phi}{ds^2}=\frac{d}{ds}\left(\frac{\cos\alpha}{M}\right)=-\frac{1}{M}\sin\alpha\frac{d\alpha}{ds}-\frac{\cos\alpha}{M^2}\frac{dM}{d\phi}\frac{d\phi}{ds}\,, \tag{2.113}$$

which becomes, upon substituting equations (2.107), (2.109), and (2.110):

$$\left.\frac{d^2\phi}{ds^2}\right|_1=-\frac{\sin^2\alpha_1}{M_1N_1}\tan\phi_1-\frac{3e^2N_1^2\cos^2\alpha_1\sin\phi_1\cos\phi_1}{a^2M_1^2}\,. \tag{2.114}$$

Similarly, for the longitude,

$$\frac{d^2\lambda}{ds^2}=\frac{d}{ds}\left(\frac{\sin\alpha}{N\cos\phi}\right)=\frac{\cos\alpha}{N\cos\phi}\frac{d\alpha}{ds}-\frac{\sin\alpha}{N^2\cos^2\phi}\frac{d}{d\phi}\left(N\cos\phi\right)\frac{d\phi}{ds}\,, \tag{2.115}$$

which, with appropriate substitutions as above, leads after simplification (left to the reader) to

$$\left.\frac{d^2\lambda}{ds^2}\right|_1=\frac{2\sin\alpha_1\cos\alpha_1}{N_1^2\cos\phi_1}\tan\phi_1\,. \tag{2.116}$$

Finally, for the azimuth,

$$\frac{d^2\alpha}{ds^2}=\frac{d}{ds}\left(\frac{\sin\alpha}{N}\tan\phi\right)=\frac{\cos\alpha}{N}\tan\phi\frac{d\alpha}{ds}-\frac{\sin\alpha}{N^2}\tan\phi\frac{dN}{d\phi}\frac{d\phi}{ds}+\frac{\sin\alpha}{N}\sec^2\phi\frac{d\phi}{ds}\,, \tag{2.117}$$

that with the substitutions for the derivatives as before and after considerable simplification (left to the reader) yields

$$\left.\frac{d^2\alpha}{ds^2}\right|_1=\frac{\sin\alpha_1\cos\alpha_1}{N_1^2}\left(1+2\tan^2\phi_1+e'^2\cos^2\phi_1\right). \tag{2.118}$$

Clearly, higher-order derivatives become more complicated, but could be derived by the same procedures. Expressions up to fifth order, also given below, are found in (Jordan, 1962) and (Rapp, 1991).

With the following abbreviations

$$v=\frac{s_{12}}{N_1}\sin\alpha_1,\quad u=\frac{s_{12}}{N_1}\cos\alpha_1,\quad \eta^2=e'^2\cos^2\phi_1,\quad t=\tan\phi_1\,, \tag{2.119}$$

the final solution to the direct problem up to fifth order in s_{12}/N_1 is thus given as follows, the details of which are left to the reader (see also Problem 3, Section 2.1.41).

$$\begin{aligned}\frac{\phi_2-\phi_1}{1+\eta^2} &= u-\frac{1}{2}v^2t-\frac{3}{2}u^2\eta^2t-\frac{v^2u}{6}\left(1+3t^2+\eta^2-9\eta^2t^2\right)-\frac{u^3}{2}\eta^2\left(1-t^2\right)\\ &\quad+\frac{v^4}{24}t\left(1+3t^2+\eta^2-9\eta^2t^2\right)-\frac{v^2u^2}{12}t\left(4+6t^2-13\eta^2-9\eta^2t^2\right)+\frac{u^4}{2}\eta^2t\\ &\quad+\frac{v^4u}{120}\left(1+30t^2+45t^4\right)-\frac{v^2u^3}{30}\left(2+15t^2+15t^4\right)\end{aligned} \tag{2.120}$$

$$\begin{aligned}\left(\lambda_2-\lambda_1\right)\cos\phi_1 &= v+uvt-\frac{v^3}{3}t^2+\frac{vu^2}{3}\left(1+3t^2+\eta^2\right)\\ &\quad-\frac{v^3u}{3}t\left(1+3t^2+\eta^2\right)+\frac{vu^3}{3}t\left(2+3t^2+\eta^2\right)\\ &\quad+\frac{v^5}{15}t^2\left(1+3t^2\right)+\frac{vu^4}{15}\left(2+15t^2+15t^4\right)-\frac{v^3u^2}{15}\left(1+20t^2+30t^4\right)\end{aligned} \tag{2.121}$$

$$\begin{aligned}\bar{\alpha}_2-\left(\alpha_1+\pi\right) &= vt+\frac{vu}{2}\left(1+2t^2+\eta^2\right)-\frac{v^3}{6}t\left(1+2t^2+\eta^2\right)+\frac{vu^2}{6}t\left(5+6t^2+\eta^2-4\eta^4\right)\\ &\quad-\frac{v^3u}{24}\left(1+20t^2+24t^4+2\eta^2+8\eta^2t^2\right)+\frac{vu^3}{24}\left(5+28t^2+24t^4+6\eta^2+8\eta^2t^2\right) \quad .\\ &\quad+\frac{v^5}{120}t\left(1+20t^2+24t^4\right)-\frac{v^3u^2}{120}t\left(58+280t^2+240t^4\right)+\frac{vu^4}{120}t\left(61+180t^2+120t^4\right)\end{aligned} \tag{2.122}$$

The inverse solution can be obtained from these series by *iteration*. Equations (2.120) and (2.121) are written as

$$\Delta\phi=\phi_2-\phi_1=\left(1+\eta^2\right)u+\delta\phi\,, \tag{2.123}$$

$$\Delta\lambda=\lambda_2-\lambda_1=\frac{v}{\cos\phi_1}+\delta\lambda\,, \tag{2.124}$$

where $\delta\phi$ and $\delta\lambda$ are the residuals with respect to the first-order terms. Now, solving for u and v, we have

$$u = \frac{\Delta\phi - \delta\phi}{1+\eta^2}, \quad v = \cos\phi_1 \left(\Delta\lambda - \delta\lambda\right); \tag{2.125}$$

and, with equation (2.119), the equation for the forward-azimuth is

$$\alpha_1 = \tan^{-1}\frac{v}{u} = \tan^{-1}\left(\left(1+\eta^2\right)\cos\phi_1 \frac{\Delta\lambda - \delta\lambda}{\Delta\phi - \delta\phi}\right). \tag{2.126}$$

For the geodesic distance, there are two choices, e.g., if $\alpha_1 \neq 0$, then from equations (2.119) and (2.125)

$$s_{12} = \frac{N_1 \cos\phi_1}{\sin\alpha_1}\left(\Delta\lambda - \delta\lambda\right). \tag{2.127}$$

Similarly, if $\alpha_1 \neq \pi/2$, then $s_{12} = \dfrac{N_1}{\cos\alpha_1}\dfrac{\Delta\phi - \delta\phi}{1+\eta^2}$. Both equations (2.126) and (2.127) are solved together by iteration with starting values obtained by initially setting $\delta\phi^{(0)} = 0$ and $\delta\lambda^{(0)} = 0$:

$$\alpha_1^{(0)} = \tan^{-1}\left(\left(1+\eta^2\right)\cos\phi_1 \frac{\Delta\lambda}{\Delta\phi}\right), \quad s_{12}^{(0)} = \frac{N_1 \cos\phi_1}{\sin\alpha_1^{(0)}}\Delta\lambda. \tag{2.128}$$

Then, using these values of $\alpha_1^{(0)}$ and $s_{12}^{(0)}$, new values, $\delta\phi^{(1)}$ and $\delta\lambda^{(1)}$ are computed using the definitions of $\delta\phi$ and $\delta\lambda$ from equations (2.120) through (2.124). This procedure continues according to

$$\alpha_1^{(j)} = \tan^{-1}\left(\left(1+\eta^2\right)\cos\phi_1 \frac{\Delta\lambda - \delta\lambda^{(j)}}{\Delta\phi - \delta\phi^{(j)}}\right), \quad s_{12}^{(j)} = \frac{N_1 \cos\phi_1}{\sin\alpha_1^{(j)}}\left(\Delta\lambda - \delta\lambda^{(j)}\right), \quad j = 1, 2, \ldots. \tag{2.129}$$

Note that the updates $\delta\phi^{(j)}$ and $\delta\lambda^{(j)}$ are computed using both $s_{12}^{(j-1)}$ and $\alpha_1^{(j-1)}$; and, therefore, the iteration must be done in concert for both s_{12} and α_1. Also, $\bar{\alpha}_2$ is computed using the solution of the direct problem, (2.122), once α_1, u, and v have been determined through the iteration. The correct quadrant of the azimuth should be determined by inspecting the signs of u and v.

The iteration continues until the differences between consecutive values of s_{12} and α_1 are smaller than some pre-defined tolerance. Note, however, that the accuracy of the result depends ultimately on the number of terms retained in $\delta\phi$ and $\delta\lambda$. Rapp (1991) reports that the accuracy of the fifth-order solutions is about 0.01 arcsec in the angles for distances of 200 km. Again, exact solutions exist that are valid for any distance, and which are only marginally more complicated mathematically, as derived in Rapp (1992).

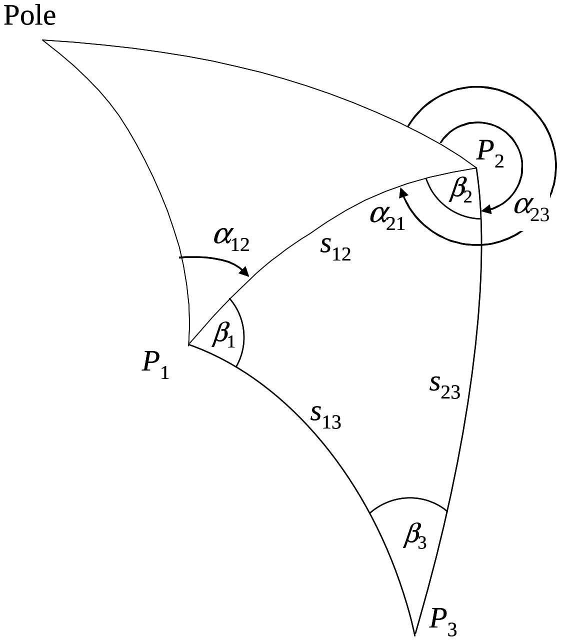
2.1.4.1 Problems

1. Derive equations (2.110) through (2.112).

2. Derive equations (2.116) and (2.118).

3. Derive equations (2.120) through (2.122) up to second order in products of *u* and *v*.

4. Consider an ellipsoidal triangle, $\Delta 123$, with sides being geodesics of arbitrary length. The following are given: lengths of sides, s_{12} and s_{13}, the angle, β_1, the latitude and longitude of point 1, (ϕ_1, λ_1), and the azimuth, α_{12} (see the Figure at the right and note the minor change in notation from the main text). Provide a detailed procedure (i.e., what problems have to be solved and provide input and output to each problem solution) to determine the other two angles, β_2, β_3, and the remaining side of the triangle, s_{23}.



5. Provide an algorithm that ensures proper quadrant determination for the azimuth in the direct and inverse problems.

6. For two points on an ellipsoid, with known coordinates, give a procedure to determine the constant in Clairaut's equation for the geodesic that connects them.

2.1.5 Transformations Between Geodetic and Cartesian Coordinates

The transformation between Cartesian and spherical coordinates is straightforward (Problem 1.4-1). Transforming from the geodetic coordinates, (ϕ, λ, h), for points in space and related to the ellipsoid, (a, f), to Cartesian coordinates, (x, y, z), is also relatively simple, as shown below. The reverse transformation from Cartesian to geodetic coordinates is not as easy. In all cases, for the sake of simplicity, it is assumed that the Cartesian origin is at the ellipsoid center and that the Cartesian coordinate axes are mutually orthogonal along the minor axis and in the equator of the ellipsoid. Referring to Figure 2.15a, it is seen that

$$\begin{aligned} x &= p\cos\lambda \\ y &= p\sin\lambda \end{aligned} \tag{2.130}$$

where $p = \sqrt{x^2 + y^2}$. Since also (compare with equation (2.47))

$$p = (N + h)\cos\phi, \tag{2.131}$$

from Figure 2.15b, it follows that

$$x = (N + h)\cos\phi\cos\lambda, \tag{2.132}$$

$$y = (N + h)\cos\phi\sin\lambda. \tag{2.133}$$

Now, from equations (2.22) and (2.48), we also have:

$$z = \left(N\left(1 - e^2\right) + h\right)\sin\phi. \tag{2.134}$$

In summary, given geodetic coordinates, (ϕ, λ, h), and the ellipsoid to which they refer, the Cartesian coordinates, (x, y, z), are computed according to:

$$\begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} (N + h)\cos\phi\cos\lambda \\ (N + h)\cos\phi\sin\lambda \\ \left(N\left(1 - e^2\right) + h\right)\sin\phi \end{pmatrix}. \tag{2.135}$$

It is emphasized that the transformation from geodetic coordinates to Cartesian coordinates cannot be done using equation (2.135) without knowing the ellipsoid parameters, including the presumptions on the origin and orientation of the axes. These obvious facts are sometimes forgotten, but are extremely important when considering different geodetic datums and reference systems.

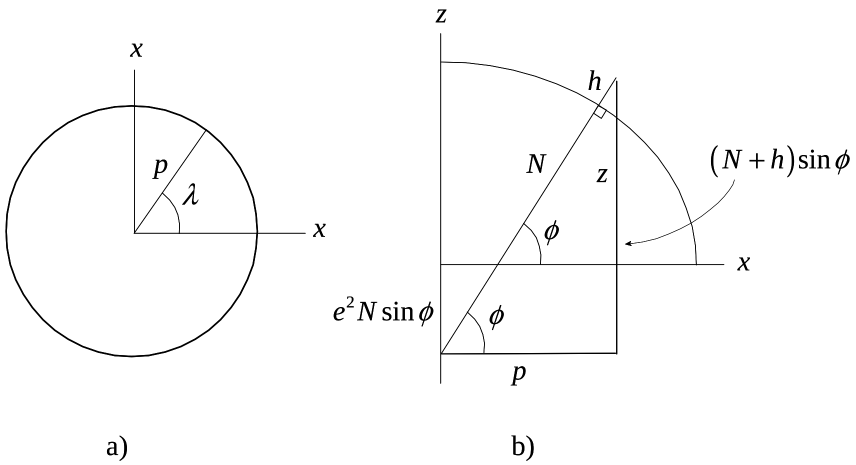


Figure 2.15: Geodetic latitude vs. Cartesian coordinates.

The reverse transformation from Cartesian to geodetic coordinates may be accomplished by various techniques. The usual method is by iteration, but closed formulas also exist. The longitude is easily computed from equations (2.130):

$$\lambda = \tan^{-1}\frac{y}{x}. \qquad (2.136)$$

The problem is in the computation of the geodetic latitude, but only if $h \neq 0$. From Figure 2.15b, we find

$$\tan\phi = \frac{(N+h)\sin\phi}{\sqrt{x^2+y^2}}; \qquad (2.137)$$

and, from equation (2.134), there is

$$(N+h)\sin\phi = z + Ne^2\sin\phi. \qquad (2.138)$$

Therefore, equation (2.137) can be re-written as

$$\phi = \tan^{-1}\left(\frac{z}{\sqrt{x^2+y^2}}\left(1+\frac{e^2 N \sin\phi}{z}\right)\right), \tag{2.139}$$

for $z \neq 0$. If $z = 0$, then, of course, $\phi = 0$. Formula (2.139) is iterated on ϕ, with starting value obtained by initially setting $h = 0$ in equation (2.134) and substituting the resulting $z = N\left(1-e^2\right)\sin\phi$ into equation (2.139):

$$\phi^{(0)} = \tan^{-1}\left(\frac{z}{\sqrt{x^2+y^2}}\left(1+\frac{e^2}{1-e^2}\right)\right). \tag{2.140}$$

Then, the iterations proceed as follows:

$$\phi^{(j)} = \tan^{-1}\left(\frac{z}{\sqrt{x^2+y^2}}\left(1+\frac{e^2 N^{(j-1)} \sin\phi^{(j-1)}}{z}\right)\right), \quad j = 1, 2, \ldots, \tag{2.141}$$

where $N^{(j-1)}$ is the prime vertical radius of curvature for the latitude, $\phi^{(j-1)}$. The iteration continues until the difference between the new and old values of ϕ is less than some pre-defined tolerance. This procedure is known as the *Hirvonen/Moritz algorithm*. Rapp (1991, p.123-124) gives another iteration scheme developed by Bowring that converges faster. However, the scheme above is also sufficiently fast for most practical applications (usually no more than two iterations are required to obtain mm-accuracy), and with today's computers the rate of convergence is not an issue. Finally, a closed (non-iterative) scheme has been developed by several geodesists; the one currently recommended by the International Earth Rotation and Reference Systems Service (IERS) is given by Borkowski (1989). In essence, the solution requires finding the roots of a quartic equation.

Once ϕ is known, the ellipsoid height, h, can be computed according to several formulas. From equations (2.131), we have

$$h = \frac{\sqrt{x^2+y^2}}{\cos\phi} - N, \quad \phi \neq 90°; \tag{2.142}$$

and, from equation (2.134), there is

$$h = \frac{z}{\sin\phi} - N\left(1-e^2\right), \quad \phi \neq 0°. \tag{2.143}$$

From Figure 2.16 and simple trigonometric relationships, a formula that holds for all latitudes is

$$h = \left(\sqrt{x^2+y^2} - N\cos\phi\right)\cos\phi + \left(z - N\left(1-e^2\right)\sin\phi\right)\sin\phi\,, \tag{2.144}$$

which simplifies, using equation (2.48) to

$$h = \sqrt{x^2+y^2}\,\cos\phi + z\sin\phi - a\sqrt{1-e^2\sin^2\phi}\,. \tag{2.145}$$

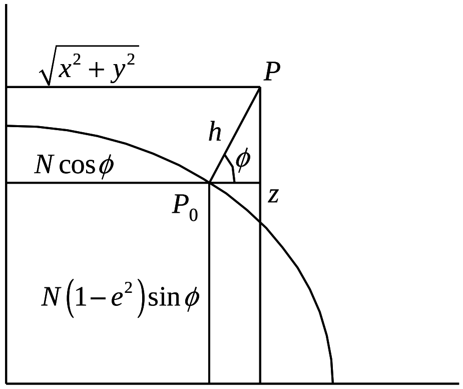


Figure 2.16: Determination of h from $\left(x, y, z\right)$ and ϕ.

2.1.5.1 Problems

1. Derive equation (2.144).

2. Show that the Cartesian coordinates, (x, y, z), can be computed from given ellipsoidal coordinates, (β, λ, u), according to

$$\begin{aligned} x &= \sqrt{u^2 + E^2}\cos\beta\cos\lambda \\ y &= \sqrt{u^2 + E^2}\cos\beta\sin\lambda \\ z &= u\sin\beta \end{aligned} \tag{2.146}$$

3. Show that the ellipsoidal coordinates, (β, λ, u), referring to an ellipsoid with linear eccentricity, E, can be computed from given Cartesian coordinates, (x, y, z), according to

$$\begin{aligned} \lambda &= \tan^{-1}\frac{y}{x} \\ u &= \left(\frac{1}{2}\left(r^2 - E^2\right) + \frac{1}{2}\sqrt{\left(r^2 + E^2\right)^2 - 4E^2p^2}\right)^{1/2} \\ \beta &= \tan^{-1}\frac{z\sqrt{u^2 + E^2}}{u\,p} \end{aligned} \tag{2.147}$$

where $r^2 = x^2 + y^2 + z^2$ and $p^2 = x^2 + y^2$. [Hint: Show that $p^2 = \left(u^2 + E^2\right)\cos^2\beta$ and $z^2 = u^2\sin^2\beta$; and use these two equations to solve for u^2 and then β.]

2.2 Astronomic Coordinates

Traditionally, for example with a theodolite (a telescope that rotates with respect to vertical and horizontal graduated circles), angular measurements (horizontal angles, directions, and vertical angles) are made with respect to the direction of gravity at a point, that is, with respect to the tangent to the *local plumb line*. The direction of gravity at any point is determined naturally by the Earth's somewhat arbitrary mass distribution and the plumb line is defined by this direction. Correspondingly, the plane that is perpendicular to the plumb line at a point defines the local horizontal, or level, plane. The direction of gravity changes from point to point, even along the vertical, making the plumb line a curved line in space, and one speaks of the *tangent* to the plumb line at a point when identifying it with the direction of gravity. Making such angular measurements as described above when the target points are the stars with known coordinates, in fact, leads to the determination of a type of azimuth and a type of latitude and longitude. These latter terrestrial coordinates are known, therefore, as *astronomic coordinates*, or also *natural coordinates*, because they are defined by nature (the direction of the gravity vector) and not by some adopted ellipsoid.

We start by defining a system for these coordinates. This definition has changed as the realization of the system has evolved with technological advancements from purely optical observations (theodolites, transit telescopes, astrolabes, zenith tubes) to global satellite and space observations (satellite Doppler, satellite laser tracking, GPS, very long baseline Interferometry (VLBI), lunar laser ranging). Prior to the satellite and space age, the z-axis of this system was defined in some conventional way by the Earth's spin axis. Saving the details for Chapters 4, it is noted that the spin axis is not fixed relative to the Earth's surface (polar motion). Therefore, the z-axis was defined by the *mean* motion of the pole and was called the *Conventional International Origin* (CIO). Today, the former astronomic system is replaced by the *IERS Terrestrial Reference System* (ITRS), which is established and maintained by the International Earth Rotation and Reference Systems Service (IERS). The ITRS is also known as a *Conventional Terrestrial Reference System* (one that is established by international agreement). The corresponding z-axis is precisely defined according to slightly different adopted conventions (Chapter 3) and is referred to as the *IERS (International) Reference Pole* (IRP). The origin for longitudes is defined by the direction of a meridian through the Greenwich Observatory in both cases; again, there is a subtle difference in the conventions and the realization (Chapter 3).

To understand the traditional astronomic system, we define the *astronomic meridian plane* for any specific point, analogous to the geodetic meridian plane for point coordinates associated with an ellipsoid. However, there is one essential and important difference. The astronomic meridian plane is the plane that contains the tangent to the plumb line at a point and is (only) parallel to the z-axis. Recall that the geodetic meridian plane contains the normal to the ellipsoid, as well as the minor axis of the ellipsoid. The astronomic meridian plane does not, generally, contain the z-axis. To show that this plane always exists, simply consider the vector at any point, P, that is parallel to the z-axis (Figure 2.17). This vector and the vector tangent to the plumb line together form a plane, the astronomic meridian plane, and it is parallel to the z-axis. Again, it is emphasized that the tangent to the plumb line does not intersect Earth's center of mass (nor its spin axis) due to the arbitrary direction of gravity. The *Greenwich meridian plane* is the plane that is parallel to the z-axis and contains the tangent to the plumb line at the Greenwich Observatory. The x-axis is parallel to this plane by definition.

Now, the *astronomic latitude*, Φ, of a point is the angle in the astronomic meridian plane from the equator (plane perpendicular to the z-axis) to the tangent of the plumb line. And, the *astronomic longitude*, Λ, is the angle in the equator from the Greenwich meridian plane to the local astronomic meridian plane. The astronomic coordinates, (Φ, Λ), determine the direction of the tangent to the plumb line, just like the geodetic coordinates, (ϕ, λ), define the direction of the ellipsoid normal. The difference between these two directions at a point is known as the *deflection of the vertical*. This angle is explored in detail in Section 2.2.3.

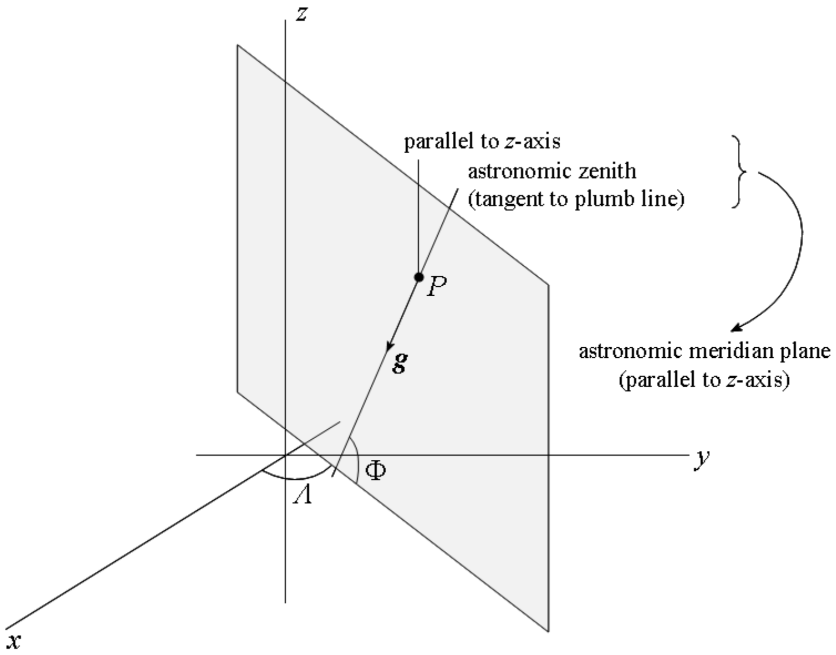


Figure 2.17: Astronomic meridian plane and astronomic coordinates.

To complete the analogy with previously defined geodetic quantities, we also consider the astronomic azimuth. The *astronomic azimuth* is the angle in the *astronomic horizon* (the plane perpendicular to the tangent of the plumb line) from the northern half of the astronomic meridian, easterly, to the plane containing both the plumb line tangent and the target point (the *vertical plane*); see Figure 2.19. Finally, the *astronomic zenith angle* (also known as the *zenith distance*) is the angle in the vertical plane from the tangent to the (outward) plumb line (*astronomic zenith*) to the target point. We note that heights are not part of the astronomic coordinates, but that heights may be included in the definition of natural coordinates, where in this case the height is based on the geopotential; this is treated later briefly in connection with vertical datums (Chapter 3).

2.2.1 Problems

1. Provide a justification that, theoretically, two distinct points on a surface (like the ellipsoid, or geoid) could have the same astronomic latitude and longitude, Φ and Λ.

2. Determine which of the following would affect the astronomic coordinates of a fixed point on the Earth's surface: i) a translation of the coordinate origin of the (x, y, z) system; ii) a general rotation of the (x, y, z) system. Determine which of the following would be affected by a rotation about the z-axis: astronomic latitude, Φ; astronomic longitude, Λ; astronomic azimuth, A. Justify your answers in all cases.

3. Assume that the ellipsoid axes are parallel to the (x, y, z) system. Geometrically determine if the geodetic and astronomic meridian planes for a point are parallel; provide a drawing with sufficient discussion to justify your answer. What are the most general conditions under which these two planes would be parallel?

2.2.2 Local Terrestrial Coordinates

This set of coordinates forms the basis for traditional three-dimensional geodesy and for close-range, local surveys. It is the local system in which we make traditional geodetic measurements of distance and angles, or directions, using distance measuring devices, theodolites, and combinations thereof (total station). It is also still used for modern measurement systems, such as in photogrammetry, for local referencing of geospatial data, and in assigning directions for navigation. The local coordinate system can be defined with respect to the local ellipsoid normal (*local geodetic system*) or the local gravity vector (*local astronomic system*). The developments for both are identical, where the only difference in the end is the specification at one point of the type of latitude and longitude, i.e., the direction of the vertical. The local system is Cartesian, consisting of three mutually orthogonal axes; however, their principal directions do not always follow conventional definitions (in surveying the directions are north, east, and up; in navigation, they are north, east, and down, or north, west, and up).

For the sake of practical visualization, consider first the *local astronomic system* (Figure 2.18). The third axis, w, is aligned with the tangent to the plumb line at the local origin point, P, which is also the observer's point. The first axis, u, is orthogonal to w and in the direction of north, defined by the astronomic meridian. And, the second axis, v, is orthogonal to w and u and points east. Note that u, v, w are coordinates in a *left-handed* system. Let Q be a target point and consider the coordinates of Q in this local astronomic system.

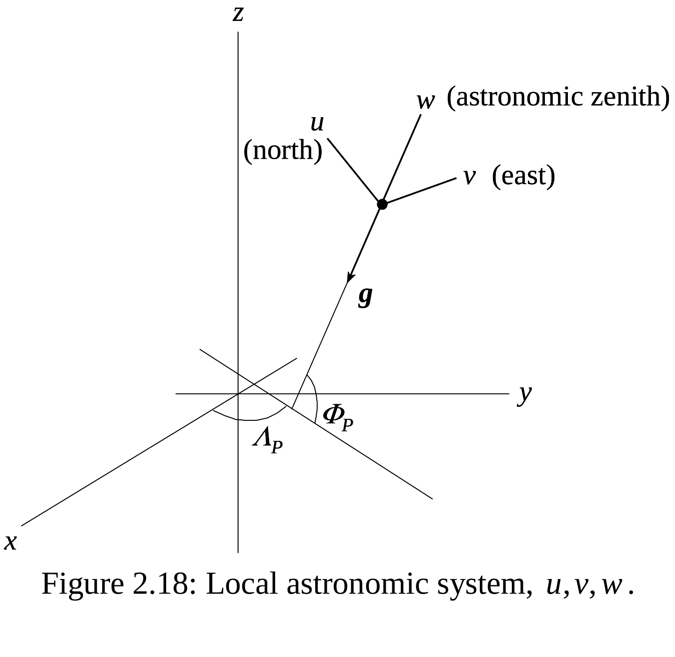


Figure 2.18: Local astronomic system, u, v, w.

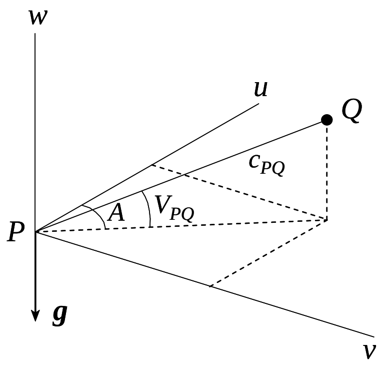


Figure 2.19: Local astronomic coordinates and measured quantities.

With reference to Figure 2.19, the measured quantities are the distance from P to Q, denoted by c_{PQ}; the astronomic azimuth of Q at P, denoted A_{PQ} (it is discussed in Section 2.3 how to determine azimuths from astronomic observations); and the vertical angle of Q at P, denoted, V_{PQ}. The local Cartesian coordinates of Q in the system centered at P are given in terms of these measured quantities by

$$\begin{aligned} u_{PQ} &= c_{PQ}\cos V_{PQ}\cos A_{PQ} \\ v_{PQ} &= c_{PQ}\cos V_{PQ}\sin A_{PQ} \\ w_{PQ} &= c_{PQ}\sin V_{PQ} \end{aligned} \tag{2.148}$$

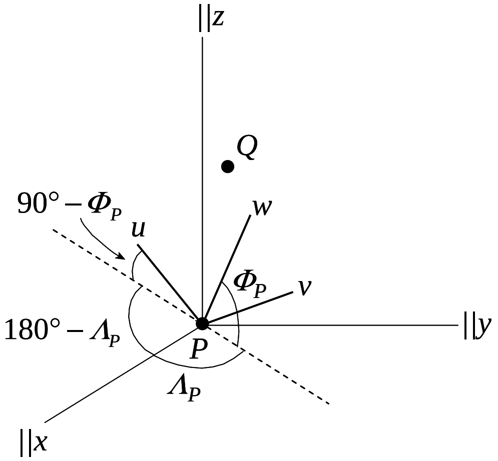


Figure 2.20: The relationship between the u, v, w and $\| x, \| y, \| z$ systems.

Consider now a Cartesian coordinate system at P that is parallel to the global x, y, z system (Figure 2.20); denote its axes, respectively, by $\| x$, $\| y$, and $\| z$. Note that the v-axis is always in the plane generated by $\| x$ and $\| y$ since the u,w-plane is perpendicular to the equator because of the definition of the meridian plane. The Cartesian coordinates of the point Q in this system are simply

$$\begin{aligned} &\| x_{PQ} \equiv \mathit{\Delta} x_{PQ} = x_Q - x_P \\ &\| y_{PQ} \equiv \mathit{\Delta} y_{PQ} = y_Q - y_P \\ &\| z_{PQ} \equiv \mathit{\Delta} z_{PQ} = z_Q - z_P \end{aligned} \qquad (2.149)$$

The relationship between the u,v,w and $\| x, \| y, \| z$ systems is one of rotation and accounting for the different handedness of the two systems. The following transformations change the local coordinates of the point, Q, from (u,v,w) to $(\| x, \| y, \| z)$:

$$\begin{pmatrix} \mathit{\Delta} x_{PQ} \\ \mathit{\Delta} y_{PQ} \\ \mathit{\Delta} z_{PQ} \end{pmatrix} = R_3\left(180° - \mathit{\Lambda}_P\right) R_2\left(90° - \mathit{\Phi}_P\right) \begin{pmatrix} 1 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} u_{PQ} \\ v_{PQ} \\ w_{PQ} \end{pmatrix}, \qquad (2.150)$$

where the right-most matrix transforms from a left-handed system to a right-handed system; only then can one apply the rotation matrices, R_1 and R_2, defined by equations (1.4) and (1.5). The resulting transformation is (left to the reader to verify):

$$\begin{pmatrix} \mathit{\Delta} x_{PQ} \\ \mathit{\Delta} y_{PQ} \\ \mathit{\Delta} z_{PQ} \end{pmatrix} = \begin{pmatrix} -\sin \mathit{\Phi}_P \cos \mathit{\Lambda}_P & -\sin \mathit{\Lambda}_P & \cos \mathit{\Phi}_P \cos \mathit{\Lambda}_P \\ -\sin \mathit{\Phi}_P \sin \mathit{\Lambda}_P & \cos \mathit{\Lambda}_P & \cos \mathit{\Phi}_P \sin \mathit{\Lambda}_P \\ \cos \mathit{\Phi}_P & 0 & \sin \mathit{\Phi}_P \end{pmatrix} \begin{pmatrix} u_{PQ} \\ v_{PQ} \\ w_{PQ} \end{pmatrix}. \qquad (2.151)$$

Therefore, substituting equation (2.148) gives

$$\begin{pmatrix} \mathit{\Delta} x_{PQ} \\ \mathit{\Delta} y_{PQ} \\ \mathit{\Delta} z_{PQ} \end{pmatrix} = \begin{pmatrix} -\sin \mathit{\Phi}_P \cos \mathit{\Lambda}_P & -\sin \mathit{\Lambda}_P & \cos \mathit{\Phi}_P \cos \mathit{\Lambda}_P \\ -\sin \mathit{\Phi}_P \sin \mathit{\Lambda}_P & \cos \mathit{\Lambda}_P & \cos \mathit{\Phi}_P \sin \mathit{\Lambda}_P \\ \cos \mathit{\Phi}_P & 0 & \sin \mathit{\Phi}_P \end{pmatrix} \begin{pmatrix} c_{PQ} \cos V_{PQ} \cos A_{PQ} \\ c_{PQ} \cos V_{PQ} \sin A_{PQ} \\ c_{PQ} \sin V_{PQ} \end{pmatrix}, \qquad (2.152)$$

which describes the transformation from measured quantities, c_{PQ}, V_{PQ}, A_{PQ}, to Cartesian coordinate *differences* in a global system, provided also astronomic latitude and longitude of the observer's point are known.

It is remarkable that conventional determinations of astronomic latitude and longitude (see Section 2.3), as well as of astronomic azimuth, vertical angle, and distance can be used to determine these relative Cartesian coordinates – this is the basis for traditional three-dimensional geodesy, that is, the computation of all three coordinates of points from terrestrial geometric measurements. It is noted, again, that these determinations are relative, not absolute, where the latter can be obtained only by specifying the coordinates, $\left(x_P, y_P, z_P\right)$, of the observer's point in the global system. Nowadays, of course, satellite systems, such as the Global Positioning System (GPS), provide the three-dimensional Cartesian coordinates virtually effortlessly in a global system. Historically (before satellites), however, three-dimensional geodesy could not be realized very accurately because of the difficulty of obtaining the vertical angle without significant atmospheric refraction error. This is one of the principal reasons that traditional geodetic control for a country was separated into horizontal and vertical networks, where the latter is achieved by leveling (and is, therefore, not strictly geometric, but based on the geopotential).

The reverse transformation from $\left(\Delta x_{PQ}, \Delta y_{PQ}, \Delta z_{PQ}\right)$ to $\left(c_{PQ}, V_{PQ}, A_{PQ}\right)$ is easily obtained since the transformation matrix is orthogonal. From equation (2.151), we have

$$\begin{pmatrix} u_{PQ} \\ v_{PQ} \\ w_{PQ} \end{pmatrix} = \begin{pmatrix} -\sin \Phi_P \cos \Lambda_P & -\sin \Lambda_P & \cos \Phi_P \cos \Lambda_P \\ -\sin \Phi_P \sin \Lambda_P & \cos \Lambda_P & \cos \Phi_P \sin \Lambda_P \\ \cos \Phi_P & 0 & \sin \Phi_P \end{pmatrix}^{\mathrm{T}} \begin{pmatrix} \Delta x_{PQ} \\ \Delta y_{PQ} \\ \Delta z_{PQ} \end{pmatrix}; \tag{2.153}$$

and, with equation (2.148), it is easily verified that

$$\tan A_{PQ} = \frac{v_{PQ}}{u_{PQ}} = \frac{-\Delta x_{PQ} \sin \Lambda_P + \Delta y_{PQ} \cos \Lambda_P}{-\Delta x_{PQ} \sin \Phi_P \cos \Lambda_P - \Delta y_{PQ} \sin \Phi_P \sin \Lambda_P + \Delta z_{PQ} \cos \Phi_P}, \tag{2.154}$$

$$\sin V_{PQ} = \frac{w_{PQ}}{c_{PQ}} = \frac{1}{c_{PQ}}\left(\Delta x_{PQ} \cos \Phi_P \cos \Lambda_P + \Delta y_{PQ} \cos \Phi_P \sin \Lambda_P + \Delta z_{PQ} \sin \Phi_P\right), \tag{2.155}$$

$$c_{PQ} = \sqrt{\Delta x_{PQ}^2 + \Delta y_{PQ}^2 + \Delta z_{PQ}^2}\ . \tag{2.156}$$

Analogous equations hold in the case of the *local geodetic coordinate system*. In this case the ellipsoid normal serves as the third axis, as shown in Figure 2.21, and the other two axes are mutually orthogonal and positioned similar to the axes in the local astronomic system. We assume that the ellipsoid is centered at the origin of the x, y, z system and designate the local geodetic coordinates by $\left(r, s, t\right)$. It is easily seen that the only difference between the local geodetic and the local astronomic coordinate systems is the direction of corresponding axes, specifically the direction of the third axis; and, this is defined by the geodetic latitude and longitude. This means that the analogues to equations (2.152) and (2.154) through (2.156) for the local geodetic system are obtained simply by replacing the astronomic coordinates with the geodetic latitude and longitude, ϕ_P and λ_P:

$$\begin{pmatrix} \Delta x_{PQ} \\ \Delta y_{PQ} \\ \Delta z_{PQ} \end{pmatrix} = \begin{pmatrix} -\sin\phi_P \cos\lambda_P & -\sin\lambda_P & \cos\phi_P \cos\lambda_P \\ -\sin\phi_P \sin\lambda_P & \cos\lambda_P & \cos\phi_P \sin\lambda_P \\ \cos\phi_P & 0 & \sin\phi_P \end{pmatrix} \begin{pmatrix} c_{PQ} \cos v_{PQ} \cos\alpha_{PQ} \\ c_{PQ} \cos v_{PQ} \sin\alpha_{PQ} \\ c_{PQ} \sin v_{PQ} \end{pmatrix}, \tag{2.157}$$

where α_{PQ} is the normal section azimuth and v_{PQ} is the vertical angle in the normal plane of Q. The reverse relationships are given by

$$\tan\alpha_{PQ} = \frac{-\Delta x_{PQ} \sin\lambda_P + \Delta y_{PQ} \cos\lambda_P}{-\Delta x_{PQ} \sin\phi_P \cos\lambda_P - \Delta y_{PQ} \sin\phi_P \sin\lambda_P + \Delta z_{PQ} \cos\phi_P}, \tag{2.158}$$

$$\sin v_{PQ} = \frac{1}{c_{PQ}} \left(\Delta x_{PQ} \cos\phi_P \cos\lambda_P + \Delta y_{PQ} \cos\phi_P \sin\lambda_P + \Delta z_{PQ} \sin\phi_P \right), \tag{2.159}$$

$$c_{PQ} = \sqrt{\Delta x_{PQ}^2 + \Delta y_{PQ}^2 + \Delta z_{PQ}^2} \,. \tag{2.160}$$

The latter equations have application, in particular, when determining normal section azimuth, distance, and vertical angle (in the normal plane) from satellite-derived Cartesian coordinate differences between points (such as from GPS). Note that the formulas hold for any point, not necessarily on the ellipsoid, and, again, that it is the normal section azimuth, *not* the geodesic azimuth in these formulas.

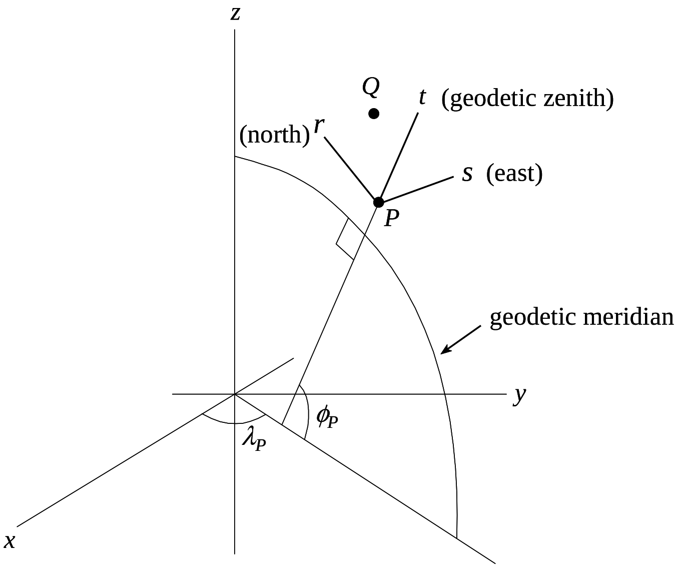


Figure 2.21: Local geodetic coordinate system.

2.2.2.1 Problems

1. Derive equation (2.151).

2. Show that the transformation from local geodetic to local astronomic coordinates (same origin point, P) is given by

$$\begin{pmatrix} u_{PQ} \\ v_{PQ} \\ w_{PQ} \end{pmatrix} = \begin{pmatrix} 1 & -\left(\mathit{\Lambda}_P - \lambda_P\right)\sin\phi_P & -\left(\mathit{\Phi}_P - \phi_P\right) \\ \left(\mathit{\Lambda}_P - \lambda_P\right)\sin\phi_P & 1 & -\left(\mathit{\Lambda}_P - \lambda_P\right)\cos\phi_P \\ \mathit{\Phi}_P - \phi_P & \left(\mathit{\Lambda}_P - \lambda_P\right)\cos\phi_P & 1 \end{pmatrix} \begin{pmatrix} r_{PQ} \\ s_{PQ} \\ t_{PQ} \end{pmatrix}, \qquad (2.161)$$

where second and higher powers in the differences, $\left(\mathit{\Phi}_P - \phi_P\right)$ and $\left(\mathit{\Lambda}_P - \lambda_P\right)$, have been neglected. (Hint: the coordinates in the two systems have the same Cartesian coordinate differences.)

3. Suppose the geodetic coordinates, $\left(\phi_P, \lambda_P\right)$ and $\left(\phi_Q, \lambda_Q\right)$, of two points on the ellipsoid are given and the distance between them is under 200 km. Develop a procedure to test the computation of the *geodesic* azimuths, $\tilde{\alpha}_{PQ}$ and $\tilde{\alpha}_{QP}$, obtained by the solution to the inverse geodetic problem (Section 2.1.4). Discuss the validity of your procedure also from a numerical viewpoint.

4.a) Derive the following two equalities:

$$\tan\left(A_{PQ} - \alpha_{PQ}\right) = \frac{\tan A_{PQ} - \tan\alpha_{PQ}}{1 + \tan A_{PQ}\tan\alpha_{PQ}} = \frac{r_{PQ}v_{PQ} - s_{PQ}u_{PQ}}{r_{PQ}u_{PQ} + s_{PQ}v_{PQ}}. \qquad (2.162)$$

b) Now, show that to first-order approximation, i.e., neglecting second and higher powers in the differences, $\left(\mathit{\Phi}_P - \phi_P\right)$ and $\left(\mathit{\Lambda}_P - \lambda_P\right)$:

$$\tan\left(A_{PQ} - \alpha_{PQ}\right) \simeq \left(\mathit{\Lambda}_P - \lambda_P\right)\sin\phi_P + \frac{s_{PQ}t_{PQ}}{r_{PQ}^2 + s_{PQ}^2}\left(\mathit{\Phi}_P - \phi_P\right) - \frac{r_{PQ}t_{PQ}}{r_{PQ}^2 + s_{PQ}^2}\left(\mathit{\Lambda}_P - \lambda_P\right)\cos\phi_P. \qquad (2.163)$$

(Hint: use equation (2.161).)

c) Finally, with the same approximation show that

$$A_{PQ} - \alpha_{PQ} \simeq \left(\mathit{\Lambda}_P - \lambda_P\right)\sin\phi_P + \left(\left(\mathit{\Phi}_P - \phi_P\right)\sin\alpha_{PQ} - \left(\mathit{\Lambda}_P - \lambda_P\right)\cos\phi_P\cos\alpha_{PQ}\right)\tan v_{PQ}. \qquad (2.164)$$

The latter is known as the (extended) *Laplace condition*, which is derived from a more geometric perspective in Section 2.2.3.

2.2.3 Differences Between Geodetic and Astronomic Quantities

As we will see in Section 2.3, the astronomic latitude, longitude, and azimuth are observable quantities based on a naturally defined and realized coordinate system, such as the astronomic system or the terrestrial reference system alluded to in Section 2.2. These quantities also depend on the direction of gravity at a point (another naturally defined and realizable direction). However, the quantities used for mapping purposes are the geodetic coordinates, based on a mathematically defined ellipsoid. Therefore, it is necessary to develop equations for the difference between the geodetic and astronomic coordinates (and azimuths), in order to relate observed quantities to mathematically and geographically useful quantities. These equations are also extremely important in realizing the proper orientation of one system relative to the other.

Already in Problem 2.2.2.1-4, the reader was asked to derive the difference between astronomic and geodetic azimuth. This is now done using spherical trigonometry that also shows more clearly the differences between astronomic and geodetic latitude and longitude. In fact, however, the latter coordinate differences are not derived, per se, and essentially are just given names, i.e., the components of the astro-geodetic deflection of the vertical, under the following fundamental assumption. Specifically, it is assumed that the two systems, the astronomic (or terrestrial) and geodetic systems, are parallel, meaning that the minor axis of the ellipsoid is parallel to the z-axis of the astronomic system and the corresponding x-axes are parallel. Under this assumption we derive the difference between the azimuths. Alternatively, one could derive the relationships under more general conditions of non-parallelism and subsequently set the orientation angles between axes to zero. The result would obviously be the same, but the procedure is outside the present scope (the relevant equations are given in Section 3.1).

Figure 2.22 depicts the plan view of a sphere of undefined radius as seen from the outside, along the tangent to the plumb line or along the local astronomic coordinate axis, w, that is, from the *astronomic zenith*. The origin of this sphere could be the center of mass of the Earth or the center of mass of the solar system, or even the observer's location. Insofar as the radius is unspecified, it may be taken as sufficiently large so that the origin, for present purposes, is immaterial. This is called the *celestial sphere*; see also Section 2.3. All points on this sphere are projections of radial directions and since one is only concerned with directions, the value of the radius is not important and may, as well, be assigned a value of 1 (unit radius), so that angles between radial directions are equivalent to great circle arcs on the sphere in terms of radian measure.

Clearly, the circle shown in Figure 2.22 is the (astronomic) *horizon*. Z_a denotes the astronomic zenith, and Z_g is the geodetic zenith, being the projection of the ellipsoidal normal through the observer, P (see Figure 2.21). As noted earlier, the angular arc between the two zeniths is the *astro-geodetic deflection of the vertical*, Θ (the deflection of the tangent to the plumb line from a mathematically defined vertical, the ellipsoid normal). It may be decomposed into two angles, one in the south-to-north direction, ξ, and one in the west-to-east direction, η (Figure 2.23). The projections of the astronomic meridian and the geodetic meridian intersect on the celestial sphere because the polar axes of the two systems are parallel by assumption (even though the astronomic meridian plane does not *contain* the z-axis, the fact that both meridian planes are parallel to the z-axis implies that on the celestial sphere, their projections intersect in the projection of the north pole). On the horizon, however, there is a difference, Δ_1, between astronomic and geodetic north.

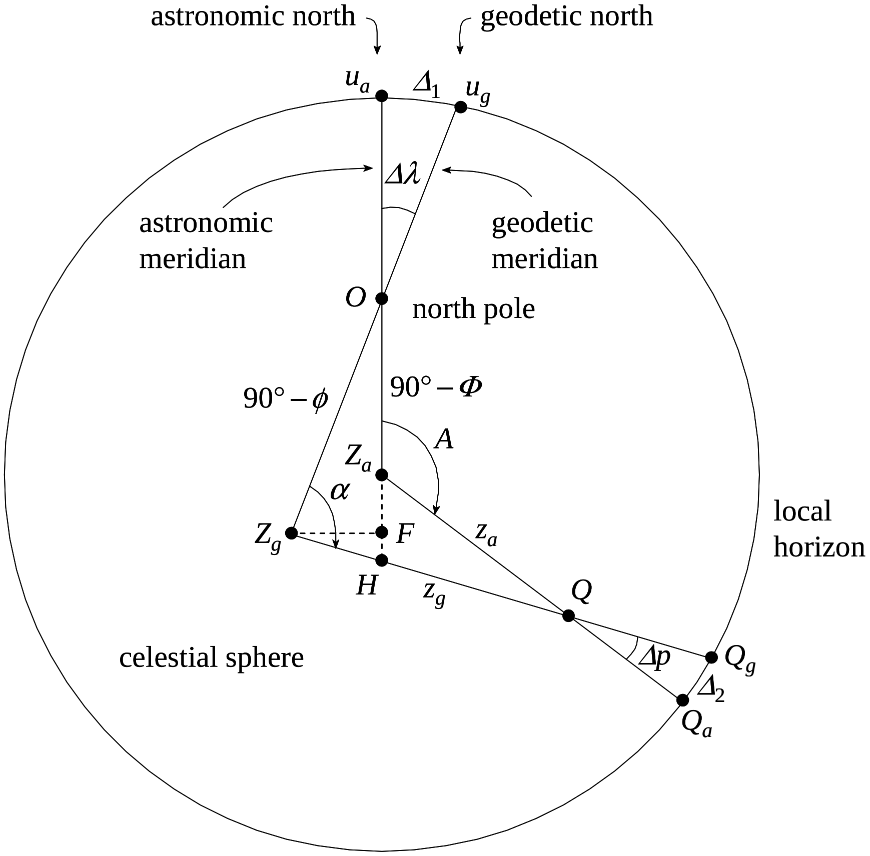


Figure 2.22: Astronomic and geodetic azimuths.

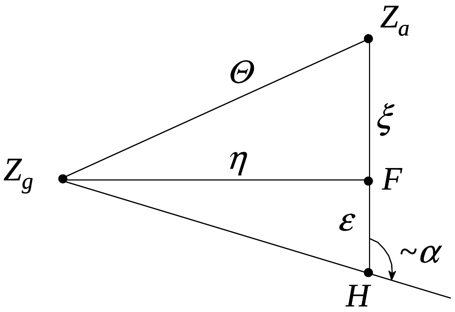


Figure 2.23: Deflection of the vertical components.

Now, the angle at the north pole between the meridians is $\Delta\lambda = \Lambda - \lambda$, again, because the two systems presumably have parallel x-axes (common origin on the celestial sphere). From the indicated astronomic and geodetic latitudes, application of the law of cosines, equation (1.2), to the triangle $Z_g OF$ yields

$$\cos\left(90° - \phi\right) = \cos\eta \cos\left(90° - \Phi + \xi\right) + \sin\eta \sin\left(90° - \Phi + \xi\right)\cos 90° . \tag{2.165}$$

Since η is a small angle (usually of the order of 10 arcsec), this simplifies to

$$\sin\phi \simeq \sin\left(\Phi - \xi\right), \tag{2.166}$$

and hence

$$\xi \simeq \Phi - \phi . \tag{2.167}$$

Applying the law of sines, equation (1.1), to the same triangle, $Z_g OF$, one finds

$$\frac{\sin\eta}{\sin\Delta\lambda} = \frac{\sin\left(90° - \phi\right)}{\sin 90°}; \tag{2.168}$$

and, hence, with the same approximations,

$$\eta \simeq \left(\Lambda - \lambda\right)\cos\phi . \tag{2.169}$$

Thus, the north and east components, ξ and η, of the deflection of the vertical are essentially the differences between the astronomic and the geodetic latitudes and longitudes, respectively.

The great circle arc, $\overset{\frown}{u_a Q_a}$, in Figure 2.22 is the same as the astronomic azimuth, A, to the target point, Q, while the great circle arc (approximately, since the two zeniths are close), $\overset{\frown}{u_g Q_g}$, is the same as the geodetic (normal section) azimuth, α, of the target point. Thus, from Figure 2.22, one obtains

$$A - \alpha = \overset{\frown}{u_a Q_a} - \overset{\frown}{u_g Q_g} = \Delta_1 + \Delta_2 . \tag{2.170}$$

It remains to find expressions for Δ_1 and Δ_2.

From the law of sines applied to triangle $u_g O u_a$, we find

$$\frac{\sin \Delta_1}{\sin \Delta\lambda} = \frac{\sin \phi}{\sin 90°} \quad \Rightarrow \quad \Delta_1 = \Delta\lambda \sin \phi , \tag{2.171}$$

with the usual small-angle approximation. Similarly, in triangle $Q_g Q Q_a$, the law of sines yields

$$\frac{\sin \Delta_2}{\sin \Delta p} = \frac{\sin\left(90° - z_g\right)}{\sin 90°} \quad \Rightarrow \quad \Delta_2 = \Delta p \cos z_g \, . \tag{2.172}$$

Also, triangle $Z_a Q H$ (see also Figure 2.23) gives

$$\frac{\sin \Delta p}{\sin\left(\xi + \varepsilon\right)} = \frac{\sin \alpha}{\sin z_a} \quad \Rightarrow \quad \Delta p = \left(\xi + \varepsilon\right) \frac{\sin \alpha}{\sin z_a} \, . \tag{2.173}$$

Finally, from the approximately planar triangle $Z_g F H$ we obtain

$$\varepsilon \simeq \frac{\eta}{\tan\left(180° - \alpha\right)} , \tag{2.174}$$

which could also be derived by rigorously applying the laws of cosines and sines on the spherical triangle and making the usual small-angle approximations.

Substituting equations (2.173) and (2.174) into equation (2.172) results in

$$\begin{aligned} \Delta_2 &= \left(\xi + \varepsilon\right) \sin \alpha \cot z \\ &= \left(\xi \sin \alpha - \eta \cos \alpha\right) \cot z \end{aligned} \tag{2.175}$$

where the approximation $z \simeq z_g \simeq z_a$ is legitimate because of the small magnitude of Δ_2. We come to the final result by combining equations (2.171) and (2.175) with equation (2.170):

$$A - \alpha = \left(\Lambda - \lambda\right) \sin \phi + \left(\xi \sin \alpha - \eta \cos \alpha\right) \cot z \, , \tag{2.176}$$

which, of course, in view of equations (2.167) and (2.169) is the same as equation (2.164). Equation (2.176) is known as the (extended) *Laplace condition*. Again, it is noted that α is the normal section azimuth. The second term on the right side of equation (2.176) is the extended part that vanishes (or nearly so) for target point on (or close to) the horizon, where the zenith angle is 90°. Even though this relationship between astronomic and geodetic azimuths at a point is a consequence of the assumed parallelism of the corresponding system axes, its application to observed astronomic azimuths, in fact, also ensures this parallelism, i.e., it is a sufficient condition. This can be proved by deriving the equation under a general rotation between the systems and specializing to parallel systems. The geodetic (normal section) azimuth, α, determined according to equation (2.176) from observed astronomic quantities is known as the *Laplace azimuth*.

The *simple* Laplace condition (for $z = 90°$),

$$A - \alpha = (\Lambda - \lambda)\sin\phi\,, \tag{2.177}$$

describes the difference in azimuths that is common to all target points, Q, with respect to a given point, P, and is due to the non-parallelism of the astronomic and geodetic meridian planes at the observer's location, i.e., at P (Figure 2.22). Interestingly, the simple Laplace condition is also the Bessel equation derived for geodesics, equation (2.95), which, however, is unrelated to the present context. The second term in the extended Laplace condition (2.174) (for target points with non-zero vertical angle) depends on the azimuth of the target. It is analogous to the error in angles measured by a theodolite whose vertical is out of alignment (leveling error).

A final note on the origin of longitudes is needed to distinguish between directions in space and points on the Earth's surface. Specifically, by definition the zero astronomic and the zero geodetic meridian planes are parallel, hence, so are the astronomic and geodetic x-axes. This is clearly required to make the two systems parallel and it is assumed in the derivations of the deflection of the vertical and Laplace's condition (see also Figure 2.22). However, for a global geodetic system that is *also geocentric*, e.g., that defined by GPS, the deflection of the vertical at the Greenwich Observatory is not zero. By equation (2.169), this means that the zero geodetic longitude in this system on the Earth's surface is not at the Greenwich observatory, which by definition has maintained a zero astronomic longitude. In fact, the geodetic longitude is about 102 m to the east of the Greenwich observatory. This geometry is elaborated in Section 3.3 and Figure 3.3.

2.2.3.1 Problems

1. Suppose the geodetic system is rotated with respect to the astronomic system by the small angle, ω_z, about the polar axis. Repeat all derivations and thus show that the components of the deflection of the vertical and the Laplace condition are now given by

$$\begin{aligned}
&\xi = \Phi - \phi \\
&\eta = \left(\Lambda - \lambda - \omega_z\right)\cos\phi \qquad (2.178) \\
&A - \alpha = \left(\Lambda - \lambda - \omega_z\right)\sin\phi + \left(\left(\Phi - \phi\right)\sin\alpha - \left(\Lambda - \lambda - \omega_z\right)\cos\phi\cos\alpha\right)\cot \mathrm{z}
\end{aligned}$$

2. Suppose that an observer determines the astronomic azimuth of a target. Describe in review fashion all the systematic corrections that must be applied to obtain the corresponding *geodesic azimuth* of the target that has been projected (mapped) along the normal onto an ellipsoid whose axes are parallel to the astronomic system.

2.3 Celestial Coordinates

In order to determine astronomic coordinates of points on the Earth, angular observations of stars are made relative to naturally defined directions on the Earth and combined with the known coordinates of the stars. Therefore, it is necessary to understand how the celestial coordinates of stars are defined and how they can be related through terrestrial observations to the astronomic coordinates. Later we also discuss the orientation of the terrestrial coordinate systems with respect to inertial space and, again, there is need for the celestial coordinates.

For the moment, we deal only with directions, or angles, because all celestial objects that concern us (such as stars or natural radio sources) are extremely distant from the observer on the Earth. Thus, as in Section 2.2, the coordinate directions of observable objects, as well as general directions, are projected radially onto the *celestial sphere*. At the risk of being too repetitive, this is a fictitious sphere having infinite or arbitrary (e.g., unit) radius; and, formally the center of this sphere is at the center of mass of the solar system. However, it can have any of a number of centers (e.g., the geocenter), where transformation from one to the other may or may not require a correction, depending on the accuracy of our computations. Certainly, this is of no consequence for the most distant objects in the universe, the quasars (quasi-stellar radio sources). The main point is that the celestial sphere should not rotate in time, meaning that it defines an *inertial system* (we ignore the effects of general relativity).

To implement the transformation from celestial to astronomic coordinates on the basis of astronomic observations, three coordinate systems are introduced: 1) the *horizon system*, in which astronomic observations are made; 2) the *equatorial, right ascension system*, in which the celestial coordinates of objects are defined; and 3) the *equatorial, hour angle system*, that connects 1) and 2). Each coordinate system is defined by mutually orthogonal axes that are related to naturally occurring directions; two such directions are needed for each system. Each system is either right-handed, or left-handed, by convention.

2.3.1 Horizon System

The horizon system of coordinates is defined on the celestial sphere by the direction of local gravity and by the direction of Earth's spin axis, intersecting the celestial sphere at the *north celestial pole* (NCP) (Figure 2.24). (For the moment it is assumed that the spin axis is fixed to the Earth and in space; see Chapters 3 and 4 for a more precise definitions of the polar direction, both for terrestrial and for celestial systems.) The positive third axis of the horizon system is the negative (upward) direction of gravity (the zenith is in the positive direction). The first axis is defined as perpendicular to the third axis and in the astronomic meridian plane, positive northward. And, the second axis is perpendicular to the first and third axes and positive eastward, so as to form a *left-handed* system. The intersection of the celestial sphere with the plane that contains both the zenith direction and an object is called the *vertical circle*.

The (instantaneous) coordinates of stars (or other celestial objects) in this system are the zenith angle and the astronomic azimuth. These are also the observed quantities; however, instead of azimuth, one typically observes only a horizontal angle with respect to some other accessible reference direction. Both are "astronomic" in the sense of being an angle that is turned about the direction defined by the astronomic zenith. The horizon system is fixed to the Earth and the coordinates of celestial objects change in time as the Earth rotates.

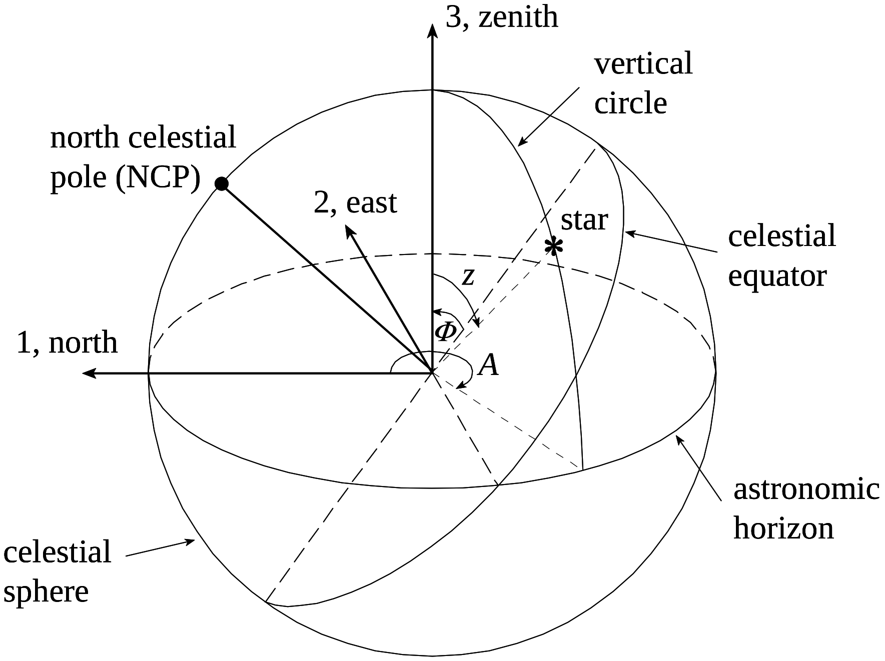


Figure 2.24: Horizon system.

2.3.2 Equatorial, Right Ascension System

The equatorial, right ascension system of coordinates is defined on the celestial sphere by the direction of Earth's spin axis (the north celestial pole) and by the direction of the *north ecliptic pole* (NEP), both of which, again, are naturally defined directions. And, again, it is assumed for the moment that the NEP is fixed in space. Figure 2.25 shows the (mean) ecliptic plane, which is the plane of the average Earth orbit around the sun. The direction perpendicular to this plane is the north ecliptic pole. A point where the ecliptic crosses the celestial equator on the celestial sphere is called an equinox; the *vernal equinox,* Υ, is the equinox at which the sun crosses the celestial equator from south to north as viewed from the Earth. It is the point on the Earth's orbit when Spring starts in the northern hemisphere. The angle between the celestial equator and the ecliptic is the *obliquity of the ecliptic,* ε; its value is approximately $\varepsilon = 23.44°$.

The first axis of the right ascension system is defined by the direction of the vernal equinox and the third axis is defined by the north celestial pole (NCP). By definition these two axes are perpendicular since the vector defining the direction of the vernal equinox lies in the equatorial plane with respect to which the polar axis is perpendicular. The second axis is perpendicular to the other two axes so as to form a *right-handed* system. The intersection of the celestial sphere with the plane that contains both the third axis (NCP) and the object is called the *hour circle* of the object (Figure 2.26), the reason for which will become apparent in Section 2.3.3. The right ascension system is assumed to be fixed in space, i.e., it is an *inertial system* in the sense that it does not rotate in space (again, this is made more precise in Chapter 4).

The coordinates of stars (or other celestial objects) in the right ascension system are the celestial coordinates: *declination* and *right ascension*. Very much analogous to the spherical coordinates of latitude and longitude on the Earth, the declination, δ, is the angle in the plane of the hour circle from the equatorial plane to the object; and the right ascension, α, is the angle in the equatorial plane from the vernal equinox, counterclockwise (as viewed from the NCP), to the hour circle of the object (despite the same notation, no confusion should arise between right ascension and azimuth). For geodetic applications, these coordinates for stars and other celestial objects are assumed given. Since the right ascension system is fixed in space, so are the coordinates of objects that are fixed in space. However, stars actually do have small lateral motion in this system and this must be known for precise work (see Section 4.2.1).

For later reference, we also define the *ecliptic system* which is a right-handed system with the same first axis (vernal equinox) as the right ascension system. Its third axis, however, is the north ecliptic pole. Coordinates in this system are the *ecliptic latitude* (angle in the *ecliptic meridian* from the ecliptic plane to the celestial object), and the *ecliptic longitude* (angle in the ecliptic plane from the vernal equinox to the ecliptic meridian of the celestial object).

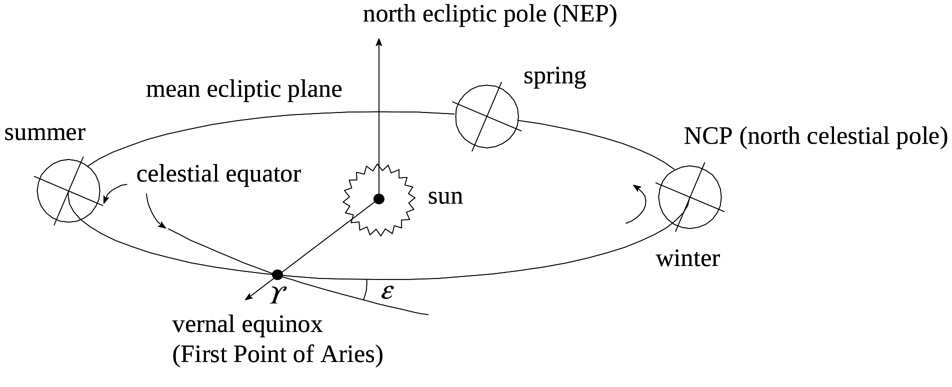


Figure 2.25: Mean ecliptic plane (seasons are for the northern hemisphere).

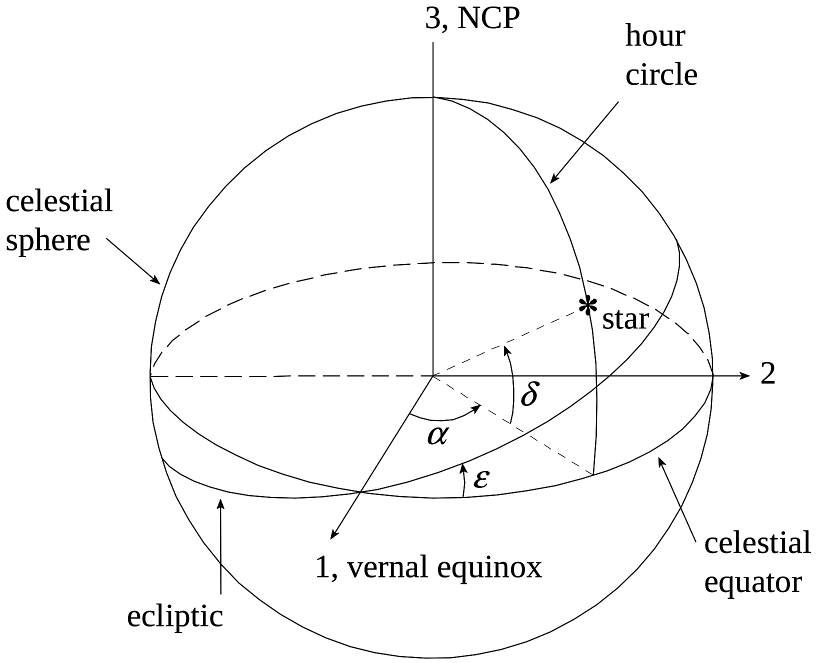


Figure 2.26: Equatorial, right ascension system.

2.3.3 Equatorial, Hour Angle System

The equatorial, hour angle system of coordinates is introduced as a link between the horizon system, in which observations are made, and the right ascension system, in which coordinates of observed objects are given. As with the previous systems, the hour angle system is defined by naturally occurring directions: the direction of Earth's spin axis (NCP), which is the third axis of the system, and the local direction of gravity, which together with the NCP defines the astronomic meridian plane. The first axis of the system is the intersection of the astronomic meridian plane with the celestial equatorial plane; and, the second axis is perpendicular to the other two axes and positive westward, so as to form a *left-handed* system (Figure 2.27). As in the case of the horizon system, the hour angle system is fixed to the Earth.

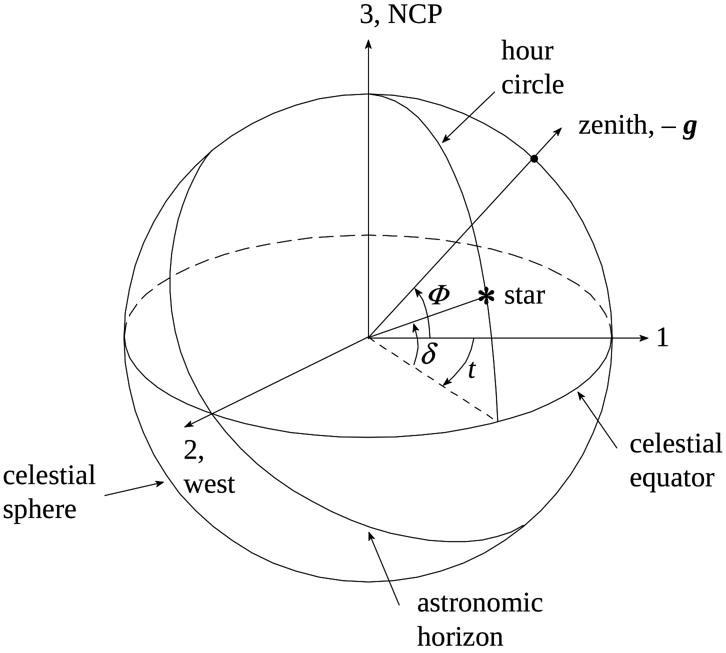


Figure 2.27: Equatorial, hour angle system.

The (instantaneous) coordinates of stars (or other celestial objects) in this system are the *declination* (the same as in the right ascension system) and the *hour angle*. The hour angle, t, that gives this system its name, is the angle in the equatorial plane from the local astronomic meridian to the hour circle of the celestial object. It is reckoned clockwise as viewed from the NCP and increases with time. In fact, it changes by $360°$ with a complete rotation of the Earth relative to inertial space for objects fixed on the celestial sphere (note that the declination remains constant as the Earth rotates – assuming the direction of the spin axis remains fixed; it does not, as discussed in Chapter 4).

2.3.4 Coordinate Transformations

Transformations between coordinates of the horizon and right ascension systems can be accomplished with rotation matrices, provided due care is taken first to convert the left-handed horizon system to a right-handed system. We take another approach that is equally valid and makes use of spherical trigonometry on the celestial sphere. Consider the so-called *astronomic triangle* (Figure 2.28) whose vertices are the three important points on the celestial sphere common to the two systems: the north celestial pole, the zenith, and the star (or other celestial object). It is left to the reader to verify that the labels of the sides and angles of the astronomic triangle, as depicted in Figure 2.28, are correct (the *parallactic angle*, p, is not needed in the present context). Using spherical trigonometric formulas, such as the law of sines, equation

(1.1), and the law of cosines, equation (1.2), it is also left to the reader to show that the following relationship holds:

$$\begin{pmatrix} \sin z \cos A \\ \sin z \sin A \\ \cos z \end{pmatrix} = \begin{pmatrix} -\sin \Phi & 0 & \cos \Phi \\ 0 & -1 & 0 \\ \cos \Phi & 0 & \sin \Phi \end{pmatrix} \begin{pmatrix} \cos \delta \cos t \\ \cos \delta \sin t \\ \sin \delta \end{pmatrix} . \qquad (2.179)$$

The matrix on the right side is orthogonal, so that the following inverse relationship also holds

$$\begin{pmatrix} \cos \delta \cos t \\ \cos \delta \sin t \\ \sin \delta \end{pmatrix} = \begin{pmatrix} -\sin \Phi & 0 & \cos \Phi \\ 0 & -1 & 0 \\ \cos \Phi & 0 & \sin \Phi \end{pmatrix} \begin{pmatrix} \sin z \cos A \\ \sin z \sin A \\ \cos z \end{pmatrix} . \qquad (2.180)$$

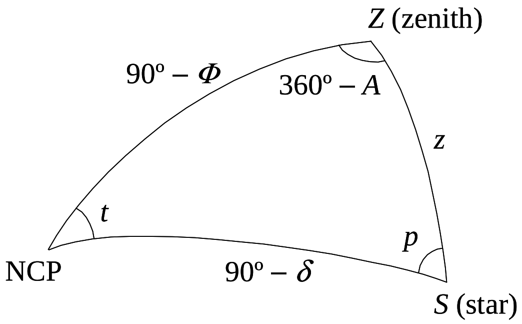


Figure 2.28: Astronomic triangle on the celestial sphere.

Figure 2.29 completes the transformation between systems by showing the relationship between the right ascension and the hour angle. Because the hour angle also is a measure of Earth's rotation with respect to a reference on the celestial sphere, the hour angle is identified with a particular type of time, called *sidereal time* (it is discussed in more detail in Chapter 5 on time systems). We define:

$$t_{\Upsilon} = \text{ hour angle of the vernal equinox} = \textit{local sidereal time } (LST) \ . \qquad (2.181)$$

It is a local time since it applies to the astronomic meridian of the observer. Clearly, from Figure 2.29, the local sidereal time of an observer viewing an arbitrary celestial object with right ascension, α, and hour angle, t, is given by

$$LST = \alpha + t \ . \qquad (2.182)$$

It is noted that 24 hours of sidereal time is the same as 360 degrees of the hour angle. Also, the hour angle of the vernal equinox at the Greenwich meridian, $t_{\Upsilon}^{\mathrm{G}}$, is known as *Greenwich Sidereal Time* (GST).

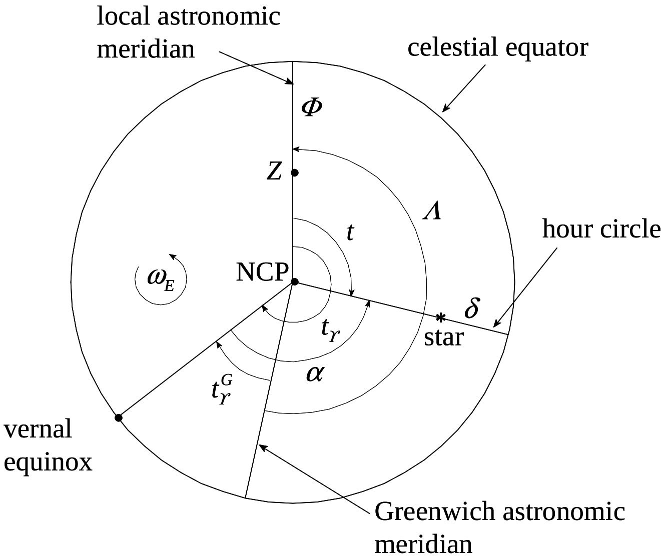


Figure 2.29: Transformation between right ascension and hour angle systems.

2.3.5 Determination of Astronomic Coordinates and Azimuth

The following is a very much abbreviated discussion of the determination of astronomic coordinates, (Φ, Λ), and astronomic azimuth, A, from terrestrial observations of stars. For more details the interested reader is referred to Mueller (1969). In the case of astronomic latitude, Φ, suppose a star crosses the local astronomic meridian of the observer. At the time of transit, the hour angle of the star is $t = 0°$, and according to Figure 2.28, one has for stars passing north and south of the zenith,

$$\begin{aligned} 90° - \Phi &= 90° - \delta_N + z_N \quad \Rightarrow \quad \Phi = \delta_N - z_N \\ 90° - \delta_S &= 90° - \Phi + z_S \quad \Rightarrow \quad \Phi = \delta_S + z_S \end{aligned} \tag{2.183}$$

where δ_N, δ_S and z_N, z_S refer to the corresponding declinations and zenith angles. The declinations of the stars are assumed given and the zenith angles are measured. Combining these, the astronomic latitude of the observer is given by

$$\Phi = \frac{1}{2}\left(\delta_N + \delta_S\right) - \frac{1}{2}\left(z_N - z_S\right). \qquad (2.184)$$

The reason for including stars on both sides of the zenith is that atmospheric refraction in the observed zenith angle will tend to cancel in the second term in equation (2.184) if the corresponding zenith angles are approximately equal. Also, it can be shown (Problem 2.3.6-2) that knowing the exact location of the astronomic meridian (i.e., knowing that $t = 0°$) is not a critical factor when measuring the zenith angle of a star at its *culmination* (the point of maximum elevation above the horizon, which the star attains as it crosses the meridian).

Determining the astronomic longitude of an observer requires that a reference meridian be established (the reference for latitudes is the equator which is established by nature). Historically, this is the meridian through the Greenwich Observatory near London, England. The longitude of an observer at any other point is simply the difference between LST and GST, each converted to angular measure (see Figure 2.29):

$$\Lambda = LST - GST\ . \qquad (2.185)$$

If one waits until a star crosses the local astronomic meridian, when $t = 0°$, then from equation (2.182), $LST = \alpha$, where the right ascension of the star must be given. Alternatively, using the law of cosines applied to the astronomic triangle (Figure 2.28), one can calculate the hour angle for any sighting of a star by measuring its zenith angle and having already determined the astronomic latitude,

$$\cos t = \frac{\cos z - \sin\Phi \sin\delta}{\cos\Phi \cos\delta}. \qquad (2.186)$$

It can be shown (Problem 2.3.6-3) that errors in the zenith measurement and the astronomic latitude have minimal effect when the star is observed near the *prime vertical*. With t thus calculated, the LST is obtained, again, from equation (2.182) and the known right ascension of the observed star.

Either way, with the hour angle known or calculated, one needs a reference for longitudes, and this is provided by the GST. It means that the observer must have a clock (chronometer) that keeps Greenwich Sidereal Time that is recorded locally at the moment of observation.

The determination of astronomic azimuth is less straightforward and can be accomplished using either a calculation of the hour angle from a time measurement or the measurement of the zenith angle. For the first case, the hour angle, t, of a star can be calculated using equation (2.182), where LST is determined from equation (2.185) based on a previous determination of the observer's longitude and a recording of GST at the moment of observation. Now, from equation (2.179), one has

$$\tan A_S = \frac{\sin t}{\sin \Phi \cos t - \cos \Phi \tan \delta}, \tag{2.187}$$

where A_S is the (instantaneous) astronomic azimuth of the star at the time of observation. The observer's astronomic latitude and, as always, the declination and right ascension of the star are assumed to be.

Alternatively, using a star's observed zenith angle, its astronomic azimuth from the law of cosines applied to the astronomic triangle (Figure 2.28) is given by

$$\cos A_S = \frac{\sin \delta - \sin \Phi \cos z}{\cos \Phi \sin z}. \tag{2.188}$$

This does not require a determination of the hour angle (hence no longitude and recording of GST), but is influenced by refraction errors in the zenith angle measurement.

Of course, z or t and, therefore, A_S will change if the same star is observed at a different time. To determine the astronomic azimuth of a terrestrial target, Q, one first sets up the theodolite so that it sights Q. Then at the moment of observing the star (with the theodolite), the horizontal angle, D, between the target and the vertical circle of the star is also measured. The astronomic azimuth of the terrestrial target is thus given by

$$A_Q = A_S - D. \tag{2.189}$$

Having established the astronomic azimuth of a suitable, fixed target, one has also established, indirectly, the location of the local astronomic meridian – it is the vertical circle at a horizontal angle, A_Q, counterclockwise (as viewed from the zenith) from the target.

2.3.6 Problems

1. Derive equation (2.179).

2. a) Starting with the third component in equation (2.179), and also using the first component, show that (assuming $d\delta = 0$)

$$d\Phi = -\frac{dz}{\cos A} - \tan A \cos\Phi\, dt \,. \tag{2.190}$$

b) Determine the optimal azimuth for the observation of a star so as to minimize the error in calculating the astronomic latitude due to errors in the zenith angle measurement and in the determination of the hour angle.

3. a) As in Problem 2, use equation (2.179) and other relationships from the astronomic triangle to show that

$$dt = -\frac{dz}{\sin A \cos\Phi} - \frac{\cot A}{\cos\Phi} d\Phi \,. \tag{2.191}$$

b) Determine the optimal azimuth for calculating a star’s hour angle so as to minimize the error in calculating the astronomic longitude due to errors in the zenith angle measurement and in the determination of the astronomic latitude.

4. a) As in Problem 2, use equation (2.179) and further trigonometric relations derived from Figure 2.28, to show that

$$dA_S = \frac{\cos p \cos\delta}{\sin z} dt + \cot z \sin A_S \, d\Phi \,, \tag{2.192}$$

where p is the parallactic angle.

b) Determine optimal conditions (declination of the star and azimuth of observation) to minimize the error in the determination of astronomic azimuth due to errors in the calculations of hour angle and astronomic latitude.

5. a) From equation (2.188), show that

$$\sin A_S dA_S = \left(\cot z - \cos A_S \tan\Phi\right) d\Phi - \left(\tan\Phi - \cos A_S \cot z\right) dz \,. \tag{2.193}$$

b) Show that the effect of a latitude error is minimized if the hour angle is $t = 90°$ or $t = 270°$; and that the effect of a zenith angle error is minimized when the parallactic angle is $p = 90°$.

Chapter 3

Terrestrial Reference Systems

Geodetic control at local, regional, national, and international levels has been revolutionized by the advent of satellite systems that provide accurate positioning capability to terrestrial observers at all scales, where, of course, the Global Positioning System (GPS) has had the most significant impact. The terrestrial reference systems and frames for geodetic control have evolved correspondingly over the last several decades. Countries and continents around the world are revising, re-defining, and updating their fundamental networks to take advantage of the high accuracy, the ease of establishing and densifying the control, and critically important, the uniformity of the accuracy and the connectivity of the control that can be achieved basically in a global setting.

These reference systems and their realizations are considered in this chapter, from the traditional to the modern, where it is discovered that the essential concepts hardly vary, but the implementation and utility clearly have changed with the tools that have become available. Even though the traditional geodetic reference systems have been or are in the process of being replaced by their modern counterparts in many economically developed regions, they are still an important component for many other parts of the world. It is important, therefore, to understand them and how they relate to the modern systems.

The starting point is the definition of the *geodetic datum*. Unfortunately, the definition is neither consistent nor explicit in the literature and is now even more confusing vis-à-vis the more precise definitions of reference system and reference frame (Section 1.2). The National Geodetic Survey (NGS, 1986), defines the geodetic datum as "a set of constants specifying the coordinate system used for geodetic control, i.e., for calculating coordinates of points on the Earth." The definition given there continues with qualifications regarding the number of such constants under traditional and modern implementations (which tends to confuse the essential definition and reduces it to specialized cases rather than providing a conceptual foundation). Other sources are less deliberate, and add no clarification. For example, Torge (1991) states that a geodetic datum "defines the orientation of a conventional [coordinate] system with respect to the global X, Y, Z-system, and hence, with respect to the body of the earth." Moritz (1978), the title of his paper notwithstanding, only states that a geodetic datum "is usually defined in terms of five parameters ..."; Ewing and Mitchell (1970) are also vague about the definition: "a geodetic datum is comprised of an ellipsoid of revolution fixed in some manner to the physical earth"; while Bomford (1971) states that a datum is the ellipsoid and/or the three coordinates of an origin point relative to the ellipsoid. Finally, Rapp (1992) attempts to bring some perspective to the definition by giving a "simple definition" for a horizontal datum, which is analogous to the discussion by Moritz.

All of these endeavors to define a geodetic datum are targeted toward the horizontal geodetic datum (i.e., for horizontal geodetic control). A more systematic definition of the geodetic datum is given below with an attempt to relate this to the definition of reference systems and frames given earlier in Section 1.2. The NGS definition, in fact, provides a reasonably good basis. Thus:

A *Geodetic Datum* is a set of parameters and constants that defines a coordinate system, including its origin and (where appropriate) its orientation and scale, in such a way as to make these accessible for geodetic applications.

This general definition may be used as a basis for defining traditional horizontal and vertical datums. It conforms to the rather vaguely stated definitions found in the literature (quoted above) and certainly to the concepts of the traditional datums established for geodetic control. Note, however, that the definition alludes to both the definition of a *system* of coordinates and its realization, that is, the *frame* of coordinates. Conceptually, the geodetic datum defines a coordinate system, but once the parameters that constitute a particular datum are specified, it takes on the definition of a frame. Because of the still wide usage of the term, we continue to talk about the geodetic datum as defined above, but realize that a more proper foundation of coordinates for geodetic control is provided by the definitions of reference system and reference frame. In fact, the word "datum" by itself still formally connotes the definition of parameters for the origin, orientation, and scale of a system, and thus is more closely associated with its frame. Indeed, the IERS extends the datum to include also temporal rates of change of these fundamental parameters (see Section 3.3).

It is now a simple matter to define the traditional geodetic datum for horizontal and vertical control:

A *horizontal geodetic datum* is a geodetic datum for horizontal geodetic control in which points are mapped onto a specified ellipsoid.

A *vertical geodetic datum* is a geodetic datum for vertical geodetic control in which points are mapped to the geopotential with a specified geoid.

The horizontal datum is two-dimensional in the sense that two coordinates, latitude and longitude, are necessary and sufficient to identify a point in the network; however, the geometry of the surface on which these points are mapped is such that its realization, or accessibility, requires a three-dimensional conceptualization. The vertical datum, on the other hand, is one-dimensional and requires the value of only a single parameter, the origin point, to be realizable. Vertical datums are discussed only briefly in this text (however, see Section 3.5).

3.1 Horizontal Geodetic Datum

The definition of any terrestrial coordinate system requires the specification of its origin and its orientation with respect to the Earth. If geodetic coordinates are used one must specify in addition the ellipsoid to which they refer. For three-dimensional systems, it is seen later that scale is also important; however, for horizontal systems describing only the angles, latitude and longitude, the coordinate system scale is not as critical since it is basically associated with heights. The scale parameters associated with horizontal distance measurements are part of the instrument error models, not part of the coordinate system scale. Therefore, the definition of the traditional horizontal geodetic datum is based on *eight* parameters: three to define its origin, three to define its orientation, and two to define the ellipsoid. More than that, however, the definition of the *datum* requires that these coordinate system attributes be accessible; that is, for its practical utilization, the coordinate system must be realizable as a frame.

The origin could be defined by placing the ellipsoid center at the center of mass of the Earth. This very natural definition had one important defect before the existence of observable artificial satellites — this origin was not accessible with sufficient accuracy. In addition, the ellipsoid thus positioned relative to the Earth rarely "fit" the region in which geodetic control was to be established. A good fit means that the ellipsoid surface should closely approximate a regional reference surface for heights - the *geoid*, or approximately mean sea level. This was important in the past since observations on the surface of the Earth need to be reduced to the ellipsoid, and the height required to do this was only known (measurable) with respect to the geoid. Therefore, a good fit of the ellipsoid to the geoid implied that the difference between these two surfaces regionally was not as important, or might be neglected, in the reduction of observations. Nevertheless, it should be recognized that the neglect of the geoid, even with a good fit, can produce systematic errors of the order of a meter, or more, that with today's accuracy requirements certainly are very significant.

This alternative definition of the "origin", so as to provide a good local fit, places the ellipsoid with respect to the Earth such that a specific point on the Earth's surface has given (i.e., specified or defined) geodetic coordinates. This *datum origin point*, also called the *initial datum point*, is then obviously accessible – it is a monumented marker on the Earth's surface (see Figure 3.1).

The coordinates, $\left(\phi_0, \lambda_0, h_0\right)$, of the origin point can be chosen arbitrarily, but usually they are determined under an imposed additional condition that the separation between the ellipsoid and the geoid in the particular region should be minimized. In the former case, one could choose

$$\phi_0 = \Phi_0, \quad \lambda_0 = \Lambda_0, \quad h_0 = H_0 , \tag{3.1}$$

where H_0 is the height of the origin point above the geoid (the *orthometric height*); this is a measurable quantity, again defined by nature. Recalling equations (2.167) and (2.169), repeated here for convenience,

$$\begin{aligned} \xi &= \Phi - \phi \\ \eta &= \left(\Lambda - \lambda\right)\cos\phi \end{aligned} \tag{3.2}$$

this choice for the origin point coordinates defines the deflection of the vertical at this point to be zero (the normal to the ellipsoid is tangent to the plumb line at this point). The ellipsoid/geoid separation (the *geoid height*, or *geoid undulation*, N_0) at this one point is also zero (Figure 3.2). Alternatively, one could also specify particular values for the deflection of the vertical and the geoid undulation at the origin point, ξ_0, η_0, N_0. Then the geodetic latitude, longitude, and ellipsoidal height of the origin point are given by

$$\phi_0 = \Phi_0 - \xi_0, \quad \lambda_0 = \Lambda_0 - \frac{\eta_0}{\cos\phi_0}, \quad h_0 = H_0 + N_0 . \tag{3.3}$$

Whether the origin point coordinates are defined by equation (3.1) or by equation (3.3), the assumption is that the geodetic and astronomic systems are parallel, because this was assumed in deriving the first two equations (Section 2.2.3).

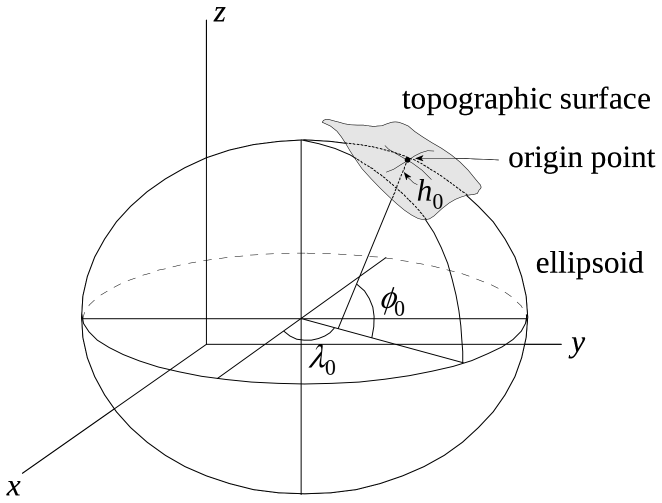


Figure 3.1: Datum origin point.

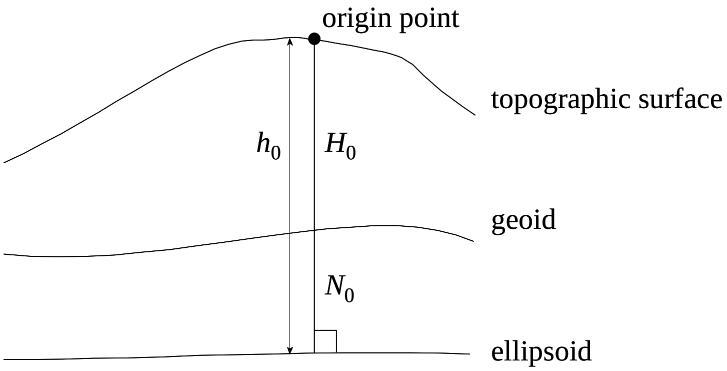


Figure 3.2: Geoid undulation, N_0, at the origin point, in general.

Indeed, the only logical definition of the orientation of the datum is to make the ellipsoid axes parallel to the fundamental astronomic (conventional terrestrial reference) system (Section 2.2); and, this is how the orientation is always *defined*. The three parameters associated with the orientation are the angles, $\omega_x, \omega_y, \omega_z$, between the ellipsoidal and the x, y, z-axes of the astronomic system; their values are zero in the case of parallelism,

$$\omega_x = 0, \quad \omega_y = 0, \quad \omega_z = 0\,. \tag{3.4}$$

The *definition* of orientation is thus simple enough, but the practical realization of this condition is less straightforward. Section 2.2.3 developed the relationships between astronomic and geodetic quantities under the assumption that the two systems are parallel and that, basically, they are concentric (i.e., the placement of the origin was considered to have no effect). In particular, and in addition to equations (3.2), it was found that the astronomic and geodetic azimuths are related by Laplace's condition (2.176),

$$A-\alpha=\left(\Lambda-\lambda\right)\sin\phi+\left(\left(\Phi-\phi\right)\sin\alpha-\left(\Lambda-\lambda\right)\cos\phi\cos\alpha\right)\cot z\,. \tag{3.5}$$

Not only are these necessary conditions for parallelism, they are also sufficient. In other words, using equations (3.2) and (3.5) to relate astronomic and geodetic coordinates and azimuth for points in a networks ensures (in theory) that the two systems are parallel.

To show sufficiency, suppose that the two systems are not parallel. Then each of the equations (3.2) and (3.5) would contain additional terms involving the angles $\omega_x, \omega_y, \omega_z$ It is outside the scope of this exposition to derive the following formulas; however, they may be found, in some fashion, in (Heiskanen and Moritz, 1967, p.213) and (Pick et al., 1973, p.436); see also the analogous polar motion equations for the astronomic coordinates and azimuth (Section 4.3.1). Neglecting second-order terms in the small rotation angles, $\omega_x, \omega_y, \omega_z$, the geodetic coordinates and azimuth become

$$\phi_{\text{rot}}=\phi-\omega_x\sin\lambda+\omega_y\cos\lambda\,, \tag{3.6}$$

$$\lambda_{\text{rot}}=\lambda+\left(\omega_x\cos\lambda+\omega_y\sin\lambda\right)\tan\phi-\omega_z\,, \tag{3.7}$$

$$\alpha_{\text{rot}}=\alpha+\left(\omega_x\cos\lambda+\omega_y\sin\lambda\right)\sec\phi\,, \tag{3.8}$$

where ϕ,λ and α refer to the geodetic coordinates and azimuth at a point for the non-rotated ellipsoid, and $\phi_{\text{rot}},\lambda_{\text{rot}}$ and α_{rot} are corresponding quantities for the same point when the ellipsoid is rotated about its center. The astrogeodetic deflections at a given point with respect to a rotated ellipsoid then become

$$\xi_{\text{rot}}=\Phi-\phi_{\text{rot}}-\omega_x\sin\lambda+\omega_y\cos\lambda\,, \tag{3.9}$$

$$\eta_{\text{rot}}=\left(\Lambda-\lambda_{\text{rot}}\right)\cos\phi+\left(\omega_x\cos\lambda+\omega_y\sin\lambda\right)\sin\phi-\omega_z\cos\phi\,; \tag{3.10}$$

and the azimuth, α_{rot}, with respect to the rotated meridian becomes

$$\begin{aligned}\alpha_{\text{rot}}&=A-\left(\Lambda-\lambda_{\text{rot}}\right)\sin\phi-\left(\left(\Phi-\phi_{\text{rot}}\right)\sin\alpha-\left(\Lambda-\lambda_{\text{rot}}\right)\cos\phi\cos\alpha\right)\cot z\\&\quad-\left(\left(-\omega_x\sin\lambda+\omega_y\cos\lambda\right)\sin\alpha-\left(\left(\omega_x\cos\lambda+\omega_y\sin\lambda\right)\sin\phi-\omega_z\cos\phi\right)\cos\alpha\right)\cot z\\&\quad+\left(\omega_x\cos\lambda+\omega_y\sin\lambda\right)\cos\phi+\omega_z\sin\phi\end{aligned} \tag{3.11}$$

To first order, the non-parallelism effect is independent of a possible origin off-set, $(\Delta x, \Delta y, \Delta z)$. Substituting equations (3.6) and (3.7) into equations (3.9) and (3.10), it is found that $\xi_{\text{rot}} = \Phi - \phi = \xi$ and $\eta_{\text{rot}} = (\Lambda - \lambda)\cos\phi = \eta$. Thus, the deflection of the vertical does not change at a point (in first-order approximation) due to a small rotation of the ellipsoid. But it does change if computed using determined quantities, $\phi_{\text{rot}}, \lambda_{\text{rot}}$, in a rotated geodetic system, as seen in equations (3.9) and (3.10). Substituting $\xi_{\text{rot}} = \Phi - \phi_{\text{rot}}$, $\eta_{\text{rot}} = (\Lambda - \lambda_{\text{rot}})\cos\phi_{\text{rot}}$, and $\alpha_{\text{rot}} = A - (\Lambda - \lambda_{\text{rot}})\sin\phi - ((\Phi - \phi_{\text{rot}})\sin\alpha - (\Lambda - \lambda_{\text{rot}})\cos\phi\cos\alpha)\cot z$ on the left sides of equations (3.9), (3.10), and (3.11), respectively, yields three equations in $\omega_x, \omega_y, \omega_z$ that for arbitrary points can only be satisfied if these rotations are zero. Therefore, the use of (3.2) and (3.5) is also sufficient to yield parallelism of the systems.

As an aside, the deflections and azimuth are directly sensitive to a displacement of the ellipsoid, since the ellipsoid normal through a point then changes direction. Assuming only a displacement and no rotation, equations (3.2) and (3.5) hold for the new deflection components and the geodetic azimuth. Neglecting effects due to the ellipsoidal eccentricity (i.e., using the mean Earth radius, R, equation (2.64)), it can be shown that (Heiskanen and Moritz 1967, p.207)

$$\begin{aligned}\xi_{\text{dis}} &= \Phi - \phi_{\text{dis}} \\ &= \Phi - \phi + \sin\phi\left(\frac{\Delta x}{R}\cos\lambda + \frac{\Delta y}{R}\sin\lambda\right) - \frac{\Delta z}{R}\cos\phi\end{aligned} \tag{3.12}$$

$$\begin{aligned}\eta_{\text{dis}} &= (\Lambda - \lambda_{\text{dis}})\cos\phi \\ &= (\Lambda - \lambda)\cos\phi + \left(\frac{\Delta x}{R}\sin\lambda - \frac{\Delta y}{R}\cos\lambda\right)\end{aligned} \tag{3.13}$$

$$\begin{aligned}\alpha_{\text{dis}} &= A - (\Lambda - \lambda_{\text{dis}})\sin\phi - ((\Phi - \phi_{\text{dis}})\sin\alpha - (\Lambda - \lambda_{\text{dis}})\cos\phi\cos\alpha)\cot z \\ &= \alpha + \tan\phi\left(-\frac{\Delta x}{R}\sin\lambda + \frac{\Delta y}{R}\cos\lambda\right) \\ &\quad - \left(\left(\sin\phi\left(\frac{\Delta x}{R}\cos\lambda + \frac{\Delta y}{R}\sin\lambda\right) - \frac{\Delta z}{R}\cos\phi\right)\sin\alpha - \left(\frac{\Delta x}{R}\sin\lambda - \frac{\Delta y}{R}\cos\lambda\right)\cos\alpha\right)\cot z\end{aligned} \tag{3.14}$$

where $\phi_{\text{dis}}, \lambda_{\text{dis}}$ are geodetic coordinates that refer to an ellipsoid with its center displaced by $(\Delta x, \Delta y, \Delta z)$ from the geocenter.

Of particular importance in realizing the parallelism of the horizontal datum relative to the astronomic system is the determination of the geodetic azimuth of a target, Q, from the origin point according to equation (3.5),

$$\alpha_{0,Q} = A_{0,Q} - (\Lambda_0 - \lambda_0)\sin\phi - \left((\Phi_0 - \phi_0)\sin\alpha_{0,Q} - (\Lambda_0 - \lambda_0)\cos\phi_0 \cos\alpha_{0,Q}\right)\cot z_{0,Q}, \qquad (3.15)$$

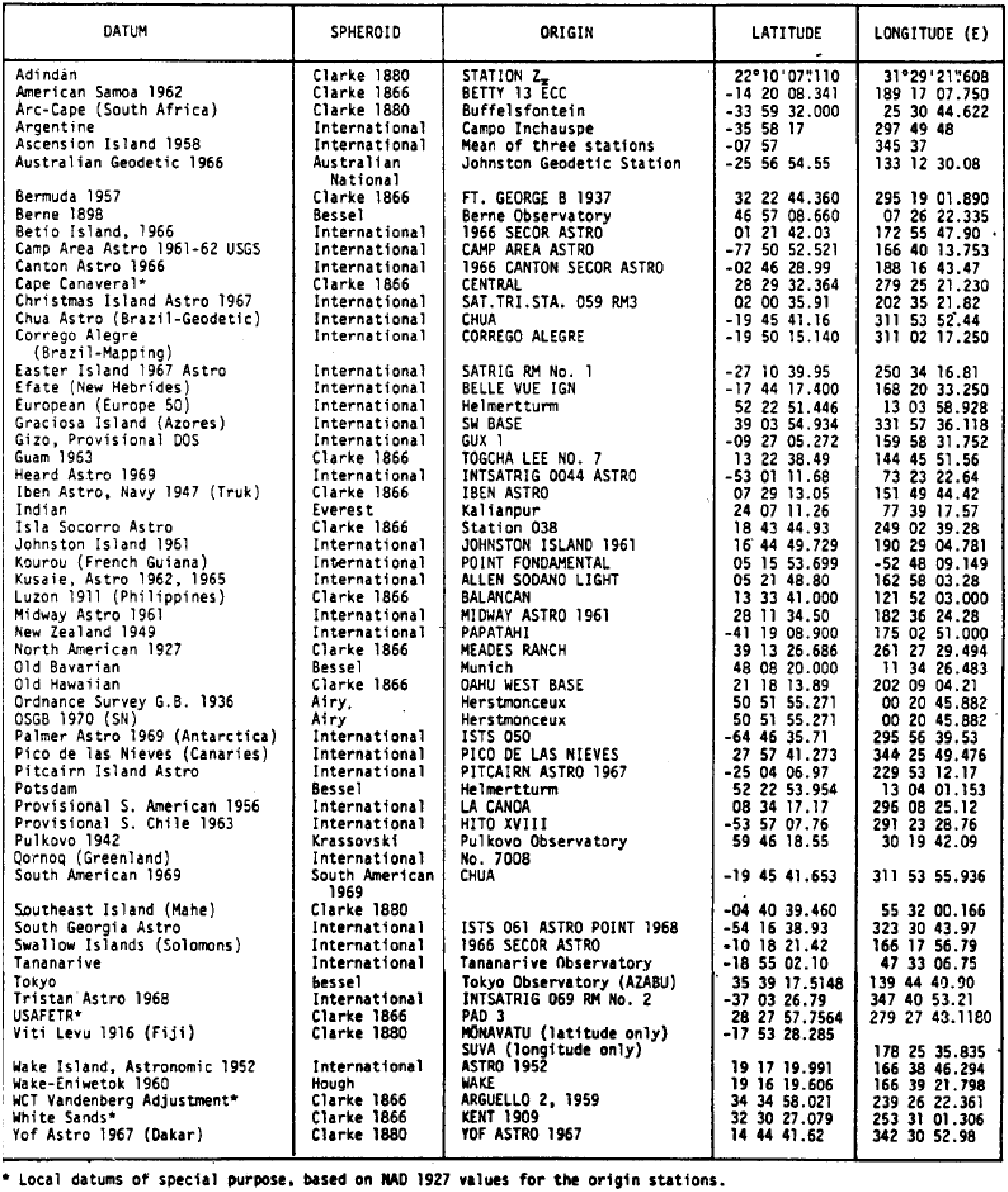
where the coordinates, ϕ_0, λ_0, have already been chosen, and the quantities, $\Phi_0, \Lambda_0, A_{0,Q}$, have been observed (i.e., they are not arbitrary, but are defined by nature); see also Section 2.2.3. The zenith angle, $z_{0,Q}$, is also obtained by observation. It is sometimes stated that the Laplace azimuth, $\alpha_{0,Q}$, at the origin is a parameter of the horizontal geodetic datum. However, we see with equation (3.15), that, in fact, this is not a parameter in the sense that it is given an arbitrarily specified value. Only by *computing* the geodetic (Laplace) azimuth according to equation (3.15) (in general at other points, equation (3.5)) can one be assured that the datum is realized as being parallel to the astronomic system. In theory, only one Laplace azimuth in a geodetic network is necessary to ensure the parallelism; but, in practice, several are interspersed throughout the region to reduce the effect of observation error (Moritz, 1978). That is, a single error in azimuth propagates in a systematic way through the network, causing significant rotational distortions, unless controlled by other azimuth observations and correspondingly computed Laplace azimuths elsewhere in the network.

To summarize, the horizontal geodetic datum as a reference *system* is defined as a system of coordinates referring to an ellipsoid, with specified parameters (e.g., a, f), whose origin is fixed to the Earth in some prescribed way (e.g, by “attaching” the ellipsoid to a monument on the Earth’s surface), and whose orientation is defined with respect to the astronomic system, always by equation (3.4). The datum as a reference *frame* is realized by the three origin point coordinates (as illustrated above), and by the three orientation parameters indirectly through the utilization of equations (3.2) and (3.5) at all points in the network where astronomic observations are related to geodetic quantities. Here, the azimuth plays the most critical role in datum orientation.

3.1.1 Examples of Horizontal Geodetic Datums

Table 3.1, taken from (Rapp, 1992), lists many of the horizontal geodetic datums of the world (not all are still in service). NIMA (1997) also lists over 100 datums (however, without datum origin point parameters). Note that the datum origin coordinates (Table 3.1) were chosen either according to equations (3.1) or (3.3), or by minimizing the deflections or the geoid undulations (geoid heights) over the region of horizontal control; or, they were simply adopted from a previous network adjustment. Again, it is beyond the present scope to explore the details of these minimization procedures and adjustments.

Table 3.1: Selected Horizontal Geodetic Datums (NASA 1978)



DATUM	SPHEROID	ORIGIN	LATITUDE	LONGITUDE (E)
Adindán	Clarke 1880	STATION Z	22°10'07".110	31°29'21".608
American Samoa 1962	Clarke 1866	BETTY 13 ECC	-14 20 08.341	189 17 07.750
Arc-Cape (South Africa)	Clarke 1880	Buffelsfontein	-33 59 32.000	25 30 44.622
Argentine	International	Campo Inchauspe	-35 58 17	297 49 48
Ascension Island 1958	International	Mean of three stations	-07 57	345 37
Australian Geodetic 1966	Australian National	Johnston Geodetic Station	-25 56 54.55	133 12 30.08
Bermuda 1957	Clarke 1866	FT. GEORGE B 1937	32 22 44.360	295 19 01.890
Berne 1898	Bessel	Berne Observatory	46 57 08.660	07 26 22.335
Betío Island, 1966	International	1966 SECOR ASTRO	01 21 42.03	172 55 47.90
Camp Area Astro 1961-62 USGS	International	CAMP AREA ASTRO	-77 50 52.521	166 40 13.753
Canton Astro 1966	International	1966 CANTON SECOR ASTRO	-02 46 28.99	188 16 43.47
Cape Canaveral*	Clarke 1866	CENTRAL	28 29 32.364	279 25 21.230
Christmas Island Astro 1967	International	SAT.TRI.STA. 059 RM3	02 00 35.91	202 35 21.82
Chua Astro (Brazil-Geodetic)	International	CHUA	-19 45 41.16	311 53 52.44
Corrego Alegre (Brazil-Mapping)	International	CORREGO ALEGRE	-19 50 15.140	311 02 17.250
Easter Island 1967 Astro	International	SATRIG RM No. 1	-27 10 39.95	250 34 16.81
Efate (New Hebrides)	International	BELLE VUE IGN	-17 44 17.400	168 20 33.250
European (Europe 50)	International	Helmertturm	52 22 51.446	13 03 58.928
Graciosa Island (Azores)	International	SW BASE	39 03 54.934	331 57 36.118
Gizo, Provisional DOS	International	GUX 1	-09 27 05.272	159 58 31.752
Guam 1963	Clarke 1866	TOGCHA LEE NO. 7	13 22 38.49	144 45 51.56
Heard Astro 1969	International	INTSATRIG 0044 ASTRO	-53 01 11.68	73 23 22.64
Iben Astro, Navy 1947 (Truk)	Clarke 1866	IBEN ASTRO	07 29 13.05	151 49 44.42
Indian	Everest	Kalianpur	24 07 11.26	77 39 17.57
Isla Socorro Astro	Clarke 1866	Station 038	18 43 44.93	249 02 39.28
Johnston Island 1961	International	JOHNSTON ISLAND 1961	16 44 49.729	190 29 04.781
Kourou (French Guiana)	International	POINT FONDAMENTAL	05 15 53.699	-52 48 09.149
Kusaie, Astro 1962, 1965	International	ALLEN SODANO LIGHT	05 21 48.80	162 58 03.28
Luzon 1911 (Philippines)	Clarke 1866	BALANCAN	13 33 41.000	121 52 03.000
Midway Astro 1961	International	MIDWAY ASTRO 1961	28 11 34.50	182 36 24.28
New Zealand 1949	International	PAPATAHI	-41 19 08.900	175 02 51.000
North American 1927	Clarke 1866	MEADES RANCH	39 13 26.686	261 27 29.494
Old Bavarian	Bessel	Munich	48 08 20.000	11 34 26.483
Old Hawaiian	Clarke 1866	OAHU WEST BASE	21 18 13.89	202 09 04.21
Ordnance Survey G.B. 1936	Airy,	Herstmonceux	50 51 55.271	00 20 45.882
OSGB 1970 (SN)	Airy	Herstmonceux	50 51 55.271	00 20 45.882
Palmer Astro 1969 (Antarctica)	International	ISTS 050	-64 46 35.71	295 56 39.53
Pico de las Nieves (Canaries)	International	PICO DE LAS NIEVES	27 57 41.273	344 25 49.476
Pitcairn Island Astro	International	PITCAIRN ASTRO 1967	-25 04 06.97	229 53 12.17
Potsdam	Bessel	Helmertturm	52 22 53.954	13 04 01.153
Provisional S. American 1956	International	LA CANOA	08 34 17.17	296 08 25.12
Provisional S. Chile 1963	International	HITO XVIII	-53 57 07.76	291 23 28.76
Pulkovo 1942	Krassovski	Pulkovo Observatory	59 46 18.55	30 19 42.09
Qornoq (Greenland)	International	No. 7008		
South American 1969	South American 1969	CHUA	-19 45 41.653	311 53 55.936
Southeast Island (Mahe)	Clarke 1880		-04 40 39.460	55 32 00.166
South Georgia Astro	International	ISTS 061 ASTRO POINT 1968	-54 16 38.93	323 30 43.97
Swallow Islands (Solomons)	International	1966 SECOR ASTRO	-10 18 21.42	166 17 56.79
Tananarive	International	Tananarive Observatory	-18 55 02.10	47 33 06.75
Tokyo	Bessel	Tokyo Observatory (AZABU)	35 39 17.5148	139 44 40.90
Tristan Astro 1968	International	INTSATRIG 069 RM No. 2	-37 03 26.79	347 40 53.21
USAFETR*	Clarke 1866	PAD 3	28 27 57.7564	279 27 43.1180
Viti Levu 1916 (Fiji)	Clarke 1880	MÖNAVATU (latitude only)	-17 53 28.285	
		SUVA (longitude only)		178 25 35.835
Wake Island, Astronomic 1952	International	ASTRO 1952	19 17 19.991	166 38 46.294
Wake-Eniwetok 1960	Hough	WAKE	19 16 19.606	166 39 21.798
WCT Vandenberg Adjustment*	Clarke 1866	ARGUELLO 2, 1959	34 34 58.021	239 26 22.361
White Sands*	Clarke 1866	KENT 1909	32 30 27.079	253 31 01.306
Yof Astro 1967 (Dakar)	Clarke 1880	YOF ASTRO 1967	14 44 41.62	342 30 52.98

* Local datums of special purpose, based on NAD 1927 values for the origin stations.

3.1.2 Problems

1. Describe a step-by-step procedure to compute the geodetic latitudes and longitudes of points in a network of measured horizontal angles and straight-line distances. Use diagrams and flowcharts to show how the coordinates could be computed from the coordinates of other points and the measurements (hint: direct problem!). Assume that the astronomic coordinates are observed at every point, but that the astronomic azimuth is observed only at the origin point. We already discussed all corrections needed to transform observed azimuths to *geodesic* azimuths; assume similar procedures exist to transform straight-line distances and angles to geodesic distances and angles between points on the ellipsoid. For helpful discussions of this problem, see (Moritz 1978).

2. a) The software for a GPS receiver gives positions in terms of geodetic latitude, longitude, and height above the ellipsoid GRS80 (the ellipsoid for WGS84). For $\phi = 40^\circ$, $\lambda = -83^\circ$, and $h = 200\text{ m}$, compute the equivalent (x, y, z) coordinates of the point in the corresponding Cartesian coordinate system.

b) Compute the geodetic coordinates (ϕ, λ, h) of that point in the NAD27 system, assuming that it, like GRS80, is geocentric (which it is not!).

c) Now compute the coordinates (ϕ, λ, h) of that point in the NAD27 system, knowing that the center of the NAD27 ellipsoid is offset from that of the WGS84 ellipsoid by $x_{\text{WGS84}} - x_{\text{NAD27}} = -4\text{ m}$, $y_{\text{WGS84}} - y_{\text{NAD27}} = 166\text{ m}$, $z_{\text{WGS84}} - z_{\text{NAD27}} = 183\text{ m}$. Compare your result with 2.b).

3. Suppose the origin of a horizontal datum is defined by a monumented point on the Earth's surface.

a) The deflection of the vertical at the origin point is *defined* to be zero. If the geodetic coordinates of the point are $\phi = 40^\circ$ and $\lambda = -83^\circ$, what are the corresponding astronomic latitude and longitude at this point? What assumptions about the orientation of the datum does this involve?

b) Suppose the ellipsoid of the datum is shifted in the z-direction by 4 m, which datum parameters will change, and by how much (give an estimate for each one based on geometrical considerations; i.e., draw a figure showing the consequence of a change in the datum)?

3.2 Geodetic Control in the U.S. (and North America)

Each datum in the world has a history that reflects the economic development of the region. In the U.S., national geodetic control is the responsibility of the National Geodetic Survey (NGS, part of NOAA, the National Oceanic and Atmospheric Administration, under the Department of Commerce); in Canada, this responsibility falls to the Geodetic Survey Division of the Department of Natural Resources (Natural Resources Canada). The North American Datum interestingly chronicles the westward expansion and globalization from its initial definition for the eastern U.S. to the present-day definition. The New England Datum of 1879 used the Clarke 1866 ellipsoid with origin point at Station Principio in Maryland. This datum was adopted for the entire country as the U.S. Standard Datum of 1901 soon after the trans-continental triangulation was completed, 1871-1897 (32 years after the completion of the trans-continental railroad in 1869!). In 1909 the datum origin was chosen to be at Meades Ranch, Kansas, upon an adjustment of the coordinates to fit the observed deflections of the vertical at hundreds of points throughout the country. When Canada and Mexico adopted this datum for their triangulations in 1913, it became the North American Datum.

In 1927, a major re-adjustment of the horizontal networks across the continent was undertaken by holding the coordinates at Meades Ranch fixed. However, these coordinates have no special significance in the sense of equations (3.1) or (3.3), being simply the determined coordinates in the previous triangulations and adjustments. The datum was named the North American Datum of 1927 (NAD27). The orientation of the datum was controlled by numerous Laplace stations throughout the network. It was estimated later with new satellite observations that the orientation was accurate to about 1 arcsec (Rapp, 1992, p.A-6). The adjustment was done in parts, primarily treating the western and eastern parts of the country separately. Errors were distributed by the residuals between observed astronomic and geodetic latitude, longitude, and azimuth along survey triangulation arcs, much like leveling residuals are distributed along leveling loops. Geoid undulations were kept small in this way, since, in essence, this amounts to a minimization of the deflections, which is equivalent to minimizing the slope of the geoid relative to the ellipsoid, and thus minimizing the variations of the geoid undulation over the network. Even though the new, more representative International Ellipsoid (Table 2.1) was available, based on Hayford’s 1909 determinations, the Clarke Ellipsoid of 1866 was retained for the datum since it was used for most of the computations over the preceding years.

In the reduction of coordinates of points in NAD27 to the ellipsoid, the geoid undulation was neglected, and thus all lengths technically refer to the geoid and not the ellipsoid, or conversely, the ellipsoid distances have a systematic error due to this neglect. This error manifested itself regionally as distortions of relative positions separated by several hundreds and thousands of kilometers within the network. Similarly, most angles were not corrected for the deflection of the vertical and were reduced to the ellipsoid as if they were turned about the ellipsoid normal. These approximate procedures and other deficiencies in the adjustment caused distortions of sections of NAD27 (i.e., locally) up to 1 part in 15,000 (1 m over 15 km)!

Because of its realization, fundamentally at a terrestrial monument, the NAD27 ellipsoid is not geocentric. This was the situation for all datums in the world prior to the use of satellites for geodetic positioning. However, once satellites entered the picture, it was possible to realize the $(0,0,0)$ origin of a datum at Earth's center, recognizing that satellites orbit around the center of mass of the Earth. Of course, this realization of the origin is indirect and is subject to errors in determining the satellite orbit and other observational errors. In addition to the new satellite data for point positioning, extensive gravity observations in North America (particularly the U.S., propelled by the search for oil) yielded good models for the geoid undulation and the deflection of the vertical. Also, early satellite altimetry and satellite perturbation analyses yielded much better values for Earth's size and its dynamic flattening.

Hence, in the 1970's and 1980's a major re-adjustment, as well as a *re-definition*, of the North American Datum was undertaken. The ellipsoid was changed to that of the Geodetic Reference System 1980 (GRS80) and was assumed to be geocentric (*system* definition). That is, the Meades Ranch station was abandoned as the origin point in favor of the geocenter (center of mass of the Earth). This geocentric realization was achieved by satellite Doppler observations which yield three-dimensional coordinates of points with respect to the centroid of the satellite orbits (i.e., the center of mass). Although astronomic observations of azimuth still served to realize the orientation of the new datum, specifically the *z*-axis rotation angle (ω_z), the satellite observations could now also provide orientation, especially the other angles, ω_x and ω_y. In addition, very long-baseline interferometry (VLBI) began to deliver very accurate orientation on a continental scale. Since geoid undulations could now be estimated with reasonable accuracy, they were used in all reductions of distances and angles to the ellipsoid. This was, in fact, an important element of the re-adjustment, since now the ellipsoid/geoid separation was not minimized in any way. The geoid undulation over the conterminous U.S. varies between about -7 m (southern Montana and Wyoming) and -37 m (over the Great Lakes). The result of this vast re-adjustment and re-definition was the North American Datum of 1983 (NAD83). For further details of the re-adjustment, the reader is directed to Schwarz (1989) and Schwarz and Wade (1990).

New realizations of NAD83 (now viewed as a 3-D reference system) were achieved with satellite positioning techniques, originally the Doppler-derived positions, but then with satellite and lunar laser ranging, and significantly with the Global Positioning System (GPS) that all provided increased accuracy of the origin and orientation. The NAD83(1986) realization is based on a transformation of the Doppler station coordinates by a 4.5 m translation in the z-direction, a 0.814 arcsec rotation about the z-axis, and a scale change of -0.6 ppm. Improvements in the realization continued with High-Accuracy Regional Networks (HARNs) derived from GPS, where the realizations NAD83(HARN) (1989 - 1997) changed the scale by -0.0871 ppm, but retained the known origin and orientation offsets of approximately 2 m and 0.03 arcsec, respectively, from the geocenter and the origin point for longitudes as realized by observations using satellite and space techniques (see also Table 3.3). Nationally, new realizations of NAD83 made use of the Continuously Operating Reference Stations (CORS), based on GPS, throughout the U.S., yielding NAD83(CORS93), NAD83(CORS94), and NAD83(CORS96) with each new adjustment. In all these realizations, the origin and orientation of the NAD83(1986) frame were, again, basically retained. Further realizations that re-adjusted the HARNs as close as possible to NAD83(CORS96) are designated NAD83(NSRS2007), where NSRS stands for the *National Spatial Reference System* and represents the fundamental geodetic control in the United States in all dimensions (horizontal and vertical) and aspects (such as providing accurate control of shorelines). A subsequent reprocessing of all CORS station data from 1994 to 2011 resulted in the realization, NAD83(2011), with published coordinates given for the epoch 2010.0.

The National Geodetic Survey (NGS) has planned[1,2,3] a modernization of the NSRS within the next decade that is based on yet another "paradigm shift" in terms of defining and realizing the coordinate systems. The many conventional, "passive," fixed benchmarks that surveyors have employed for centuries to access the coordinate frame will no longer be maintained by NGS and will not form the primary control. Instead, the NAD83, already viewed as a three-dimensional system, will be replaced by a system that is defined and actively maintained using Global Navigation Satellite Systems (GNSS). These include firstly GPS, but also the Russian GLONASS (GLObal'naya NAvigatsionnaya Sputnikovaya Sistema), the European Galileo System, the Chinese BeiDou (Compass) System, and others as they come on line. The system definitions of origin, orientation, and scale now will be the same as for the International Terrestrial Reference System. The realization will be actively maintained using an extensive foundational CORS network that is accurately tied to the International Terrestrial Reference Frame (Section 3.3). Thus, the 2-meter origin offset will finally disappear and the system will be truly geocentric. NGS will make available mathematical tools (software accessible on the internet, similar to the current Online Positioning User Service, OPUS[4]) that allow users to obtain coordinates for any point for which they can provide GNSS (e.g., GPS) data. In this way, the user community will be responsible for any local monumentation of control; and, all such control will be tied unambiguously and with precision defined by the user to the national CORS network. The motivation behind this planned mode of operation is the realization that permanently emplaced monuments on the Earth's surface can no longer be viewed as associated with constant coordinates. Plate tectonics, subsidence, and other deformation of the crust due to natural and anthropogenic causes make this concept obsolete at the centimeter level of precision. In fact, NGS has already been migrating to this new mode by providing coordinates of CORS

[1] NGS (2008). The National Geodetic Survey Ten-Year Plan, Mission, Vision and Strategy, 2008-2018. http://www.ngs.noaa.gov/INFO/NGS10yearplan.pdf

[2] Proceedings of the 2010 Federal Geospatial Summit on Improving the National Spatial Reference System. http://www.ngs.noaa.gov/2010Summit/2010FederalGeospatialSummitProceedings.pdf

[3] Report from the 2015 Geospatial Summit on Improving the National Spatial Reference System. http://www.geodesy.noaa.gov/2015GeospatialSummit/ReportFromThe2015GeospatialSummitv7.pdf

[4] www.ngs.noaa.gov/OPUS/

sites at a near current epoch (e.g., 2010.0) together with velocities due to known motions within the frame of NAD83(2011) (see Section 3.4.1).

With the replacement of NAD83, NGS also plans to replace the vertical datum NAVD88 (Section 3.5) by a geopotential model, where again the control is achieved actively without the need, at least on a national level, to maintain passive markers.

3.3 International Terrestrial Reference System

The international efforts to define a terrestrial system can be traced back to the turn of the last century (1900's) when the International Latitude Service (ILS) (that was established in 1899 by the International Association of Geodesy (IAG)) organized observations of astronomic latitude in order to detect and monitor the motion of the pole (Section 4.3.1). The ILS was reorganized into the International Polar Motion Service (IPMS) in 1962 by resolution of the International Astronomical Union (IAU); and, the IPMS officially continued the work of the ILS. Also, the Rapid Latitude Service (RLS) of the Bureau International de l'Heure (BIH) in Paris, France, was established in 1955 again by the IAU, and predicted coordinates of the instantaneous pole and served primarily to help in the time keeping work of the BIH[5]. In addition, the U.S. Navy and the Defense Mapping Agency (U.S.) published polar motion results based on the latest observing technologies (such as lunar laser ranging (LLR) and very long baseline interferometry (VLBI)).

In 1960, it was decided finally at the General Assembly of the International Union of Geodesy and Geophysics (I.U.G.G.) to adopt as terrestrial pole the average of the true celestial pole during the period 1900-1905 (a six-year period over which the Chandler period of 1.2 years would repeat five times; see Section 4.3.1). This average was named the *Conventional International Origin* (CIO) starting in 1968 (not to be confused with the Celestial Intermediate Origin, Chapter 4). Even though more than 50 observatories ultimately contributed to the determination of the pole through latitude observations, the CIO was defined and monitored by the original 5 latitude observatories under the original ILS (located approximately on the 39th parallel; including Gaithersburg, Maryland; Ukiah, California, Carloforte, Italy; Kitab, former U.S.S.R.; and Mizusawa, Japan).

The reference meridian was originally defined as the astronomic meridian through the Greenwich observatory, near London, England. However, from the 1950s until the 1980s, the BIH monitored the variation in longitudes (due to polar motion and variations in Earth's spin rate, or length-of-day) of many observatories (about 50) and a mean "Greenwich" meridian was defined, based on an average of zero-meridians, as implied by the variation-corrected longitudes of these observatories. The basis of these longitudes ultimately was an accurate determination of time as observed from Earth rotation ($UT1$, Chapter 5).

[5] For a history of the BIH, see (Guinot 2000).

These early conventions and procedures to define and realize a terrestrial reference system addressed astronomic *directions* only; no attempt was made to define a realizable origin, although implicitly it could be thought of as being geocentric. From 1967 until 1988, the BIH was responsible for determining and monitoring the CIO and reference meridian. In 1979 the BIH Conventional Terrestrial System (CTS) replaced the 1968 BIH system with a better reference to the CIO. However, the CIO as originally defined was not entirely satisfactory because it could be accessed only through 5 latitude observatories. As of 1984, the BIH defined the BIH CTS (or BTS) based on satellite laser ranging, VLBI, and other space techniques. The alignment was based on maintaining continuity in the astronomic longitude origin for time determination based on Earth rotation (mean solar time corrected for polar motion, UT1; Chapter 5). In addition, the mean pole during 1980-1983 as monitored by the BIH served to define the third axis of the BTS84. With the inclusion of satellite observations, an (indirectly) accessible origin of the system could also be defined (geocentric). As new and better satellite and VLBI observations became available from year to year, the BIH published new realizations of its system: BTS84, BTS85, BTS86, and BTS87.

One consequence of defining a *geocentric*, essentially geometric, system not based on astronomic observations, but on the systems realized by satellite and space techniques, is that the geodetic longitude origin does not coincide with the astronomic longitude origin. As noted above, the primary consideration by BIH for defining the orientation of the BTS was that the historical time system should remain unbroken (continuity in the Earth-rotation based time, UT1). Thus, the astronomic longitude origin was maintained approximately at the Greenwich Observatory. However, in aligning the BTS frame to the frames realized by the satellite and space techniques, which presumably maintained an orientation to previously determined geodetic coordinates that only needed to be corrected for a geocentric translational offset, the geodetic longitude of the Greenwich Observatory is not zero, but deviates from the zero astronomic longitude by the east component of the deflection of the vertical at the Observatory. This was confirmed recently by Malys et. al (2015), who also document the history of the Greenwich longitude. Figure 3.3 adapted from Bomford (1971, p.) (and Malys et al., 2015) shows the geometric relationships between astronomic and geodetic systems that result in the non-zero (geocentric) geodetic longitude at the Greenwich Observatory. Indeed, the geodetic longitude there in today's geocentric terrestrial reference frame is $-5.31''$; or, the geodetic zero meridian is about 102 m to the east of the astronomic meridian at the Greenwich Observatory.

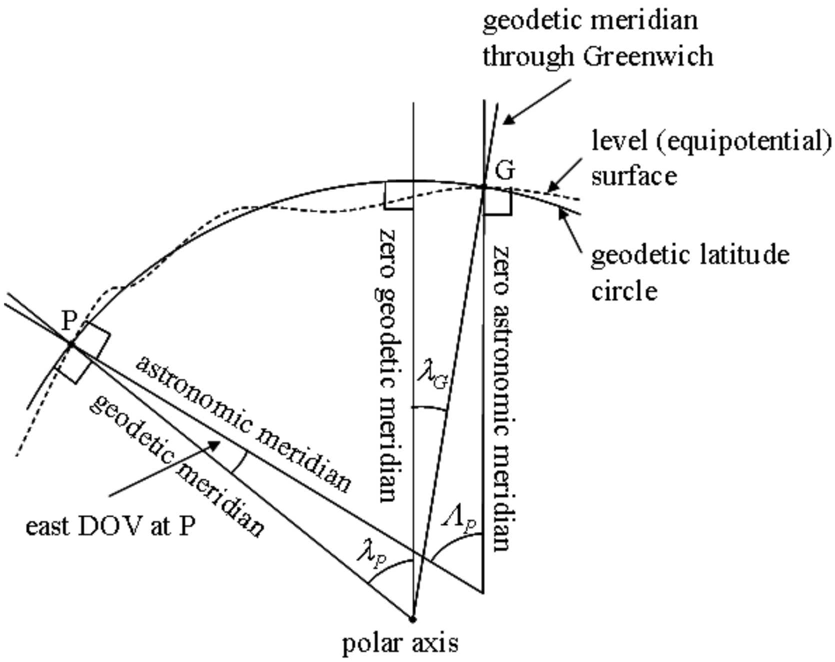


Figure 3.3: Relative geometry of geodetic and astronomic meridians at Greenwich (G) and another point, P, on the same latitude circle as G. The polar axis is the semi-minor axis of a geocentric ellipsoid. Note that the zero meridians in the geodetic and astronomic systems are parallel. The geodetic and astronomic longitudes of P would be equal if the east deflection of the vertical (DOV) at P were zero.

In 1988 the functions of monitoring the pole and the reference meridian were turned over to the newly established *International Earth Rotation Service* (IERS), thus replacing the BIH and the IPMS as the corresponding service organizations. The time service, originally also under the BIH, now resides with the Bureau International des Poids et Mésures (BIPM). The IERS, renamed in 2003 to *International Earth Rotation and Reference Systems Service* (retaining the same acronym), is responsible for defining and realizing both the *International Terrestrial Reference System* (ITRS) and the *International Celestial Reference System* (ICRS). In each case, an origin, an orientation, and a scale are defined among other conventions for the system. The system is then realized as a frame by the specification of these datum parameters and the coordinates of points worldwide. Since various observing systems (analysis centers and techniques) contribute to the overall realization of the reference system and since new realizations are obtained recurrently with improved observation techniques and instrumentation, the transformations among various realizations are of paramount importance. Especially, if one desires to combine data referring to realizations of different reference systems, or to different realizations of the same system, it is important to understand the coordinate relationships so that the data are combined ultimately in one consistent coordinate system. We first continue this section with a description of the ITRS and its realization and treat transformations in the next section.

The IERS International Terrestrial Reference System is defined by an orthogonal triad of right-handed, equally scaled axes with the following additional conventions:

a) The *origin* is geocentric, that is, at the center of mass of the Earth (including the mass of the oceans and atmosphere). Nowadays, because of the capability to detect the small (cm-level) variations due to terrestrial mass re-distributions, the origin is defined as an average location of the center of mass and referred to some epoch.

b) The *scale* is defined by the speed of light in vacuum and the time interval corresponding to one second (see Chapter 5) within the theory of general relativity and in the local Earth frame.

c) The *orientation* is defined by the directions of the CIO and the geodetic reference meridian as given for 1984 by the BIH. These principal directions are now called the IRP (International Reference Pole) and the IRM (International Reference Meridian) (also, ITRF Zero Meridian). Since it is now well established that Earth's crust (on which the observing stations are located) is divided into plates that exhibit tectonic motion (of the order of centimeters per year), it is further stipulated that the time evolution of the orientation of the reference system has no residual global rotation with respect to the crust ("no-net-rotation" condition). That is, even though the points on the crust, through which the system is realized, move with respect to each other, the net rotation of the system with respect to its initial definition should be zero.

The realization of the ITRS is the International Terrestrial Reference Frame (ITRF) and requires that three origin parameters, three orientation parameters, and a scale parameter must be identified with actual values. These seven parameters are not observable without conventions (see below) and their specification is formulated by the IERS in terms of constraints imposed on the solution of coordinates from observations. Moreover, the constraints are cast in the form of a seven-parameter transformation (see Section 3.4) from an a priori defined frame to the realized frame: three translation parameters that realize the origin; three angle parameters that realize the orientation, and a scale change parameter that realizes the scale. As a simple example (which is not practiced anymore), suppose a previous frame contains a point with defined coordinates (analogous to the Meades Ranch origin point, but known to refer to the geocenter). The next realization, based on new observations, could be related to the previous frame by constraining the translation to be zero. Because these datum (transformation) parameters are determined for points on the Earth's crust ("crust-based frame"), and because the Earth as a whole is a dynamic entity, the parameters are associated with an epoch and, today, are supplemented with rates of change, making the total number of parameters equal to 14.

Unlike the origin of the historical (traditional) geodetic datum that could be accessed at a physical point on the Earth, the geocenter is accessible only indirectly by dynamical modeling of satellite orbits and observations of distances relative to the satellites in these orbits. In either case, however, whether a marker on the Earth's surface or its geocenter, the origin is defined by a convention, just like all other parts of the coordinate system. As such it is not, a priori, an observable, or measurable, quantity like a distance or an angle. This is the classic *datum defect* problem, well known in all types of surveying, where observations of distances and angles must ultimately be *related* to a point or direction that is fixed or defined by convention.

With satellite techniques, on the other hand, there is the advantage of knowing that the center of mass is the centroid for all orbits. In that sense, the center of mass of the Earth serves as a natural origin point that, in theory, is accessible. That is, if the orbit is known, observations (e.g., distances) from points on the Earth's surface to points on the orbit are in a geocentric system, by definition. Determining the orbit by dynamical methods (using and/or solving for the gravitational field of the Earth, as well as other forces acting on the satellite) is beyond the present scope (Seeber 2003). Suffice it to say that not all origin realizations are the same as obtained by different analysis centers that, moreover, process different satellite data (satellite laser ranging, lunar laser ranging, GPS, Doppler data). Generally, the most precise methods are based on satellite laser ranging (SLR).

For the first ITRFs in the early 1990s, it was customary to relate all frames realized by particular analysis centers and/or satellite techniques to one of the satellite laser ranging (SLR) solutions from the Center for Space Research (CSR) in Austin, Texas, which was considered to be the best solution that accesses the center of mass and thus realizes the origin. The origins of solutions (i.e., realized coordinate systems) from other techniques, such as Doppler and GPS, were related by IERS to the ITRF origin through a translation determined by using stations that are common to both the CSR and the other solutions. Later, a weighted average of selected SLR and GPS solutions was used to realize the origin. For ITRF2000, the origin was realized by a weighted average of "the most consistent SLR solutions" submitted to the IERS (Petit and Luzum 2010). With ITRF2005 and ITRF2008, the IERS used a time series over 13 years and 26 years, respectively, of re-processed SLR data at selected, globally distributed sites to realize the origin. The latest realization, ITRF2014, follows the same procedures as for ITRF2005 and ITR2008, reprocessing all data up to 2014 and providing also enhanced models for post-seismic deformation at earthquake-prone sties.

The scale similarly was realized for the early ITRFs by the SLR solutions from the CSR analysis center, with the scale of other solutions transformed accordingly. For all subsequent realizations of scale, SLR was combined with Very Long Baseline Interferometry (VLBI), which accurately measures coordinate differences of stations separated by large distances (several 1000 km) using observed directions to quasars (Chapter 4). (It is noted that VLBI provides no information on the origin of coordinates.)

Satellite and space observational techniques contain no information on the absolute longitudinal orientation of a system. This orientation has no obvious natural reference and is completely arbitrary (the Greenwich meridian). One might argue that the equatorial orientation (or, equivalently, the polar direction) like the center of mass is a natural reference that is accessible indirectly from astronomic observations, VLBI, and satellite tracking (since the orbit is also defined by the figure axis of the Earth, see Section 4.3.2). However, the polar direction is complicated, a result of both polar motion with respect to the Earth's crust, and precession and nutation with respect to the celestial sphere (see Chapter 4). Besides this, the stations on the Earth's crust, which ultimately realize the ITRS, are in constant motion due to plate tectonics. Thus, the adopted convention for realizing the orientation of the ITRS is to ensure that each successive realization after 1984 is aligned with the orientation defined by the BIH in 1984 (with some early adjustments for different solutions of the Earth Orientation Parameters (Chapter 4).

The methods of combining different solutions and introducing the constraints needed to address the datum defect (i.e., specifying origin, scale, and orientation) has become increasingly complicated as more data are assimilated and analysis centers employ various weighting schemes to account for the various observational accuracies. These details are beyond the present scope and the interested reader is referred to the IERS Conventions of 2003 (McCarthy and Petit 2003) and of 2010 (Petit and Luzum 2010) and references therein (specifically also publications by Altamimi et al. 2002a,b, and references therein).

The model for the coordinates of any of the observing stations participating in the realization of ITRS is given by

$$\boldsymbol{x}(t) = \boldsymbol{x}_0 + (t - t_0)\boldsymbol{v}_0 + \sum_i \Delta\boldsymbol{x}_i(t), \tag{3.16}$$

where $\boldsymbol{x}_0$ and $\boldsymbol{v}_0$ are the coordinates and their velocities for the observing station, defined for a particular epoch, t_0. These are solved on the basis of observed coordinates, $\boldsymbol{x}(t)$, at time, t, using some type of observing system (like satellite laser ranging). The quantities, $\Delta\boldsymbol{x}_i$, are corrections applied by analysis centers to account for various, short-wavelength, local geodynamic effects, such as solid Earth tides, ocean loading, and atmospheric loading, with the objective of accounting for the non-constant velocities. Details for corresponding recommended models are provided by the IERS Conventions 2010 (Chapter 7). The coordinate vector, $\boldsymbol{x}_0$, and the linear velocity, $\boldsymbol{v}_0$, for each participating station is provided by IERS as a result of the assimilation of all data, and these represent the consequent realization of ITRS at epoch, t_0. In the past, the linear velocity was modeled largely by the tectonic plate motion model, NNR-NUVEL1A (McCarthy, 1996); thus,

$$\boldsymbol{v}_0 = \boldsymbol{v}_{\mathrm{NUVEL1A}} + \delta\boldsymbol{v}_0, \tag{3.17}$$

where $\boldsymbol{v}_{\mathrm{NUVEL1A}}$ is the velocity given as a set of rotation rates for the major tectonic plates, and $\delta\boldsymbol{v}_0$ is a residual velocity for the station. The newest ITRFs (since ITRF2000) appear to indicate significant departures of the station velocities, $\boldsymbol{v}_0$, from the NNR-NUVEL1A model, which, however, does not impact the integrity of the ITRF.

3.3.1 World Geodetic System of the U.S. Department of Defense

The World Geodetic System 1984 (WGS84) is the equivalent of the ITRS for the U.S. Department of Defense (and includes also a global gravitational model). It is the evolution of previous reference systems, WGS60, WGS66, and WGS72 (DMA 1987). The corresponding reference frame for WGS84 as originally realized in 1987 on the basis mostly of satellite Doppler observations agreed approximately with NAD83. The next realization, designated WGS84(G730), made use of observations from 12 GPS stations around the world and was aligned with the ITRF92 to an accuracy of about 20 cm in all coordinates. Here, G730 denotes the 730th week of the GPS satellite ephemerides. The next realization, WGS84(G873), improved on this and was designed to be consistent with ITRF94, which was achieved with about 10 cm accuracy. The next realization, WGS84(G1150), was based on GPS observations at 17 U.S. Air Force and NIMA (National Imagery and Mapping Agency)[6] stations, and it is consistent with ITRF2000 at the 2 cm level of accuracy (Merrigan et al. 2002)[7]. Finally, the latest realization, WGS84(G1674), is actually adjusted to be consistent with ITRF2008. That is, all the WGS84 stations adopted their ITRF2008 coordinates and velocities for epoch 2005.0[8] (Wong et al. 2012).

3.4 Transformations

With many different realizations of terrestrial reference systems, as well as local or regional datums, it is important in geodetic applications to know the relationship between the coordinates of points in these frames. Especially for the realization of ITRF, extensive use is made of transformations to define the evolution of the realizations and the relationships of ITRF to the realizations of reference systems of contributing analysis centers or space techniques. The transformations of traditional local horizontal datums (referring to an ellipsoid) with respect to each other and with respect to a global terrestrial reference frame is a topic beyond the present scope. However, for standard Cartesian systems, like the ITRS and WGS84, and even the new realizations of the NAD83 and other modern realizations of regional datums (like the European Coordinate Reference Systems[9]), a simple 7-parameter similarity transformation (*Helmert transformation*) serves as the basic model for the transformations.

According to the definition of the IERS, this transformation model is given by

[6] Renamed in 2003 to National Geospatial-Intelligence Agency (NGA)

[7] https://www.ion.org/publications/abstract.cfm?jp=p&articleID=2164

[8] http://www.unoosa.org/pdf/icg/2012/template/WGS_84.pdf

[9] http://www.euref.eu/

$$\boldsymbol{x}_{\text{to}} = \boldsymbol{T} + \left(1 + D\right) R^{\text{T}} \boldsymbol{x}_{\text{from}} , \qquad (3.18)$$

where $\boldsymbol{x}_{\text{to}}$ is the coordinate vector of a point in the frame *to* which its coordinates are transformed, and $\boldsymbol{x}_{\text{from}}$ is the vector of coordinates of that same point in the frame *from* which it is transformed. (Perhaps it is not the best notation, but it is the clearest in defining the direction of the transformation, and the reader is cautioned not to confuse "to" with the epoch, t_0.) The translation, or displacement, between frames is given by the vector, $\boldsymbol{T}$, and the scale difference is given by D. The IERS definition concerning the rotations between frames is somewhat counter-intuitive, where the rotation matrix, here denoted R^{T}, represents rotations angles in the negative (clockwise) sense, rather than the usual positive (counterclockwise) sense; see Figure 3.4. Since the rotation angles are small, we have from equation (1.9):

$$R^{\text{T}} = R_1^{\text{T}}\left(R1\right) R_2^{\text{T}}\left(R2\right) R_3^{\text{T}}\left(R3\right) = \begin{pmatrix} 1 & -R3 & R2 \\ R3 & 1 & -R1 \\ -R2 & R1 & 1 \end{pmatrix}, \qquad (3.19)$$

where $R1$, $R2$, and $R3$ are the small rotation angles, in the notation and definition of the IERS.

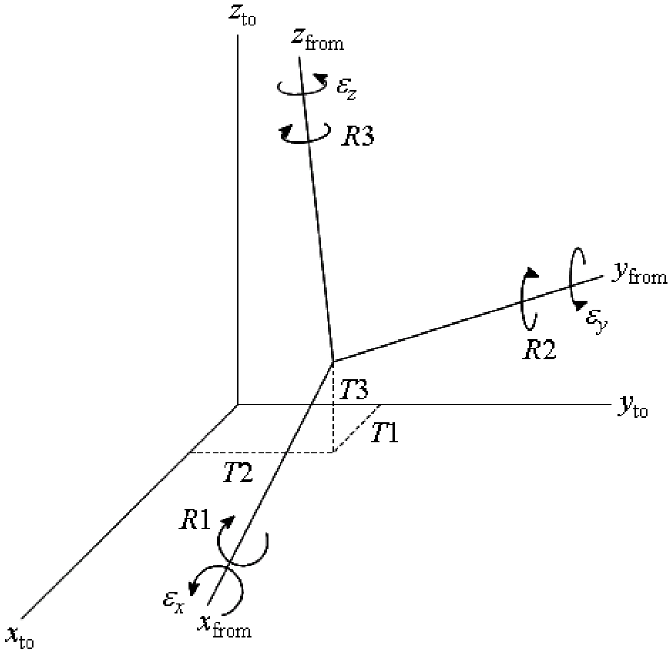


Figure 3.4: Transformation parameters for the IERS and the NGS models. IERS and NGS conventions are illustrated with $R1, R2, R3$ and $\varepsilon_x, \varepsilon_y, \varepsilon_z$, respectively.

Since D is also a small quantity, we can neglect second-order terms and write

$$\begin{pmatrix} x \\ y \\ z \end{pmatrix}_{\text{to}} = \begin{pmatrix} x \\ y \\ z \end{pmatrix}_{\text{from}} + \begin{pmatrix} T1 \\ T2 \\ T3 \end{pmatrix} + D \begin{pmatrix} x \\ y \\ z \end{pmatrix}_{\text{from}} + \begin{pmatrix} 0 & -R3 & R2 \\ R3 & 0 & -R1 \\ -R2 & R1 & 0 \end{pmatrix} \begin{pmatrix} x \\ y \\ z \end{pmatrix}_{\text{from}} . \qquad (3.20)$$

Each of the seven parameters of this model, $T1$, $T2$, $T3$, $R1$, $R2$, $R3$, and D, may have a time variation that is modeled simply as being linear,

$$\beta_i(t) = \beta_{0i} + \dot{\beta}_{0i}(t - t_0), \qquad (3.21)$$

where β_i refers to any of the parameters. The 14 parameters, β_{0i} and $\dot{\beta}_{0i}$, $i = 1, \ldots, 7$, then constitute the complete transformation. Combining equations (3.20) and (3.21), we have

$$\begin{pmatrix} x(t) \\ y(t) \\ z(t) \end{pmatrix}_{\text{to}} = \begin{pmatrix} x(t) \\ y(t) \\ z(t) \end{pmatrix}_{\text{from}} + \begin{pmatrix} T1(t) \\ T2(t) \\ T3(t) \end{pmatrix} + D(t) \begin{pmatrix} x(t) \\ y(t) \\ z(t) \end{pmatrix}_{\text{from}} + \begin{pmatrix} 0 & -R3(t) & R2(t) \\ R3(t) & 0 & -R1(t) \\ -R2(t) & R1(t) & 0 \end{pmatrix} \begin{pmatrix} x(t) \\ y(t) \\ z(t) \end{pmatrix}_{\text{from}} , \qquad (3.22)$$

noting that the transformation, as given by the parameters, $\beta_i(t)$, is valid at a particular epoch, *t*.

Thus, transformations among the various frames require careful consideration not only of velocities of the points themselves in a particular frame, but also the "velocity" of the transformation parameters, themselves. In general, the transformation of coordinates of a point between two different frames may proceed in two ways. Suppose the point coordinates in the "from"-frame vary in time (within that frame),

$$\boldsymbol{x}_{\text{from}}(t) = \boldsymbol{x}_{\text{from}}(t_0) + \dot{\boldsymbol{x}}_{\text{from}} \cdot (t - t_0), \qquad (3.23)$$

where the coordinate velocity, $\dot{\boldsymbol{x}}_{\text{from}}$, and the coordinates at t_0, both in the "from"-frame, are known. Further, suppose that the transformation parameters between the "from"- and "to"-frames area given by equation (3.21). Then the coordinates of this point in the "to"-frame at the epoch, t, can be computed either according to

$$\begin{array}{ccccc} \boldsymbol{x}_{\text{from}}(t_0) & \rightarrow & \boldsymbol{x}_{\text{to}}(t_0) & \rightarrow & \boldsymbol{x}_{\text{to}}(t) \\ & \uparrow & & \uparrow & \\ & \boldsymbol{\beta}_0 & & \dot{\boldsymbol{x}}_{\text{to}} & \end{array} \qquad (3.24)$$

or

$$\begin{array}{ccccc} \boldsymbol{x}_{\text{from}}\left(t_0\right) & \rightarrow & \boldsymbol{x}_{\text{from}}\left(t\right) & \rightarrow & \boldsymbol{x}_{\text{to}}\left(t\right) \\ & \uparrow & & \uparrow & \\ & \dot{\boldsymbol{x}}_{\text{from}} & & \boldsymbol{\beta}\left(t\right) & \end{array} \qquad (3.25)$$

where the vertical arrows indicate either Helmert transformations and simple velocity transformations, given by equation (3.23). The velocity, $\dot{\boldsymbol{x}}_{\text{to}}$, can be determined from the derivative of equation (3.22),

$$\begin{pmatrix} \dot{x} \\ \dot{y} \\ \dot{z} \end{pmatrix}_{\text{to}} = \begin{pmatrix} \dot{x} \\ \dot{y} \\ \dot{z} \end{pmatrix}_{\text{from}} + \begin{pmatrix} \dot{T}1 \\ \dot{T}2 \\ \dot{T}3 \end{pmatrix} + \dot{D} \begin{pmatrix} x(t) \\ y(t) \\ z(t) \end{pmatrix}_{\text{from}} + \begin{pmatrix} 0 & -\dot{R}3 & \dot{R}2 \\ \dot{R}3 & 0 & -\dot{R}1 \\ -\dot{R}2 & \dot{R}1 & 0 \end{pmatrix} \begin{pmatrix} x(t) \\ y(t) \\ z(t) \end{pmatrix}_{\text{from}} , \qquad (3.26)$$

which neglects second-order terms. The transformation methods described by equations (3.24) and (3.25) are equivalent if the velocities, $\dot{\boldsymbol{\beta}}$, $\dot{\boldsymbol{x}}_{\text{from}}$, $\dot{\boldsymbol{x}}_{\text{to}}$ of/within the frames are accurately related according to equation (3.26).

Table 3.2 lists the transformation parameters among the various IERS (and BIH) terrestrial Reference Frames since 1984. [These numbers were obtained from various IERS publications and internet sites and have been known to contain some inconsistencies (see also the ITRF internet site[10])]. Rates of the parameters are given only since 1993. Note that ITRF96 and ITRF97 were defined to be identical to ITRF94 with respect to epoch 1997. In order to obtain transformation parameters for other than the listed epoch, equation (3.21) should be employed. For example, using the last row of Table 3.2, the translation in *x* between ITRF2005 and ITRF2008 at the epoch, $t = 2000$, is given by

$$\begin{aligned} T1(t) &= T1(t_0) + \dot{T}1 \cdot (t - t_0) \\ &= 0.05 \text{ cm} - 0.03 \text{ cm/yr} \cdot (-5 \text{ yr}) \\ &= 0.20 \text{ cm} \end{aligned} \qquad (3.27)$$

The Petit and Luzum (2010) provide transformation parameters *from* ITRF2008 to all previous frames for the epoch, $t = 2000$; see also the IERS internet site for parameter values corresponding to the most recent (and previous) realization.

[10] http://itrf.ensg.ign.fr/

Table 3.3 lists transformation parameters from WGS84 to ITRF90 as published by McCarthy (1992) as well as from recent ITRFs to NAD83(CORS96) as published by the National Geodetic Survey·. These are no longer available on the internet; individual transformations may be found in the literature; e.g., Soler and Snay (2004). Note that the rotation parameters, $\varepsilon_{1,2,3}$, in Table 3.3 represent the more intuitive rotations from the *from*-frame to the *to*-frame. Also, note that the transformation parameters formally are estimates with given associated standard deviations (they are not listed here). Therefore, the determination of the vector of coordinates through such a transformation, in principle, should include a rigorous treatment of the propagation of errors.

Table 3.2: Transformation parameters for recent terrestrial reference frames.

From	To	$T1\,\|\,\dot{T}1$	$T2\,\|\,\dot{T}2$	$T3\,\|\,\dot{T}3$	$R1\,\|\,\dot{R}1$	$R2\,\|\,\dot{R}2$	$R3\,\|\,\dot{R}3$	$D\,\|\,\dot{D}$	t_0
		cm cm/yr	cm cm/yr	cm cm/yr	0.001" 0.001"/ yr	0.001" 0.001"/ yr	0.001" 0.001"/ yr	10^{-8} 10^{-8}/yr	
BTS84	BTS85	5.4	2.1	4.2	–0.9	–2.5	–3.1	-0.5	1984
BTS85	BTS86	3.1	–6.0	–5.0	–1.8	–1.8	–5.81	–1.7	1984
BTS86	BTS87	–3.8	0.3	–1.3	–0.4	2.5	7.5	–0.2	1984
BTS87	ITRF0	0.4	–0.1	0.2	0.0	0.0	–0.2	–0.1	1984
ITRF0	ITRF88	0.7	–0.3	–0.7	–0.3	–0.2	–0.1	0.1	1988
ITRF88	ITRF89	0.5	3.6	2.4	–0.1	0.0	0.0	–0.31	1988
ITRF89	ITRF90	–0.5	–2.4	3.8	0.0	0.0	0.0	–0.3	1988
ITRF90	ITRF91	0.2	0.4	1.6	0.0	0.0	0.0	–0.03	1988
ITRF91	ITRF92	–1.1	–1.4	0.6	0.0	0.0	0.0	–0.14	1988
ITRF92	ITRF93	–0.2 –0.29	–0.7 0.04	–0.7 0.08	–0.39 –0.11	0.80 –0.19	–0.96 0.05	0.12 0.0	1988
ITRF93	ITRF94	–0.6 0.29	0.5 –0.04	1.5 –0.08	0.39 0.11	–0.80 0.19	0.96 –0.05	–0.04 0.0	1988
ITRF94	ITRF96	0.0 0.0	0.0 0.0	0.0 0.0	0.0 0.0	0.0 0.0	0.0 0.0	0.0 0.0	1997
ITRF96	ITRF97	0.0 0.0	0.0 0.0	0.0 0.0	0.0 0.0	0.0 0.0	0.0 0.0	0.0 0.0	1997
ITRF97	ITRF2000	–0.67 0.00	–0.61 0.06	1.85 0.14	0.0 0.0	0.0 0.0	0.0 -0.02	–0.155 -0.001	1997
ITRF2000	ITRF2005	-0.01 0.02	0.08 -0.01	0.58 0.18	0.0 0.0	0.0 0.0	0.0 0.0	-0.040 -0.008	2000
ITRF2005	ITRF2008	0.05 -0.03	0.09 0.0	0.47 0.0	0.0 0.0	0.0 0.0	0.0 0.0	-0.094 0.0	2005
ITRF2008	ITRF2014	-0.16 0.00	-0.19 0.00	-0.24 0.01	0.0 0.0	0.0 0.0	0.0 0.0	0.002 -0.003	2010

$$\begin{pmatrix} x \\ y \\ z \end{pmatrix}_{\text{to}} = \begin{pmatrix} x \\ y \\ z \end{pmatrix}_{\text{from}} + \begin{pmatrix} T1(t) \\ T2(t) \\ T3(t) \end{pmatrix} + D(t)\begin{pmatrix} x \\ y \\ z \end{pmatrix}_{\text{from}} + \begin{pmatrix} 0 & -R3(t) & R2(t) \\ R3(t) & 0 & -R1(t) \\ -R2(t) & R1(t) & 0 \end{pmatrix}\begin{pmatrix} x \\ y \\ z \end{pmatrix}_{\text{from}} . \qquad (3.28)$$

Table 3.3: Transformation parameters for other terrestrial reference frames. Note that $\varepsilon_x = -R1$, $\varepsilon_y = -R2$, $\varepsilon_z = -R3$.

From	To	$T1 \mid \dot{T}1$	$T2 \mid \dot{T}2$	$T3 \mid \dot{T}3$	$\varepsilon1 \mid \dot{\varepsilon}1$	$\varepsilon2 \mid \dot{\varepsilon}2$	$\varepsilon3 \mid \dot{\varepsilon}3$	$D \mid \dot{D}$	t_0
		cm cm/yr	cm cm/yr	cm cm/yr	0.001" 0.001"/ yr	0.001" 0.001"/ yr	0.001" 0.001"/ yr	10^{-8} 10^{-8}/yr	
WGS72	ITRF90	-6.0	51.7	472.3	18.3	-0.3	–547.0	23.1	1984
WGS84[1]	ITRF90	-6.0	51.7	22.3	18.3	-0.3	7.0	1.1	1984
ITRF96	NAD83 (CORS96)	99.1 0.0	–190.7 0.0	-51.3 0.0	25.8 0.053	9.7 -0.742	11.7 -0.032	0.0 0.0	1997
ITRF97	NAD83 (CORS96)	98.9 0.07	–190.7 -0.01	-50.3 0.19	25.9 0.067	9.4 -0.757	11.6 -0.031	-0.09 -0.02	1997
ITRF2000	NAD83 (CORS96)	99.6 0.07	–190.1 -0.07	-52.2 0.05	25.9 0.067	9.4 -0.757	11.6 -0.051	0.06 -0.02	1997
IGS08	NAD83 (2011)	99.343 0.079	-190.331 -0.060	-52.655 -0.134	25.91467 0.06667	9.42645 -0.75744	11.59935 -0.05133	0.171504 -0.010201	1997

[1] original realization; sign error for ε_z has been corrected.

$$\begin{pmatrix} x \\ y \\ z \end{pmatrix}_{\text{to}} = \begin{pmatrix} x \\ y \\ z \end{pmatrix}_{\text{from}} + \begin{pmatrix} T1(t) \\ T2(t) \\ T3(t) \end{pmatrix} + D(t) \begin{pmatrix} x \\ y \\ z \end{pmatrix}_{\text{from}} + \begin{pmatrix} 0 & \varepsilon_z(t) & -\varepsilon_y(t) \\ -\varepsilon_z(t) & 0 & \varepsilon_x(t) \\ \varepsilon_y(t) & -\varepsilon_x(t) & 0 \end{pmatrix} \begin{pmatrix} x \\ y \\ z \end{pmatrix}_{\text{from}} . \qquad (3.29)$$

3.4.1 Transformations to and Realizations of NAD83

IAG resolutions (Resolutions Nos.1 and 4; IAG 1992) recommend that regional high-accuracy reference frames be tied to an ITRF, where such frames associated with large tectonic plates may be allowed to rotate with these plates as long as they coincide with an ITRF at some epoch. This procedure was adopted for NAD83, which for the conterminous U.S. and Canada lies (mostly) on the North American tectonic plate. This plate has global rotational motion estimated according to the NNR-NUVEL1A model by the following rates (McCarthy 1996),

$$\begin{aligned} \Omega_x &= 0.000258 \text{ rad/}10^6\text{yr} = 0.053 \text{ mas/yr} = 1.6 \text{ mm/yr} \\ \Omega_y &= -0.003599 \text{ rad/}10^6\text{yr} = -0.742 \text{ mas/yr} = -22.9 \text{ mm/yr} \qquad (3.30) \\ \Omega_z &= -0.000153 \text{ rad/}10^6\text{yr} = -0.032 \text{ mas/yr} = -0.975 \text{ mm/yr} \end{aligned}$$

where the last equality for each rate uses the approximation that the Earth is a sphere with radius, $R = 6371\text{ km}$. These rates are in the same sense as the IERS convention for rotations.

The transformation between a regional frame and ITRF can be determined (using the standard Helmert transformation model) if a sufficient number of points exists in both frames. Such was the case for the transformation between NAD83(HARN) and ITRF93 on the basis of 9 VLBI stations in the U.S. that had accurate 3-D coordinates in both frames (Soler and Snay 2000). The resulting transformation parameters were applied in a transformation of all CORS stations whose coordinates originally were determined in ITRF93, which thus yielded the realization NAD83(CORS93). This procedure was repeated with respect to ITRF94 and ITRF96, using also additional VLBI sites in Canada (Craymer et al., 2000). The solution for the Helmert transformation parameters from ITRF96 to NAD83(CORS96) resulted in (see also Table 3.3):

$$\begin{aligned}
&T1(1997.0) = 0.9910\text{ m}\\
&T2(1997.0) = -1.9072\text{ m}\\
&T3(1997.0) = -0.5129\text{ m}\\
&R1(1997.0) = -25.79\text{ mas}\\
&R2(1997.0) = -9.65\text{ mas}\\
&R2(1997.0) = -11.66\text{ mas}\\
&D(1997.0) = 6.62\text{ ppb}
\end{aligned}\tag{3.31}$$

where the angles refer to the convention used by IERS, and the epoch 1997.0 indicates the epoch of validity of the transformation parameters. The scale factor for these transformations to NAD83 was set to zero ($D(1997.0) = 0$) so that the two frames, by definition, have the same scale. Snay (2003) notes that this is equivalent to determining a transformation in which the transformed latitudes and longitudes of the points in one frame would best approximate the latitudes and longitudes in the other in a least-squares sense. That is, the scale is essentially the height, and the height is, therefore, not being transformed. We thus have

$$\begin{pmatrix} x \\ y \\ z \end{pmatrix}_{NAD83(CORS96)} = \begin{pmatrix} x \\ y \\ z \end{pmatrix}_{ITRF96(1997)} + \begin{pmatrix} T1(1997) \\ T2(1997) \\ T3(1997) \end{pmatrix} + \begin{pmatrix} 0 & -R3(1997) & R2(1997) \\ R3(1997) & 0 & -R1(1997) \\ -R2(1997) & R1(1997) & 0 \end{pmatrix} \begin{pmatrix} x \\ y \\ z \end{pmatrix}_{ITRF96(1997)} \tag{3.32}$$

Since the regional frame, NAD83(HARN), is attached to the North American tectonic plate, the ITRF coordinates of points on that plate have a velocity, while corresponding coordinates in NAD83 have virtually no velocity due to plate motion (unless the point is on another plate; see Problem 3.4.2-3). Now, the transformation parameters, thus determined, refer to a particular epoch (1997.0 in this case). At other epochs, the NAD83 coordinates presumably do not change at the VLBI sites used to determine the transformation parameters; but, their coordinates in the ITRF do change because the North American plate is moving (rotating) in a global frame. Therefore, the transformation between NAD83(CORS96) and ITRF96 at other epochs should account for this motion. For points on the North American plate we may incorporate the plate motion, equations (3.30), into the ITRF transformation from one epoch to the next as

$$\begin{pmatrix} x(t) \\ y(t) \\ z(t) \end{pmatrix}_{ITRF96} = \begin{pmatrix} x(1997) \\ y(1997) \\ z(1997) \end{pmatrix}_{ITRF96} + \begin{pmatrix} 0 & -\Omega_z(t-1997) & \Omega_y(t-1997) \\ \Omega_z(t-1997) & 0 & -\Omega_x(t-1997) \\ -\Omega_y(t-1997) & \Omega_x(t-1997) & 0 \end{pmatrix} \begin{pmatrix} x(1997) \\ y(1997) \\ z(1997) \end{pmatrix}_{ITRF96} , \tag{3.33}$$

where, e.g., both $x(t)$ and $x(1997.0)$ refer to the IRTF96, but at different epochs. Substituting this into the ITRF96-NAD83 transformation, we obtain

$$\begin{pmatrix} x \\ y \\ z \end{pmatrix}_{NAD83(CORS96)} = \begin{pmatrix} x(t) \\ y(t) \\ z(t) \end{pmatrix}_{ITRF96} + \begin{pmatrix} T1 \\ T2 \\ T3 \end{pmatrix} + \begin{pmatrix} 0 & -R3(t) & R2(t) \\ R3(t) & 0 & -R1(t) \\ -R2(t) & R1(t) & 0 \end{pmatrix} \begin{pmatrix} x \\ y \\ z \end{pmatrix}_{ITRF96(1997.0)} , \tag{3.34}$$

where

$$\begin{aligned} R1(t) &= R1(1997.0) - \Omega_x(t-1997.0) \\ R2(t) &= R2(1997.0) - \Omega_y(t-1997.0) \\ R3(t) &= R3(1997.0) - \Omega_z(t-1997.0) \end{aligned} \tag{3.35}$$

which agrees with Table 3.3, since the transformation, equation (3.29), uses angles defined in the reverse sense (NGS convention). Hence, e.g.,

$$\varepsilon_x(t) = -R1(t) = -R1(1997.0) + \Omega_x(t-1997.0) . \tag{3.36}$$

Using the transformation, equation (3.34), NGS thus realized NAD83 at all CORS stations and designated this realization NAD83(CORS96). By definition all temporal variations in the displacement and scale parameters in this transformation were set to zero.

For transformations to NAD83 from the next realization of ITRS, the NGS adopted slightly different transformation parameters than determined by the IERS. The transformation parameters from ITRF96 to ITRF97 are published as zero (including zero time-derivatives of these parameters); see Table 3.2. Yet, the International GNSS Service (IGS) determined the transformation ITRF96 to ITRF97 based solely on GPS stations and found non-zero transformation parameters. Since the control networks of NAD83 are now largely based on

GPS, NGS decided to use the IGS-derived ITRF96-to-ITRF97 transformation, yielding the transformation parameters between ITRF97 and NAD83 as given in Table 3.3 and obtained from:

$$\begin{aligned} ITRF97 \to NAD83(CORS96) = {} & \left(ITRF97 \to ITRF96\right)_{IGS} \\ & + \left(ITRF96 \to NAD83(CORS96)\right) \end{aligned} \tag{3.37}$$

For ITRF2000, there were only insignificant differences between the transformation parameters determined by IERS and by IGS, and thus we have

$$\begin{aligned} ITRF2000 \to NAD83(CORS96) = {} & \left(ITRF2000 \to ITRF97\right)_{IERS} \\ & + \left(ITRF97 \to ITRF96\right)_{IGS} \\ & + \left(ITRF96 \to NAD83(CORS96)\right) \end{aligned} \tag{3.38}$$

as verified by the numerical values in Tables 3.2 and 3.3.

Since the IERS-derived transformation parameters for ITRFs are time-dependent, the more general transformation to NAD83 now yields time-dependent coordinates in NAD83. However, for the most part these reflect only very small motions within the NAD83 frame. In order to determine velocities of control points within NAD83 based on their velocities in the ITRF, one can use equation (3.26). It is expected that most of the ITRF velocity associated with a point (the first term in equation (3.26)) is cancelled by the plate motion, given by the last term, so that within NAD83 there is essentially no motion, only residual motion due to local effects. For example, those points near a plate boundary (such as near the west coast of the U.S.) have significant motion within NAD83 that is determined by the total motion of ITRF minus the overall plate motion model.

Recently, NGS updated all NAD83 coordinates of its CORS stations to the epoch 2010.0, and used formula (3.26) to determine the corresponding NAD83 velocities. A utility called OPUS[11] (On-line User Positioning Service) (Soler and Snay 2004) is offered by NGS to determine 2010.0 coordinates, $\boldsymbol{x}$, in NAD83 for any point observed by static differential GPS observations. For simple examples of how the NAD83 and ITRF coordinates of CORS points are related, see Problem 3.4.2-3.

[11] http://www.ngs.noaa.gov/OPUS/

3.4.2 Problems

1. a) Rigorously derive the approximation, formula (3.20), from the exact equation (3.18) (3.15) and clearly state all approximations. Determine the error in coordinates of the point in Problem 3.1.2-2 when using equation (3.20) instead of equation (3.18) for the parameters associated with the ITRF2000 – NAD83(CORS86) transformation.

b) Given the coordinates of a point in Columbus: $\phi = 40°$, $\lambda = -83°$, $h = 200$ m, in the NAD83(CORS86) frame, compute its coordinates in the ITRF89, as well as in the ITRF94, based on the transformation parameters in Tables 3.2 and 3.3.

2. a) Which of the following remain invariant in a 7-parameter similarity transformation, equation (3.18)?

i) chord distance; ii) distance from origin; iii) longitude

b) Answer 2.a) for each of the quantities listed in case $R = I$ (identity matrix) (be careful!).

3. Using the web site: http://www.ngs.noaa.gov/CORS/, find the coordinate data sheet of CORS station Westford (WES2) in eastern Massachusetts (click on the station and then use the links "Get Site Info" and "Coordinates"; then click on "Position and Velocity" and use the Positions at ARP (antenna reference point)).

a) Using the listed IGS08(2005) coordinates and velocities for this GPS station, compute the NAD83 coordinates and their velocity for 2010.0 and compare them to the values published by NGS. Use the transformation IGS08-to-NAD83(2011) in Table 3.3.

b) Do the same for the CORS station, Point Loma 5 (PLO5), near San Diego, Southern California.

3.5 Vertical Datums

Nowadays, heights of points could be reckoned using GPS with respect to an ellipsoid; in fact, we have already introduced this height as the ellipsoidal height, h (Section 2.1.2). However, this height does not correspond with our intuitive sense of height as a measure of vertical distance with respect to a *level surface*. Two points with the same ellipsoidal height may be at different levels in the sense that water would flow from one point to the other. Ellipsoidal heights are purely geometric quantities that have no connection to the gravity potential; and, it is the *gravity potential* that determines which way water flows. An unperturbed lake surface comes closest to a physical manifestation of a level surface. Mean sea level (often quoted as a reference for heights) is also reasonably close, but not equal to a level surface, due to various non-gravitational forces that cause the hydrostatic equilibrium of the mean surface to deviate from being gravitationally level. We may *define* a level surface simply as a surface on which the gravity potential is constant. Discounting friction, no work is done in moving an object along a level surface; water does not flow on a level surface; and all points on a level surface should be at the same height – at least, this is what we intuitively would like to understand by heights. The *geoid* is defined to be that level surface that closely approximates mean sea level (mean sea level deviates from the geoid by up to 2 m due to the persistent variations in pressure, salinity, temperature, wind setup, etc., of the oceans). There is still today considerable controversy about the exact *realizability* (accessibility) of the geoid as a definite surface, and the definition given here is correspondingly (and intentionally) vague.

A vertical datum, like a horizontal datum, requires an origin, but being one-dimensional, there is no orientation; and, the scale is inherent in the measuring apparatus (leveling rods). The origin is a point on the Earth's surface where the height is a defined value (e.g., zero height at a coastal tide-gauge station); but, an alternative definition is now being considered by some countries (see below). This origin is obviously accessible and satisfies the requirement for the definition of a datum. From this origin point, heights (that is, height differences) can be measured to any other point using standard leveling procedures (which we do not discuss further). Traditionally, a point at mean sea level served as origin point, but it is not important what the absolute gravity potential is at this point, since one is interested only in height differences (potential differences) with respect to the origin. This is completely analogous to the traditional horizontal datum, where the origin point (e.g., located on the surface of the Earth) may have arbitrary coordinates, and all other points within the datum are tied to the origin in a relative way. Each vertical datum, being thus defined with respect to an arbitrary origin, is not tied to a global, internationally agreed upon, vertical datum. The latter, in fact, does not yet exist officially, although much debate, discussion, literature, and candidate models have centered on just such a datum.

Figure 3.5 shows the geometry of two local vertical datums each of whose origin is a station at mean sea level. In order to transform from one vertical datum to another requires knowing the gravity potential difference between these origin points. This difference is not zero because mean sea level is not exactly a level surface; differences in height between the origins typically are several decimeters.

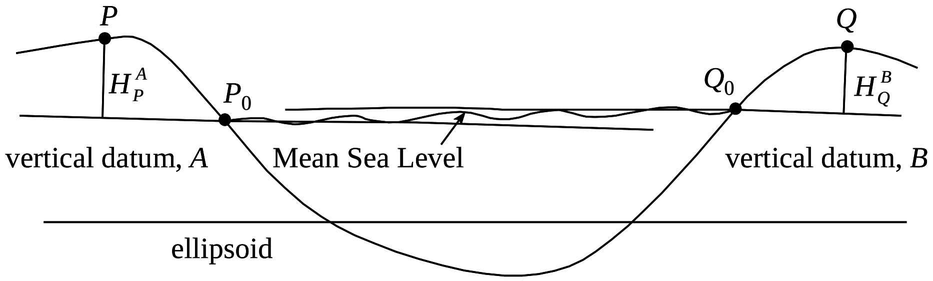


Figure 3.5: Two vertical datums with respect to mean sea level.

The heights that are measured and belong to a particular vertical datum ultimately are defined by differences in gravity potential with respect to the origin point. There are a number of options to scale the geopotential difference so that it represents a height difference (that is, with distance units). The most natural height (but not necessarily the most realizable height from a theoretical viewpoint) is the *orthometric height*, H, defined as the distance along the (curved) plumb line from the level surface (a local geoid), that passes through the datum origin, to the point in question. With sufficient accuracy, we may neglect the curvature of the plumb line and approximate the orthometric height as a distance along the ellipsoidal normal. Analogous to Figure 3.2, we then have

$$H = h - N \,, \tag{3.39}$$

where N is the distance from the ellipsoid to the level surface that passes through the origin point. This is the (local) *geoid undulation*. It is equal to the global geoid undulation minus the offset of the origin point or local vertical datum from the global geoid.

For North America, the *National Geodetic Vertical Datum of 1929* (NGVD29) served the U.S. for vertical control until the late 1980's; and Canada's Geodetic Vertical Datum of 1928 (CGVD28) until 2013 was still the official datum for vertical control. The origin of NGVD29 was actually based on several *defined* heights of *zero* at 21 coastal (mean sea level) tide-gauge stations in the U.S. and 5 in Canada. Similarly, a suitable set of coastal tide gauge stations served to define the origin of CGVD28. Defining zero height at different points of mean sea level caused distortions in the network since, as noted above, mean sea level is not a level surface. Additional distortions were introduced because leveled heights were not corrected rigorously for the non-parallelism of the level surfaces.

In 1988 a new vertical datum was introduced, the *North American Vertical Datum of 1988* (NAVD88). Its origin is a single station with a defined height (not zero) at Pointe-au-Père

(Father's Point), on the St. Lawrence river in Rimouski, Québec, which is also the origin point for the *International Great Lakes Datum* of 1985 (IGLD85). Despite this location for the origin, NAVD88 was never officially adopted by Canada. Defining the origin at a single point eliminated the theoretical problem of constraining a level surface to a non-level surface (mean sea level). Also, the leveled heights were more rigorously corrected for the non-parallelism of the level surfaces.

However, recent analyses determined[12], with improved gravity models and GPS (providing N and h, respectively, in equation (3.39)), that the entire network has a tilt error of more than 1 meter from the east coast (where the origin lies) to the west coast. This is due in part to the propagation of systematic leveling errors, but also to remaining model errors in the implementation of the theory of orthometric height determination.

To rectify these problems (among others), NGS plans to replace NAVD88 by a geopotential model. This new paradigm in vertical control constitutes a re-definition of a system analogous to the re-definitions of past horizontal datums. In essence, there will be no physical benchmark to define the origin of the datum. Instead, a chosen value, W_0, of the gravity potential will serve the function of defining the geoid. With an accurate geopotential model, it is then just a matter of determining the ellipsoidal height of a point (using GPS) and determining the geoid undulation, N, for this point from the gravity model. Making use of equation (3.39) then yields the orthometric height, H. Clearly, the geopotential model must be very accurate so that the computational error in N is commensurate with that of h. The goal is cm-level accuracy for H over the entire continent. New Zealand has already instituted such a system in 2009 based on a geoid model (Amos 2010); and The Geodetic Survey Division (GSD) of Natural Resources Canada likewise replaced CGVD28 with CGVD2013[13], which is a gravity model with one of its equipotential surfaces, $W_0 = 62,636,856.0\ \text{m}^2/\text{s}^2$, representing the reference for heights. It is close to mean level around the coasts of Canada, but not specifically tied to a mark on the ground. NGS is developing an accurate gravity model from a systematic survey of gravity by airborne systems throughout the country to bring all existing and new gravity data to a common level of accuracy. Eventually, the combined NGS and GSD vertical datums will become a unified North American Vertical Datum.

[12] Proceedings of the 2010 Federal Geospatial Summit on Improving the National Spatial Reference System. http://www.ngs.noaa.gov/2010Summit/2010FederalGeospatialSummitProceedings.pdf

[13] http://www.nrcan.gc.ca/earth-sciences/geomatics/geodetic-reference-systems/9054#_Toc372901509

Chapter 4

Celestial Reference System

Ultimately the orientation of the terrestrial reference system is tied to an astronomic system, as it has always been throughout history. The astronomic reference system, or more correctly, the *celestial reference system* is supposed to be an *inertial* reference system in which our laws of physics hold without requiring corrections for rotations. For geodetic purposes it serves as the primal reference for positioning since it has no dynamics. Conversely, it is the system with respect to which we study the dynamics of the Earth as a rotating body. And, finally, it serves, of course, also as a reference system for astrometry.

We will study primarily the transformation from the celestial reference frame to the terrestrial reference frame and this requires some understanding of the dynamics of Earth rotation and its orbital motion, as well as the effects of observing celestial objects on a moving and rotating body such as the Earth. The definition of the celestial reference system was until rather recently (1998), in fact, tied to the dynamics of the Earth, whereas, today it is defined as being almost completely independent of the Earth. The change in definition is as fundamental as that which transferred the origin of the regional terrestrial reference system (i.e., the horizontal geodetic datum) from a monument on Earth's surface to the geocenter. It is, as always, a question of accessibility or realizability. Traditionally, the orientation of the astronomic or celestial reference system was defined by two naturally occurring direction in space, the north celestial pole, basically defined by Earth's spin axis (or close to it), and the intersection of the celestial equator with the ecliptic, i.e., the vernal equinox (see Section 2.3.2). Once the dynamics of these directions were understood, it was possible to define *mean* directions that are fixed in space and the requirement of an inertial reference system was fulfilled (to the extent that we understand the dynamics). The stars provided the accessibility to the system in the form of coordinates (and their variation) as given in a fundamental catalog, which is then the celestial reference frame. Because the defining directions (the orientation) depend on the dynamics of the

Earth (within the dynamics of the mutually attracting bodies in our solar system), even the mean directions vary slowly in time. Therefore, the realization of the system included an epoch of reference; i.e., a specific time when the realization held true. For any other time, realization of the frame required transformations based on the motion of the observable axes, which in turn required a dynamical theory based on a fundamental set of constants and parameters. All this was part of the definition of the celestial reference system.

On the other hand, it is known that certain celestial objects, called *quasars* (quasi-stellar radio sources), exhibit no perceived motion on the celestial sphere due to their great distance from the Earth. These are also naturally occurring directions, but they have no dynamics, and as such would clearly be much preferred for defining the orientation of the celestial system. The problem was their accessibility and hence the realizability of the frame. However, a solid history of accurate, very-long-baseline interferometry (VLBI) measurements of these quasars has prompted the re-definition of the celestial reference system as one whose orientation is defined by a set of quasars. In this way, the definition has fundamentally changed the celestial reference system from a *dynamic* system to a *kinematic* (or, geometrical) system. The axes of the celestial reference system are still (close to) the north celestial pole and vernal equinox, but are not defined dynamically in connection with Earth's motion, rather they are tied to the defining set of quasars whose coordinates are given with respect to these axes. Moreover, there is no need to define an epoch of reference, because (presumably) these directions will never change in inertial space (at least in the foreseeable future of mankind).

The IERS *International Celestial Reference System* (ICRS), thus, is defined to be an inertial system (i.e., non-rotating) whose first and third mutually orthogonal coordinate axes (equinox and pole) were realized initially (1995) by the coordinates of 608 compact extra-galactic sources (quasars), as chosen by the Working Group on Reference Frames of the International Astronomical Union (IAU); see Feissel and Mignard (1998). Of these, 212 sources defined the orientation, and the remainder comprised candidates for additional ties to the reference frame. The origin of the ICRS is defined to be the center of mass of the solar system (*barycentric* system) and is realized by observations in the framework of the theory of general relativity.

By recommendations from the International Astronomical Union (and duly adopted) the pole and equinox of the ICRS are supposed to be close to the mean dynamical pole and equinox of J2000.0 (Julian date, 2000, see below). Furthermore, the adopted pole and equinox for ICRS, for the sake of continuity, should be consistent with the directions realized for FK5, which is the fundamental catalogue (fifth version) of stellar coordinates that refers to the epoch J2000.0 and served as realization of a previously defined celestial reference system. Specifically, the origin of right ascension for FK5 was originally defined on the basis of the mean right ascension of 23 radio sources from various catalogues, with the right ascension of one particular source fixed to its FK4 value, transformed to J2000.0. Similarly, the FK5 pole was based on its J2000.0 direction defined using the 1976 precession and 1980 nutation series (see below). The FK5 directions are estimated to be accurate to ± 50 milliarcsec for the pole and ± 80 milliarcsec for the equinox; and, it is now known, from improved observations and dynamical models

(McCarthy 1996, McCarthy and Petit 2003, Petit and Luzum 2010), that the ICRS pole and equinox are close to the mean dynamical equinox and pole of J2000.0, well within these tolerances. Thus, the definition of the ICRS origin of right ascension and pole are only qualitative with respect to FK5 – fundamentally they are defined to be geometric axes fixed by a set of quasars. The precise transformation to a dynamical system, such as defined by modern theories, is briefly discussed in Section 4.1.3.

The realization of the ICRS, the *International Celestial Reference Frame* (ICRF) is accomplished with VLBI measurements of the quasars; and, as observations improve the orientation of the ICRF will be adjusted so that it has no net rotation with respect to previous realizations (analogous to the ITRF). The original realization was designated ICRF1; and, it was extended in 1999 and again in 2002 with additional objects observed with VLBI, thus totaling 667 and 717, respectively. The next significant realization, designated ICRF2, was constructed in 2009, where now 295 quasars define the system (being more stable and better distributed in the sky than for ICRF1), and which also includes an additional 3119 extragalactic sources. Aside from VLBI, the principal realization of the ICRS is through the Hipparcos catalogue, based on recent observations of some 120,000 well-defined stars using the Hipparcos (High Precision Parallax Collecting Satellite), optical, orbiting telescope. This catalogue is tied to the ICRF with an accuracy of about 0.6 mas (milliarcsec) in each axis. Additional catalogues for up to 100 million stars are described by Petit and Luzum (2010).

4.1 Dynamics of the Pole and Equinox

Despite the simple, geometric (kinematic) definition and realization of the ICRS, we do live and operate on a dynamical body, the Earth, whose naturally endowed directions (associated with its spin and orbital motion) in space vary due to the dynamics of motion according to gravitational and geodynamical theories. Inasmuch as we observe celestial objects to aid in our realization of terrestrial reference systems, we need to be able to transform between the ICRF and the ITRF, and therefore, we need to understand these dynamics to the extent, at least, that allows us to make these transformations accurately. The description of the transformation, comprising *Earth orientation parameters*, has also changed in recent years. Here, both the traditional description and the modern transformation are treated, where the traditional one is perhaps a bit more accessible in terms of physical intuition, whereas, the latter tends to hide these concepts. Furthermore, the opportunity was taken in the new approach to implement certain nuances necessary for an unambiguous definition of Earth rotation. Thus, we start with the traditional approach and evolve this into the modern transformation formulas.

Toward this end, we need, first of all, to define a system of time (since the theoretical description of *dynamics* inherently requires it). We call the relevant time scale the *Dynamic Time*, referring to the time variable in the equations of motion describing the dynamical behavior of the massive bodies of our solar system. Rigorously (with respect to the theory of general relativity), the dynamic time scale can refer to a coordinate system (coordinate time) that is, for example, *barycentric* (origin at the center of mass of the solar system) or *geocentric*, and is thus designated barycentric coordinate time (TCB) or geocentric coordinate time (TCG); or, it refers to a *proper time*, associated with the frame of the observer (terrestrial dynamic time (TDT), or barycentric dynamic time (TBD)); see Section 5.3 on further discussions of the different dynamical time scales. The dynamic time scale, based on proper time, is the most uniform that can be defined theoretically, meaning that the time scale in our local experience, as contained in our best theories that describe the universe, is constant.

Dynamic time is measured in units of *Julian days*, which are close to our usual days based on Earth rotation, but they are *uniform*; whereas, solar days (based on Earth rotation) are not, for the simple reason that Earth rotation is not uniform. The origin of dynamic time, designated by the *Julian date*, J0.0, is defined to be Greenwich noon, 1 January 4713 B.C. Julian days, by definition, start and end when it is noon (dynamical time) in Greenwich, England. Furthermore, by definition, there are exactly 365.25 Julian days in a *Julian year*, or exactly 36525 Julian days in a *Julian century*. With the origin as given above, the Julian date, J1900.0, corresponds to the Julian day number, JD2,415,021.0, being Greenwich noon, 1 January 1900; and the Julian date, J2000.0, corresponds to the Julian day number, JD2,451,545.0, being Greenwich noon, 1 January 2000 (see Figure 4.1). We note that Greenwich noon represents mid-day in our usual designation of days starting and ending at midnight, and so JD2,451,545.0 is also 1.5 January 2000. Continuing with this scheme, 0.5 January 2000 is really Greenwich noon, 31 December 1999 (or 31.5 December 1999).

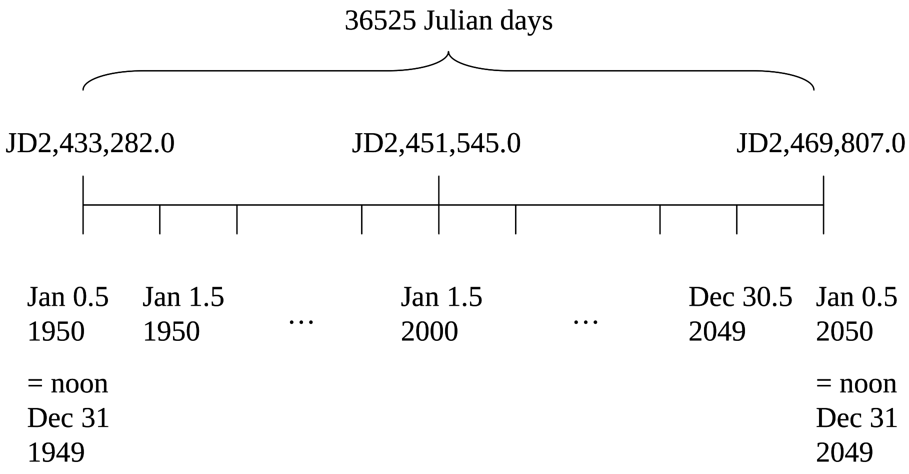


Figure 4.1: One Julian century.

For practical reasons, a *modified* Julian day number,

$$MJD = JD - 2400000.5, \tag{4.1}$$

is also defined relative to a new origin, which counts days as starting at midnight in Greenwich.

An *epoch* is an instant in time (as opposed to a *time interval* which is the difference between two epochs). We may define three epochs, as follows:

t_0: the fundamental or basic epoch for which the values of certain constants and parameters are defined that are associated with the dynamical theories of the transformation (previously, the reference system).

t: the *epoch of date*, being the current or some other time at which the dynamics should be realized (e.g., the time of observation).

t_F: an epoch that is fixed and arbitrary, representing another epoch with respect to which the theory could be developed.

The distinction between t_0 and t_F is a matter of convenience, where t_0 always refers to the epoch for which the constants are defined.

4.1.1 Precession

The gravitational interaction of the Earth with the other bodies of the solar system, including primarily the moon and the sun, but also the planets, causes Earth's orbital motion to deviate from the simple Keplerian model of motion of two point masses in space. Also, because the Earth is not a perfect homogeneous sphere, its rotation is affected likewise by the gravitational action of the bodies in the solar system. If there were no other planets (only the Earth/moon system) then the orbit of the Earth/moon system around the sun would be essentially a plane fixed in space. This plane defines the ecliptic (see also Section 2.3.2). But the gravitational actions of the planets cause this ecliptic plane to behave in a dynamic way, called *planetary precession*.

If the obliquity of the ecliptic (the tilt angle between equator and ecliptic, Section 2.3.2) were zero or the Earth were not flattened at its poles, then there would be no gravitational torques due to the sun, moon, and planets acting on the Earth. But since $\varepsilon \neq 0$ and $f \neq 0$, the sun, moon, and planets do cause a precession of the equator (and, hence, the pole) that is known as *luni-solar precession* and *nutation*, depending on the period of the motion. That is, the equatorial bulge of the Earth and its tilt with respect to the ecliptic allow the Earth to be torqued by the gravitational forces of the sun, moon, and planets, since they all lie approximately on the ecliptic. Planetary precession together with luni-solar precession is known as *general precession*.

The complex dynamics of the precession and nutation is a superposition of many periodic motions originating from the myriad of periods associated with the orbital dynamics of the corresponding bodies. Smooth, long-period motion is termed luni-solar precession, and short-periodic (up to 18.6 years) is termed nutation. The periods of nutation depend primarily on the orbital motion of the moon relative to the orbital period of the Earth. The most recent models for nutation also contain short-periodic effects due to the relative motions of the planets.

We distinguish between precession and nutation even though to some extent they have the same sources. In fact, the modern approach mentioned earlier combines the two into one model (as seen later in Section 4.1.3). Since precession is associated with very long-term motions of the Earth's reference axes in space, we divide the total motion into *mean* motion, or average motion, that is due to precession and the effect of short-period motion, due to nutation, that at a particular epoch describes the residual motion, so to speak, with respect to the mean. First, we discuss precession over an interval of time. The theory for determining the motions of the reference directions was developed by Simon Newcomb at the turn of the 20th century. Its basis lies in celestial mechanics and involves the *n*-body problem for planetary motion, for which no analytical solution has been found (or exists). Instead, iterative, numerical procedures have been developed and formulated. We cannot give the details of this (see, e.g., Woolard 1953), but can only sketch some of the results.

In the first place, planetary precession may be described by two angles, π_A and Π_A, where the subscript, *A*, refers to the "accumulated" angle from some fixed epoch, say t_0, to some other epoch, say *t*. Figure 4.2 shows the geometry of the motion of the ecliptic due to planetary precession from t_0 to *t*, as described by the angles, π_A and Π_A. The pictured ecliptics and equator are fictitious in the sense that they are affected only by precession and not nutation, and as such are called "mean ecliptic" and "mean equator". The angle, π_A, is the angle between the mean ecliptics at t_0 and *t*; while Π_A is the ecliptic longitude of the axis of rotation of the ecliptic due to planetary precession. The vernal equinox at t_0 is denoted by Υ_0.

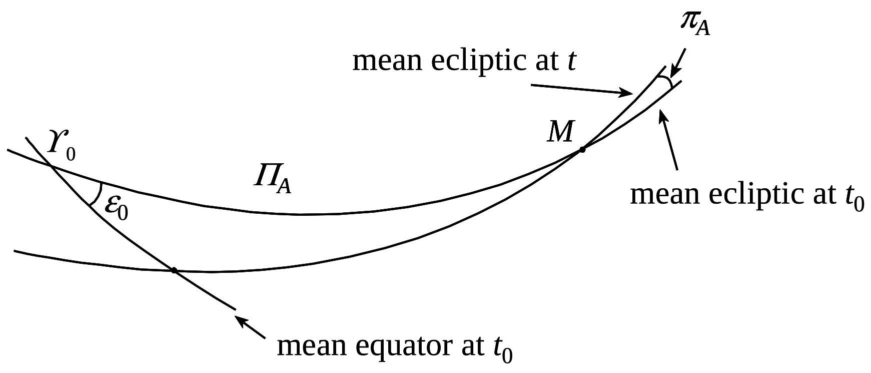


Figure 4.2: Planetary precession.

The angles, π_A and Π_A, can be expressed as time series where the coefficients are based on the celestial dynamics of the planets. Usually, the series are given in the form:

$$\begin{aligned}\sin\pi_A \sin\Pi_A &= s\left(t-t_0\right)+s_1\left(t-t_0\right)^2+s_2\left(t-t_0\right)^3+\cdots\\ \sin\pi_A \cos\Pi_A &= c\left(t-t_0\right)+c_1\left(t-t_0\right)^2+c_2\left(t-t_0\right)^3+\cdots\end{aligned} \tag{4.2}$$

The epoch about which the series is expanded could also be t_F, but then the coefficients would obviously have different values. For example, Seidelmann (1992, p.104) gives the following series based on 1976 theory and associated constants; see also Woolard, 1953, p.44):

$$\begin{aligned}\pi_A \sin\Pi_A &= \left(4.1976-0.75250T+0.000431T^2\right)\tau\\ &\qquad +\left(0.19447+0.000697T\right)\tau^2-0.000179\tau^3 \text{ [arcsec]}\\ \pi_A \cos\Pi_A &= \left(-46.8150-0.00117T+0.005439T^2\right)\tau\\ &\qquad +\left(0.05059-0.003712T\right)\tau^2+0.000344\tau^3 \text{ [arcsec]}\end{aligned} \tag{4.3}$$

where the time variables, T and τ, are fractions of a Julian century, given by

$$T=\frac{t_F-t_0}{36525}, \qquad \tau=\frac{t-t_F}{36525}, \tag{4.4}$$

and the epochs, t_0, t_F, and t, are Julian dates given in units of Julian days (specifically, $t_0 = 2,451,545.0$). The coefficients in the series have appropriate units so that each term is in units of arcsecond. It is noted that this two-epoch approach to formulating precession has been largely abandoned in modern theories with no difference in accuracy.

The luni-solar precession depends on the geophysical parameters of the Earth. No analytic formula based on theory was used for this due to the complicated nature of the Earth's shape and internal constitution. Instead, Newcomb gave an empirical parameter, (now) called *Newcomb's precessional constant*, P_N, based on observed rates of precession. In fact, this "constant" rate is not strictly constant, as it depends slightly on time according to

$$P_N = P_0 + P_1\left(t-t_0\right), \tag{4.5}$$

where $P_1 = -0.00369 \text{ arcsec/century}$ (per century) is due to changes in eccentricity of Earth's orbit (Lieske et al. 1977, p.10). Newcomb's precessional constant depends on Earth's moments of inertia and enters in the dynamical equations of motion for the equator due to the gravitational torques of the sun and moon. It is not accurately determined on the basis of geophysical theory, rather it is derived from observed general precession rates. It describes the motion of the mean equator along the ecliptic according to the *rate*:

$$\dot{\psi} = P_N \cos\varepsilon_0 - P_g \,, \qquad (4.6)$$

where ε_0 is the obliquity of the ecliptic at t_0, and P_g is a general relativistic term called the *geodesic precession*. The accumulated angle in luni-solar precession of the equator along the ecliptic is given by ψ_A.

Figure 4.3 shows the accumulated angles of planetary and luni-solar precession, as well as general precession (in longitude). The precession angles, as given in this figure, describe the motion of the mean vernal equinox as either along the mean ecliptic (the angle, ψ_A, due to motion of the mean equator, that is, luni-solar precession), or along the mean equator (the angle, χ_A, due to motion of the mean ecliptic, that is, planetary precession). The accumulated general precession in longitude is the angle, as indicated, between the mean vernal equinox at epoch, t_0, and the mean vernal equinox at epoch, *t*. Even though (for relatively short intervals of several years) these accumulated angles are small, we see that the accumulated general precession is not simply an angle in longitude, but motion due to a compounded set of rotations.

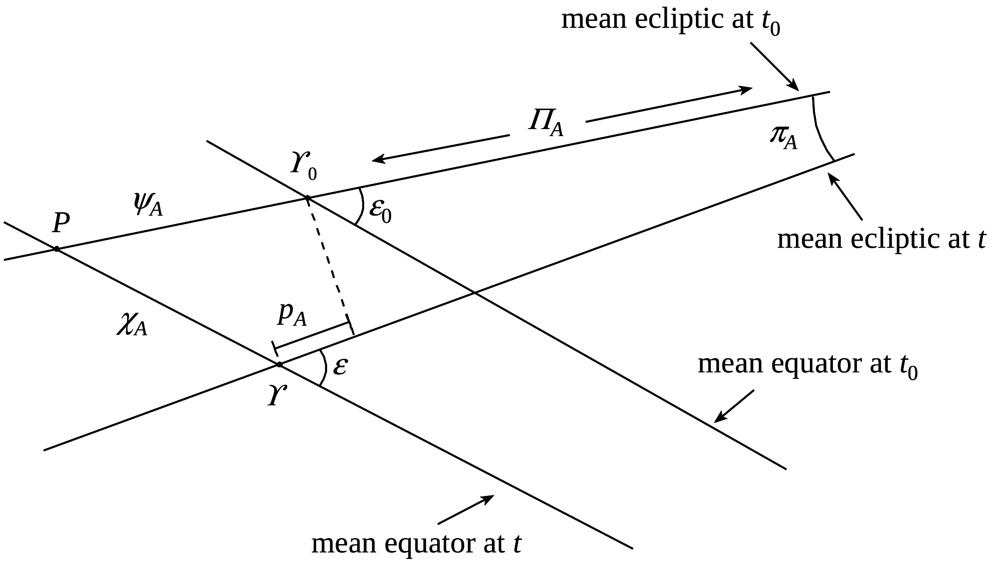


Figure 4.3: General precession = planetary precession + luni-solar precession.

It is easier to formulate the relationships between the various types of precession by considering the limits of the accumulated angles as the time interval goes to zero, that is, by considering the *rates*. Following conventional notation, we denote rates by the corresponding un-subscripted angles:

$$\chi = \left.\frac{d\chi_A}{dt}\right|_{t=t_0} , \qquad \psi = \left.\frac{d\psi_A}{dt}\right|_{t=t_0} , \qquad p = \left.\frac{dp_A}{dt}\right|_{t=t_0} . \tag{4.7}$$

From Figure 4.3, we thus have the following relationship between the precession rates (viewing the geometry of the accumulated motions in the differential sense):

$$p = \psi - \chi \cos\varepsilon_0 , \tag{4.8}$$

where the second term is merely the projection of the planetary precession onto the ecliptic. Now, applying the law of sines to the spherical triangle $MP\varUpsilon$ in Figure 4.3, we find

$$\begin{aligned} &\sin\chi_A \sin\left(180° - \varepsilon\right) \approx \sin\pi_A \sin\varPi_A \\ \Rightarrow \qquad &\chi_A \sin\varepsilon \approx \sin\pi_A \sin\varPi_A \end{aligned} \tag{4.9}$$

Substituting the first of equation (4.2) and taking the time derivative according to equation (4.7), we have for the rate in planetary precession

$$\chi = \frac{s}{\sin\varepsilon_0} , \tag{4.10}$$

where second-order terms (e.g., due to variation in the obliquity) are neglected. Putting equations (4.10) and (4.6) into equation (4.8), the rate of general precession (in longitude) is given by

$$p = P_N \cos\varepsilon_0 - P_g - s\cot\varepsilon_0 . \tag{4.11}$$

More rigorous differential equations are given by Lieske et al. (1977, p.10).

Equation (4.11) shows that Newcomb’s precessional constant, P_N, is related to the general precession rate; and, this is how it is determined, from the observed rate of general precession at epoch, t_0. This observed rate was one of the adopted constants that constituted the definition of the celestial reference system when it was defined dynamically. The other constants included P_1 (the time dependence of Newcomb’s constant), P_g (the geodesic precession term), ε_0 (the obliquity at epoch, t_0), and any other constants needed to compute the coefficients, s, s_k, c, c_k, on the basis of planetary dynamics. Once these constants are adopted, all other precessional parameters can be derived.

The rate of general precession of the vernal equinox in longitude can also be decomposed into rates (and accumulated angles) in right ascension, *m*, and declination, *n*. From Figure 4.4, we have the accumulated general precession of Υ_0 in declination, n_A, and in right ascension, m_A:

$$\begin{aligned} n_A &= \psi_A \sin\varepsilon_0, \\ m_A &= \psi_A \cos\varepsilon_0 - \chi_A; \end{aligned} \tag{4.12}$$

and, in terms of rates:

$$\begin{aligned} n &= \psi \sin\varepsilon_0, \\ m &= \psi \cos\varepsilon_0 - \chi. \end{aligned} \tag{4.13}$$

Finally, the rate of general precession in longitude is then also given by:

$$p = m\cos\varepsilon_0 + n\sin\varepsilon_0 . \tag{4.14}$$

Again, these formulas hold only for Υ_0; the precession for general points is given below.

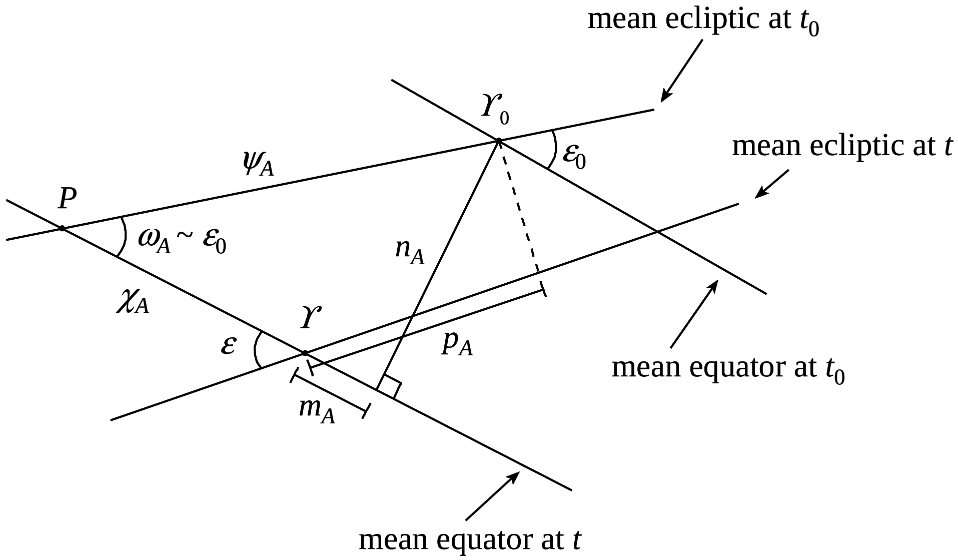


Figure 4.4: General precession in right ascension and declination.

In fact, the rate, n, (and accumulated angle, n_A) is one of three precessional elements that are used to transform coordinates of celestial points from a frame referring to the mean pole and equinox at t_0 (or some other fundamental epoch) to a frame associated with the mean pole and equinox at another epoch, t. The accumulated general precession in declination is also designated, θ_A. For the accumulated angle in right ascension one cannot use m_A, as defined above, since the polar convergence of the hour circles must be considered. Instead, two other precessional elements are used that enable the transformation. Referring to Figure 4.5, also showing the result of general precession, but now in terms of the motions of the pole and equinox, we define two angles, z_A and ζ_A, in right ascension. The mean pole, Z_0, at epoch, t_0, moves as a result of general precession to its position, Z, at epoch, t; and the connecting great circle arc clearly is the accumulated general precession in declination. The general precession rate of Υ_0 in right ascension can be decomposed formally into rates along the mean equator at epoch, t_0, and along the mean equator at a differential increment of time later:

$$m = \zeta + z \,. \tag{4.15}$$

We see that the great circle arc, $\overset{\frown}{Z_0 Z Q}$, intersects the mean equator of t_0 at right angles because it is an hour circle with respect to the pole, Z_0; and it intersects the mean equator of t at right angles because it is also an hour circle with respect to the pole, Z. Consider a point on the celestial sphere. Let its coordinates in the mean celestial reference frame of t_0 be denoted by (α_0, δ_0) and in the mean frame at epoch, t, by (α_m, δ_m). In terms of unit vectors, let

$$\boldsymbol{r}_0 = \begin{pmatrix} \cos\alpha_0 \cos\delta_0 \\ \sin\alpha_0 \cos\delta_0 \\ \sin\delta_0 \end{pmatrix}, \qquad \boldsymbol{r}_m = \begin{pmatrix} \cos\alpha_m \cos\delta_m \\ \sin\alpha_m \cos\delta_m \\ \sin\delta_m \end{pmatrix}. \tag{4.16}$$

Then, with the angles as indicated in Figure 4.5, we have the following transformation between the two frames:

$$\begin{aligned} \boldsymbol{r}_m &= \mathbf{R}_3(-z_A)\,\mathbf{R}_2(+\theta_A)\,\mathbf{R}_3(-\zeta_A)\,\boldsymbol{r}_0 \\ &= \mathbf{P}\,\boldsymbol{r}_0 \end{aligned} \tag{4.17}$$

where $\mathbf{P}$ is called the *precession transformation matrix*. Again, note that this is a transformation between *mean* frames, where the nutations have not yet been taken into account.

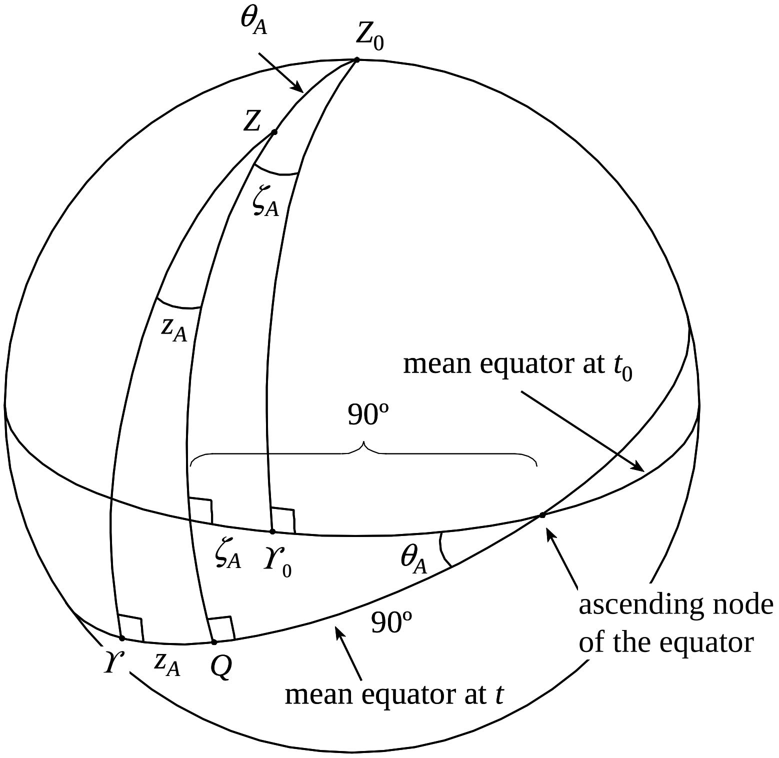


Figure 4.5: Precessional elements.

Numerical values for the precessional constants based on adopted and derived constants by the International Astronomical Union in 1976, are given by Lieske et al. (1977),

$$\begin{aligned} p &= 5029.0966 \text{ arcsec/Julian century} \\ P_1 &= -0.00369 \text{ arcsec/Julian century} \\ P_g &= 1.92 \text{ arcsec/Julian century} \\ \varepsilon_0 &= 23°26'21.448'' \end{aligned} \qquad (4.18)$$

and refer to the fundamental epoch, $t_0 = J2000.0$. Based on these, series expressions could be developed for the various precessional quantities and elements, as shown, for example, by Seidelmann (1992, p.104). These series (and adopted constants) constituted the IAU precession theory of 1976. With space improved measurements (VLBI and lunar laser ranging) it was found that the adopted constants deviated significantly with respect to the precision of the measurements (Capitaine et al. 2003). In 2000, the IAU recommended a revision of the precession model, combined with a substantial revision of the nutation model (see below) derived from a least-squares adjustment to current VLBI data, based on the work of Mathews et al. (2002). Principally, this new model corrects the longitude and obliquity precession rates. An improvement in the dynamical theory for precession was developed (Capitaine et al. 2005) and adopted by the IAU in 2006 that expresses the polynomial series for the precessional elements with terms up to fifth order. These are, with units of milli-arcseconds [mas],

$$\pi_A = 46998.973\tau - 33.4926\tau^2 - 0.12559\tau^3 + 0.000113\tau^4 - 0.0000022\tau^5 \tag{4.19}$$

$$\Pi_A = 629546793.6 - 867957.58\tau + 157.992\tau^2 - 0.5371\tau^3 - 0.04797\tau^4 + 0.000072\tau^5 \tag{4.20}$$

$$\psi_A = 5038481.507\tau - 1079.0069\tau^2 - 1.14045\tau^3 + 0.132851\tau^4 - 0.0000951\tau^5 \tag{4.21}$$

$$\chi_A = 10556.403\tau - 2381.4292\tau^2 - 1.21197\tau^3 + 0.170663\tau^4 - 0.0000560\tau^5 \tag{4.22}$$

$$p_A = 5028796.195\tau + 1105.4348\tau^2 + 0.07964\tau^3 - 0.023857\tau^4 - 0.0000383\tau^5 \tag{4.23}$$

$$\zeta_A = 2650.545 + 2306083.227\tau + 298.8499\tau^2 + 18.01828\tau^3 - 0.005971\tau^4 - 0.0003173\tau^5 \tag{4.24}$$

$$z_A = -2650.545 + 2306077.181\tau + 1092.7348\tau^2 + 18.26837\tau^3 - 0.028596\tau^4 - 0.0002904\tau^5 \tag{4.25}$$

$$\theta_A = 2004191.903\tau - 429.4934\tau^2 - 41.82264\tau^3 - 0.007089\tau^4 - 0.0001274\tau^5 \tag{4.26}$$

$$\varepsilon_A = 84381406.0 - 46836.769\tau - 0.1831\tau^2 + 200340\tau^3 - 0.000576\tau^4 - 0.0000434\tau^5 \tag{4.27}$$

where τ is given by equations (4.4) with $t_F = t_0$, hence,

$$\tau = \frac{t - t_0}{36525}; \tag{4.28}$$

and t_0 corresponds to J2000.0, i.e., $t_0 = 2,451,545.0$.

The coefficient of τ in these series represents the *rate* of the corresponding precessional element at $t = t_0$ (i.e., $\tau = 0$). For example,

$$\begin{aligned} \left.\frac{d}{d\tau}\psi_A\right|_{\tau=0} &= 5038.481507 \text{ arcsec/Julian century} \\ &\approx 50 \text{ arcsec/year} \end{aligned} \tag{4.29}$$

which is the rate of luni-solar precession, causing the Earth's spin axis to precess with respect to the celestial sphere and around the ecliptic pole with a period of about 25,800 years. The luni-solar effect is by far the most dominant source of precession. The rate of change in the obliquity of the ecliptic is given by

$$\left.\frac{d}{d\tau}\varepsilon_A\right|_{\tau=0} = -46.836769 \text{ arcsec/Julian century} \\ \approx -0.47 \text{ arcsec/year} \tag{4.30}$$

and the rate of the westerly motion of the equinox, due to planetary precession, is given by

$$\left.\frac{d}{d\tau}\chi_A\right|_{\tau=0} = 10.556403 \text{ arcsec/Julian century} \\ \approx 0.11 \text{ arcsec/year} \tag{4.31}$$

4.1.2 Nutation

Up to now we have considered only what the dynamics of the pole and equinox are in the mean over longer periods. The nutations describe the dynamics over the shorter periods, traditionally limited to the longest period of about 18.6 years associated with the lunar orbit. Also, for precession we determined the motion of the mean pole and mean equinox over an interval, from t_0 to t. The transformation due to precession was from one mean frame to another mean frame. But for nutation, we determine the difference between the mean position and the true position for a particular (usually the current) epoch, t (also known as the *epoch of date*). The transformation due to nutation is one from a mean frame to a true frame at the same epoch. Since true axes now come into the picture, rather than mean axes, it is important to define exactly the polar axis with respect to which the nutations are computed (as discussed later, one may consider the spin axis, the angular momentum axis, the “figure” axis, or an axis defined in terms of the character of its motions in the frequency domain). Without giving a specific definition at this point (see, however, Section 4.3.2), the currently defined axis is called the *Celestial Intermediate Pole* (CIP) that corresponds closely to the spin axis and represents the Earth’s axis for which nutations are computed (2003 IERS Conventions, see Section 4.1.3). A previous designation, the Celestial Ephemeris Pole (CEP), is also discussed in some detail in section 4.3.2.

Recall that nutations are due primarily to the luni-solar attractions and hence can be modeled on the basis of a geophysical model of the Earth and its motions in space relative to the sun and moon. The nutations that we thus define are also called *astronomic nutations*. The transformation for the effect of nutation is accomplished with two angles, $\Delta\varepsilon$ and $\Delta\psi$, that respectively describe (1) the change (from mean to true) in the tilt of the equator with respect to the mean ecliptic, and (2) the change (again, from mean to true) of the equinox along the mean ecliptic (see Figure 4.6). There is no need to transform from the mean ecliptic to the true ecliptic, since the interest is only in the dynamics of the true equator (and by implication the true pole). The true vernal equinox, Υ_T, is always defined to be on the mean ecliptic.

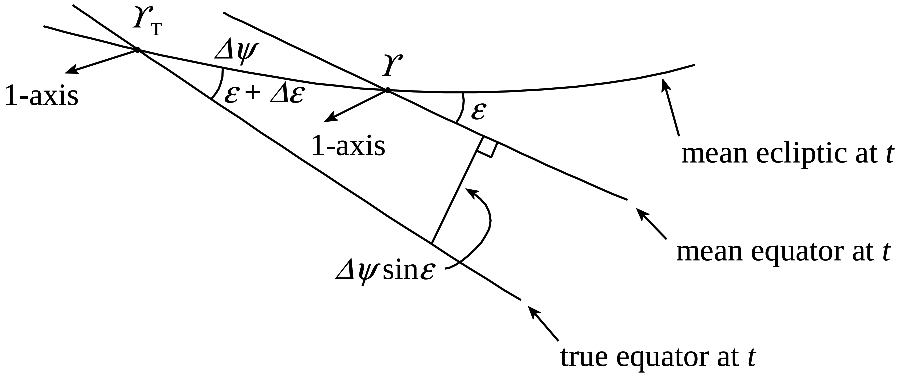


Figure 4.6: Nutational elements.

With respect to Figure 4.6, it is seen that $\Delta\psi$ is the *nutation in longitude*. It is due mainly to the orbital ellipticities of the Earth and the moon, causing non-uniformity in the luni-solar precessional effects. The *nutation in obliquity*, $\Delta\varepsilon$, is caused primarily by the moon's orbital plane being out of the ecliptic (by about 5.145 degrees). Early models for the nutation angles were given in the form (Seidelmann 1992, p.112-114),

$$\Delta\psi = \sum_{i=1}^{n} S_i \sin A_i \,, \qquad \Delta\varepsilon = \sum_{i=1}^{n} C_i \cos A_i \,, \tag{4.32}$$

where the angle,

$$A_i = n_{\ell,i}\ell + n_{\ell' i}\ell' + n_{F,i}F + n_{D,i}D + n_{\Omega,i}\Omega \,, \tag{4.33}$$

represents a linear combination of *fundamental arguments*, being Delaunay variables (angles, or ecliptic coordinates; Vinti 1998) of the sun, moon and their orbital planes on the celestial sphere (Table 4.1). The integer multipliers, $n_{\ell,i},\ldots,n_{\Omega,i}$, correspond to different linear combinations of the fundamental arguments and C_i and S_i are the amplitudes of the periodic terms.

Table 4.1: Fundamental arguments for the nutation angles (Petit and Luzum 2010, Sect. 5.7)

$\ell = 134.96340251^\circ + 1717915923.2178''\tau + 31.8792''\tau^2 + 0.051635''\tau^3 - 0.00024470''\tau^4$
$\ell' = 357.52910918^\circ + 129596581.0481''\tau - 0.5532''\tau^2 + 0.000136''\tau^3 - 0.00001149''\tau^4$
$F = 93.27209062^\circ + 1739527262.8478''\tau - 12.7512''\tau^2 - 0.0001037''\tau^3 + 0.00000417''\tau^4$
$D = 297.85019547^\circ + 1602961601.2090''\tau - 6.3706''\tau^2 + 0.006593''\tau^3 - 0.00003169''\tau^4$
$\mathit{\Omega} = 125.04455501^\circ - 6962890.5431''\tau + 7.4722''\tau^2 + 0.007702''\tau^3 - 0.00005939''\tau^4$

$\ell =$ the mean anomaly of the Moon;
$\ell' =$ the mean anomaly of the Sun;
$F =$ the mean longitude of the Moon minus the mean longitude of the Moon's node;
$D =$ the mean elongation of the Moon from the Sun;
$\mathit{\Omega} =$ the mean longitude of the ascending node of the moon.

The theory and series developed by Woolard (1953) included $n = 69$ terms for $\mathit{\Delta}\psi$ and $n = 40$ terms for $\mathit{\Delta}\varepsilon$. The subsequent theory and series (Kinoshita 1977) adopted by the IAU in 1980, which included modifications for a non-rigid Earth model (Wahr 1985) had $n = 106$ terms. The IAU1980 nutation model was replaced in 2003 by the new nutation model of Mathews et al. (2002), designated IAU2000A (2000B is an abbreviated, less precise version). This model accounts for the mantle anelasticity, the effects of ocean tides, electromagnetic couplings between the mantle, the fluid outer core, and the solid inner core, as well as various non-linear terms not previously considered. A slight revision of the model due to the new IAU 2006 precession model is designated the IAU2000A$_{\text{R06}}$ nutation model, which has $n = 1320$ terms for $\mathit{\Delta}\psi$ and $n = 1037$ terms for $\mathit{\Delta}\varepsilon$ (Petit and Luzum 2010)[1]. This current model is a refined version of equations (4.32),

$$\mathit{\Delta}\psi = \sum_{i=1}^{n} \left(a_i \sin A_i + a_i' \cos A_i \right), \qquad \mathit{\Delta}\varepsilon = \sum_{i=1}^{n} \left(b_i \cos A_i + b_i' \sin A_i \right), \tag{4.34}$$

where, the angle, A_i, includes Delaunay variables for the planets; thus, the angles, $\mathit{\Delta}\psi$, $\mathit{\Delta}\varepsilon$, for now include also *planetary nutations*. The combined IAU 2006/2000A precession-nutation model is accurate to about 0.3 milliarcsec (mas).

Table 4.2 summarizes the largest of the nutation amplitudes and associated variables and parameters according to the model given by equations (4.32). Other terms as in equations (4.34) and contributions from the planets are generally (usually much) less than these. The listed periods of the nutations may be computed from the linear coefficients of the resulting polynomial expressions for the angle, A_i. The frame bias (Sect. 4.1.3) is already incorporated in Table 4.1.

[1] Tables 5.3a,b in the electronic supplement, http://62.161.69.131/iers/conv2010/conv2010_c5.html

Table 4.2: Some terms of the series for nutation in longitude and obliquity, referred to the mean ecliptic of date (IAU2000A$_{\mathrm{R06}}$ nutation model). The time variable, τ, is defined by equation (4.28). The index, i, does not correspond to the index used by the IERS.

i	period [days]	a_i [10^{-6} arcsec]	b_i † [10^{-6} arcsec]	$n_{\ell,i}$	$n_{\ell',i}$	$n_{F,i}$	$n_{D,i}$	$n_{\Omega,i}$
1	6798.4	$-17206424.18 - 17418.82\,\tau$	$9205233.10 + 883.03\,\tau$	0	0	0	0	1
2	182.6	$-1317091.22 - 1369.60\,\tau$	$573033.60 - 458.70\,\tau$	0	0	2	-2	2
3	13.7	$-227641.81 + 279.60\,\tau$	$97846.10 + 137.40\,\tau$	0	0	2	0	2
4	3399.2	$207455.40 - 69.80\,\tau$	$-89749.20 - 29.10\,\tau$	0	0	0	0	2
5	365.3	$147587.70 + 1181.70\,\tau$	$7387.10 - 192.40\,\tau$	0	1	0	0	0
6	27.6	$71115.90 - 87.20\,\tau$	$-675.00 + 35.80\,\tau$	1	0	0	0	0
7	121.7	$-51682.10 - 52.40\,\tau$	$22438.60 - 17.40\,\tau$	0	1	2	-2	2
8	13.6	$-38730.20 + 38.00\,\tau$	$20073.00 + 31.80\,\tau$	0	0	2	0	1
9	9.1	$-30146.40 + 81.60\,\tau$	$12902.60 + 36.70\,\tau$	1	0	2	0	2

The predominant terms in the nutation series have periods of 18.6 years, 0.5 years, and 0.5 months as seen in Table 4.2. Figure 4.7 depicts the motion of the pole due to the combined luni-solar precession and the largest of the nutation terms. This diagram also shows the so-called *nutational ellipse* which describes the extent of the true motion with respect to the mean motion. The "semi-axis" of the ellipse, that is orthogonal to the mean motion, is the principal term in the nutation in obliquity and is also known as the *constant of nutation*. The values for it and for the other "axis", given by $\Delta\psi \sin\varepsilon$ (Figure 4.6), can be inferred from Table 4.2. The total motion of the pole (mean plus true) on the celestial sphere, of course, is due to the superposition of the general precession and all the nutations.

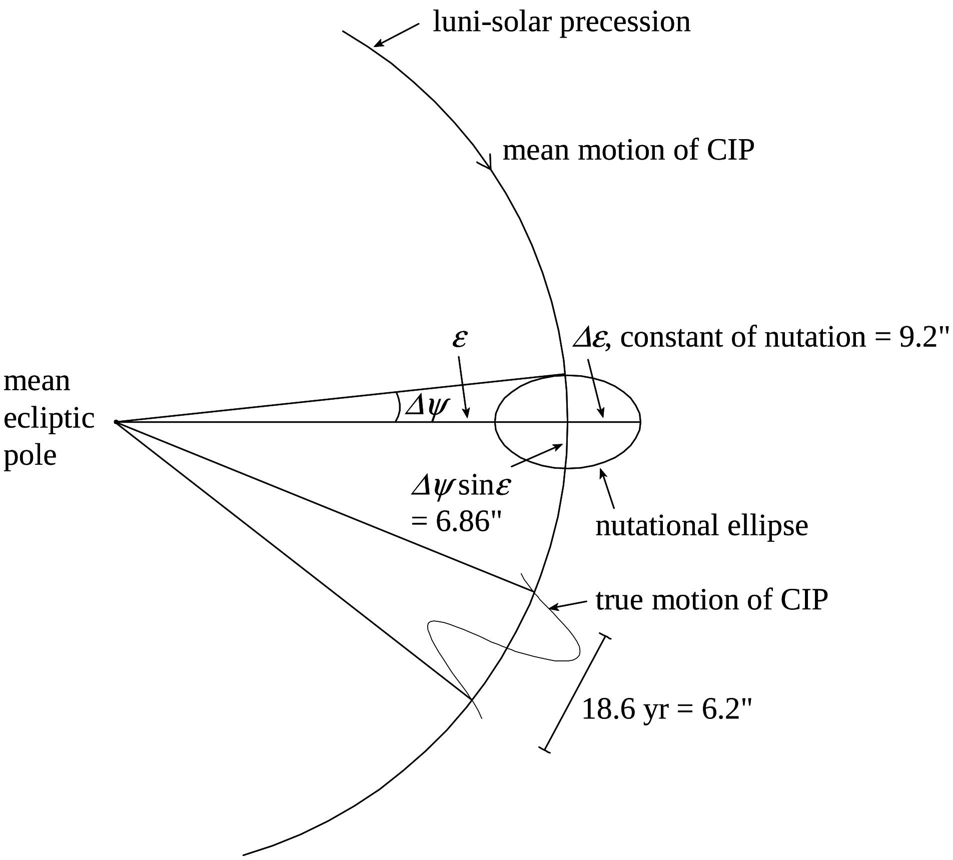


Figure 4.7: Luni-solar precession and nutation.

The transformation at the current epoch (epoch of date) from the mean frame to the true frame accounts for the nutation of the CIP. Referring to Figure 4.6, we see that this transformation is accomplished with the following rotations:

$$\begin{aligned} \boldsymbol{r} &= \mathbf{R}_1\left(-\varepsilon-\Delta\varepsilon\right)\mathbf{R}_3\left(-\Delta\psi\right)\mathbf{R}_1\left(\varepsilon\right)\boldsymbol{r}_m \\ &= \mathbf{N}\boldsymbol{r}_m \end{aligned} \tag{4.35}$$

where ε is the mean obliquity at epoch, t, and

$$\boldsymbol{r} = \begin{pmatrix} \cos\alpha\cos\delta \\ \sin\alpha\cos\delta \\ \sin\delta \end{pmatrix} \tag{4.36}$$

is the vector of coordinates in the true frame at the current epoch. The combined transformation due to precession and nutation from the mean epoch, t_0, to the current epoch, *t*, is given by the combination of equations (4.17) and (4.35):

$$\boldsymbol{r} = \mathbf{NP}\boldsymbol{r}_0 \, . \tag{4.37}$$

Approximate expressions for the nutation matrix, $\mathbf{N}$, can be formulated, for reduced accuracy, since $\Delta\varepsilon$ and $\Delta\psi$ are small angles (Seidelmann, 1992, p.120); in particular, they may be limited to just the principal (largest amplitude) terms. The new convention for the transformation, analogous to equation (4.37), was adopted in 2003 by the IERS and is discussed in Section 4.1.3.

Finally, it is noted that previous and current nutation models are supplemented for those seeking the highest accuracy and temporal resolution by small corrections (called "celestial pole offsets") obtained from continuing VLBI observations. For example, the most recent models do not contain the diurnal motion called free-core nutation caused by the interaction of the mantle and the rotating fluid outer core (Petit and Luzum 2010). IERS publishes differential elements in longitude, $\delta\psi$, and obliquity, $\delta\varepsilon$ (previously also denoted $d\Delta\psi$ and $d\Delta\varepsilon$) that can be added to the elements implied by the nutation series (see also equations (4.56) and (4.57) under the new conventions):

$$\begin{aligned} \Delta\psi &= \Delta\psi(\text{model}) + \delta\psi \\ \Delta\varepsilon &= \Delta\varepsilon(\text{model}) + \delta\varepsilon \end{aligned} \tag{4.38}$$

4.1.3 New Conventions

The method of describing the motion of the CIP on the celestial sphere according to precession and nutation, as given by the matrices in equations (4.17) and (4.35), has been critically analyzed by astronomers, in particular by N. Capitaine (Capitaine et al. 1986, Capitaine 1990) at the Paris Observatory. Several deficiencies in the conventions were indicated especially in light of new and more accurate observations and because of the new kinematical way of defining the *Celestial Reference System* (CRS). Specifically, the separation of motions due to precession and nutation was considered somewhat artificial since no clear distinction can be made between them. Also, with the kinematical definition of the Celestial Reference System, there is no longer any reason to use the mean vernal equinox on the mean ecliptic as an origin of right ascensions. In fact, doing so imparts additional rotations to right ascension due to the rotation of the ecliptic that then must be corrected when considering the rotation of the Earth with respect to inertial space (Greenwich Sidereal Time, or the hour angle at Greenwich of the vernal equinox, see Section 2.3.4; see also Section 5.2.1). Similar "imperfections" were noted when considering the relationship between the CIP and the terrestrial reference system, which will be addressed in Section 4.3.1.1.

In 2000 the International Astronomical Union (IAU) adopted a set of resolutions that precisely adhered to a new, more accurate, and simplified way of dealing with the transformation between the celestial and terrestrial reference systems. The IERS, in 2003, similarly adopted the new methods based on these resolutions (McCarthy and Petit 2003). These were reinforced with IAU resolutions in 2006 and adopted as part of the IERS Conventions 2010. In addition to revising the definitions of the Celestial Ephemeris Pole (CEP), now called the Celestial Intermediate Pole (CIP), the new conventions revised the origins for right ascensions and terrestrial longitude in the intermediate frames associated with the transformations between the Celestial and Terrestrial Reference Systems. The new definitions were designed so as to ensure continuity with the previously defined quantities and to eliminate extraneous residual rotations from their realization. These profoundly different methods and definitions simplify the transformations and solidify the paradigm of *kinematics* (rather than dynamics) upon which the celestial reference system is based. On the other hand the new transformation formulas tend to hide some of the dynamics that lead up to their development.

In essence, the position of the (instantaneous) pole, P, on the celestial sphere at the epoch of date, t, relative to the position at some fundamental epoch, t_0, can be described by two coordinates (very much like polar motion coordinates, see Section 4.3.1) in the celestial system defined by the reference pole, P_0, and by the reference origin of right ascension, Σ_0, as shown in Figure 4.8. In this figure, the pole, P, is displaced from the pole, P_0, and has celestial coordinates, d (co-declination) and E (right ascension). The true (instantaneous) equator (the plane perpendicular to the axis through P) at time, t, intersects the reference equator (associated with P_0) at two nodes that are $180°$ apart. The hour circle of the node, N, is orthogonal to the great circle arc $\overset{\frown}{P_0 P}$; therefore, the right ascension of the ascending node of the equator is $90°$ plus the right ascension of the instantaneous pole, P. The origin for right ascension at the epoch of date, t, is defined kinematically under the condition that there is no rotation *rate* in the *instantaneous coordinate system* about the pole due to precession and nutation. This is the concept of the so-called *non-rotating origin* (NRO) that is now also used to define the instantaneous origin for terrestrial longitudes (see Section 4.3.1.1). This origin for right ascensions on the instantaneous equator is now called the *Celestial Intermediate Origin* (CIO), denoted σ in Figure 4.8 (it has also been called the Celestial Ephemeris Origin, CEO).

Rather than successive transformations involving precessional elements and nutation angles, the transformation is more direct in terms of the coordinates, (d, E) (reformulated in terms of Cartesian-like elements, X and Y; see below), and the additional parameter, *s*, that defines the instantaneous origin of right ascensions. The transformation from the celestial reference frame to the instantaneous celestial frame is

$$\boldsymbol{r} = \mathbf{Q}^{\mathrm{T}} \boldsymbol{r}_0, \tag{4.39}$$

where

$$\mathbf{Q}^{\mathrm{T}} = \mathbf{R}_3\left(-s\right)\mathbf{R}_3\left(-E\right)\mathbf{R}_2\left(d\right)\mathbf{R}_3\left(E\right), \tag{4.40}$$

which is easily derived by considering the successive rotations as the origin point transforms from the CRS origin, Σ_0, to the instantaneous origin, σ (Figure 4.8). Equation (4.39) essentially replaces equation (4.37), but also incorporates the new conventions for defining the origin in right ascension. The exact relationship to the previously defined transformation is given later in Section 5.2.1. We adhere to the notation used in the IERS Conventions 2003, which defines $\mathbf{Q}$ as the transformation *from* the system of the instantaneous pole and origin *to* the CRS.

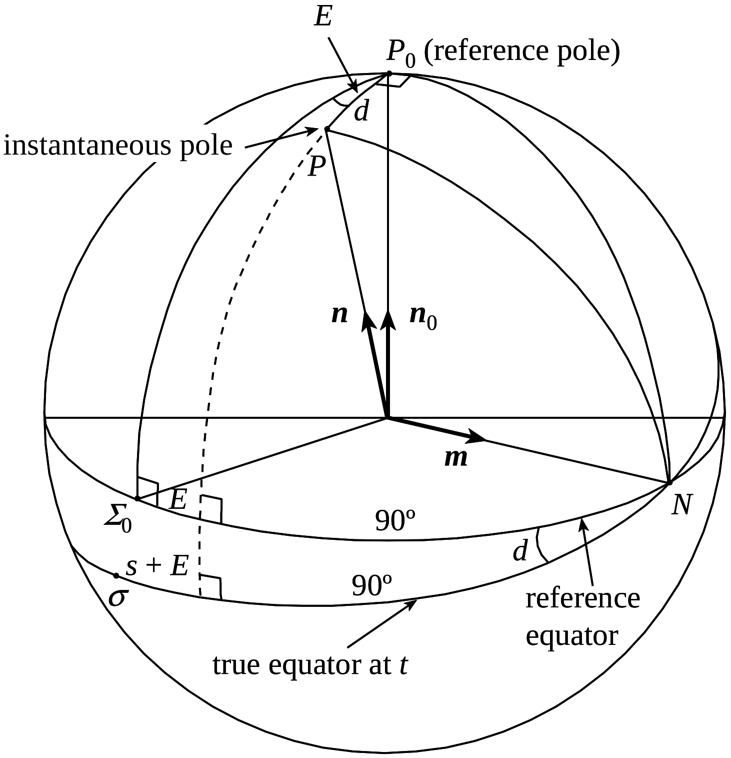


Figure 4.8: Coordinates of instantaneous pole in the celestial reference system.

It remains to determine the parameter, s. The total rotation rate of the pole, P, in inertial space is due to changes in the coordinates, $\left(d, E\right)$, and in the parameter, s. Defining three non-co-linear unit vectors, $\boldsymbol{n}_0$, $\boldsymbol{m}$, $\boldsymbol{n}$, essentially associated with these quantities, as shown in Figure 4.8, we may express the total rotation rate as follows:

$$\boldsymbol{\Theta} = \boldsymbol{n}_0 \dot{E} + \boldsymbol{m}\dot{d} - \boldsymbol{n}\left(\dot{E} + \dot{s}\right), \tag{4.41}$$

where the dots denote time-derivatives. Now, s is chosen so that the total rotation rate, $\boldsymbol{\Theta}$, has no component along $\boldsymbol{n}$. That is, s defines the origin point, $\boldsymbol{\sigma}$, on the instantaneous equator that has no rotation rate about the corresponding polar axis (*non-rotating origin*). This condition is formulated as $\boldsymbol{\Theta} \cdot \boldsymbol{n} = 0$, meaning that there is no component of the total rotation rate along the instantaneous polar axis. Therefore,

$$0 = \boldsymbol{n} \cdot \boldsymbol{n}_0 \dot{E} + \boldsymbol{n} \cdot \boldsymbol{m}\dot{d} - \left(\dot{E} + \dot{s}\right); \tag{4.42}$$

and, since $\boldsymbol{n} \cdot \boldsymbol{m} = 0$, $\boldsymbol{n} \cdot \boldsymbol{n}_0 = \cos d$, we have

$$\dot{s} = \left(\cos d - 1\right)\dot{E}\,. \tag{4.43}$$

For convenience, we define coordinates X, Y, and Z

$$\begin{pmatrix} X \\ Y \\ Z \end{pmatrix} = \begin{pmatrix} \sin d \cos E \\ \sin d \sin E \\ \cos d \end{pmatrix}. \tag{4.44}$$

Then, it is easily shown that

$$X\dot{Y} - Y\dot{X} = -\dot{E}\left(\cos^2 d - 1\right); \tag{4.45}$$

and, substituting this together with $Z = \cos d$ into equation (4.43) and integrating yields

$$s = s_0 - \int_{t_0}^{t} \frac{X\dot{Y} - Y\dot{X}}{1 + Z}\, dt\,, \tag{4.46}$$

where $s_0 = s\left(t_0\right)$ is chosen so as to ensure continuity with the previous definition of the origin point at the epoch 1 January 2003.

The transformation matrix, $\mathbf{Q}$, equation (4.40), is given more explicitly by:

$$\mathbf{Q}=\begin{pmatrix} 1-\cos^2 E\left(1-\cos d\right) & -\sin E\cos E\left(1-\cos d\right) & \sin d\cos E \\ -\sin E\cos E\left(1-\cos d\right) & 1-\sin^2 E\left(1-\cos d\right) & \sin d\sin E \\ -\sin d\cos E & -\sin d\sin E & \cos d \end{pmatrix}\mathbf{R}_3\left(s\right). \tag{4.47}$$

With the coordinates, $\left(X,Y,Z\right)$, defined by equation (4.44), and $1-\cos d = a\sin^2 d$, where $a = 1\big/\left(1+\cos d\right)$, it is easy to derive that

$$\mathbf{Q}=\begin{pmatrix} 1-aX^2 & -aXY & X \\ -aXY & 1-aY^2 & Y \\ -X & -Y & 1-a\left(X^2+Y^2\right) \end{pmatrix}\mathbf{R}_3\left(s\right), \tag{4.48}$$

Expressions for X and Y can be obtained directly from precession and nutation equations with respect to the celestial system; see references mentioned in Section 4 of (Capitaine, 1990). For the latest IAU 2006/2000A precession-nutation models, Petit and Luzum (2010) give the following:

$$\begin{aligned} X =& -0.016617 + 2004.191898\,\tau - 0.4297829\,\tau^2 \\ & -0.19861834\,\tau^3 + 0.000007578\,\tau^4 + 0.0000059285\,\tau^5 \\ & +\sum_j\Big(\left(a_{s,0}\right)_j \sin\left(\text{ARGUMENT}\right) + \left(a_{c,0}\right)_j \cos\left(\text{ARGUMENT}\right)\Big) \\ & +\sum_j\Big(\left(a_{s,1}\right)_j \tau\sin\left(\text{ARGUMENT}\right) + \left(a_{c,1}\right)_j \tau\cos\left(\text{ARGUMENT}\right)\Big) \\ & +\sum_j\Big(\left(a_{s,2}\right)_j \tau^2\sin\left(\text{ARGUMENT}\right) + \left(a_{c,2}\right)_j \tau^2\cos\left(\text{ARGUMENT}\right)\Big) + \cdots \text{ [arcsec]} \end{aligned} \tag{4.49}$$

$$\begin{aligned} Y =& -0.006951 - 0.025896\,\tau - 22.4072747\,\tau^2 \\ & +0.00190059\,\tau^3 + 0.001112526\,\tau^4 + 0.0000001358\,\tau^5 \\ & +\sum_j\Big(\left(b_{s,0}\right)_j \sin\left(\text{ARGUMENT}\right) + \left(b_{c,0}\right)_j \cos\left(\text{ARGUMENT}\right)\Big) \\ & +\sum_j\Big(\left(b_{s,1}\right)_j \tau\sin\left(\text{ARGUMENT}\right) + \left(b_{c,1}\right)_j \tau\cos\left(\text{ARGUMENT}\right)\Big) \\ & +\sum_j\Big(\left(b_{s,2}\right)_j \tau^2\sin\left(\text{ARGUMENT}\right) + \left(b_{c,2}\right)_j \tau^2\cos\left(\text{ARGUMENT}\right)\Big) + \cdots \text{ [arcsec]} \end{aligned} \tag{4.50}$$

where $\tau = (t - t_0)/36525$ with t and t_0 the Julian day numbers for the epoch of date and J2000.0, respectively; and, the coefficients $(a_{s,k})_j$, $(a_{c,k})_j$ $(b_{s,k})_j$, $(b_{c,k})_j$ are available[2] in tabulated form for each of the corresponding fundamental arguments, ARGUMENT, of the nutation model. These arguments are the same as given in equation (4.33). A full description is given by Petit and Luzum (ibid., Section 5.7).

Also, for the parameter, s, the following includes all terms larger than $0.5\ \mu\text{arcsec}$, as well as the constant, s_0:

$$\begin{aligned} s = &-\frac{1}{2}XY + 94 + 3808.65\tau - 122.68\tau^2 - 72574.11\tau^3 \\ &+\sum_k C_k \sin\alpha_k + \sum_k D_k \tau \sin\beta_k + \sum_k E_k \tau \cos\gamma_k + \sum_k F_k \tau^2 \sin\theta_k \ \ [\mu\text{arcsec}] \end{aligned} \tag{4.51}$$

where the coefficients, C_k, D_k, E_k, F_k, and the arguments, α_k, β_k, γ_k, θ_k, are elaborated by Petit and Luzum (ibid., Chapter 5, p.59).

We note that the newly adopted IAU 2006/2000A model for precession and nutation (on which expressions (4.49), (4.50), and (4.51) are based) replace the IAU 2000 model (and, of course, the old IAU 1976 precession and IAU 1980 nutation models). The new models are described in detail in (ibid.) and yield accuracy of about $0.3\ \text{mas}$ in the position of the pole. Furthermore, these transformation equations referring to the kinematic pole of the ICRS incorporate the “frame bias” described below.

To see how the coordinates, d, E, are related to the traditional precession and nutation angles, it is necessary to consider how the Celestial Reference System was defined prior to the new, current kinematic definition. The dynamic definition was based on the mean equator and mean equinox at a certain fundamental epoch, t_0. Recall that the precession and nutation of the equator relative to the mean ecliptic at t_0 is due to the accumulated luni-solar precessions in longitude, ψ_A, and in the obliquity of the ecliptic, ω_A (which differs from ε_A by the rotation of the mean ecliptic; see Figure 4.4), as well as the nutations, $\Delta\psi_1$ and $\Delta\varepsilon_1$, in longitude and in the obliquity at t_0 (again, differing from corresponding quantities at t). Let $\bar{d}, \bar{E}$ be coordinates, similar to d, E, of the instantaneous pole in the dynamic mean system. Then, defining $\bar{X}, \bar{Y}, \bar{Z}$ similar to X, Y, Z, it is easy to derive the following identity from the laws of sines and cosines applied to the spherical triangle, $\Upsilon_0 \Upsilon_1 \bar{N}$, in Figure 4.9:

$$\begin{pmatrix} \bar{X} \\ \bar{Y} \\ \bar{Z} \end{pmatrix} = \begin{pmatrix} \sin\bar{d}\cos\bar{E} \\ \sin\bar{d}\sin\bar{E} \\ \cos\bar{d} \end{pmatrix} = \begin{pmatrix} \sin(\omega_A + \Delta\varepsilon_1)\sin(\psi_A + \Delta\psi_1) \\ \sin(\omega_A + \Delta\varepsilon_1)\cos(\psi_A + \Delta\psi_1)\cos\varepsilon_0 - \cos(\omega_A + \Delta\varepsilon_1)\sin\varepsilon_0 \\ \sin(\omega_A + \Delta\varepsilon_1)\cos(\psi_A + \Delta\psi_1)\sin\varepsilon_0 + \cos(\omega_A + \Delta\varepsilon_1)\cos\varepsilon_0 \end{pmatrix}. \tag{4.52}$$

[2] ftp://tai.bipm.org/iers/conv2010/chapter5/

Further expansions of $\bar{X}$ and $\bar{Y}$ as series derivable from series expansions for the quantities, ψ_A, ω_A, $\Delta\psi_1$, and $\Delta\varepsilon_1$ may be found in Capitaine (1990).

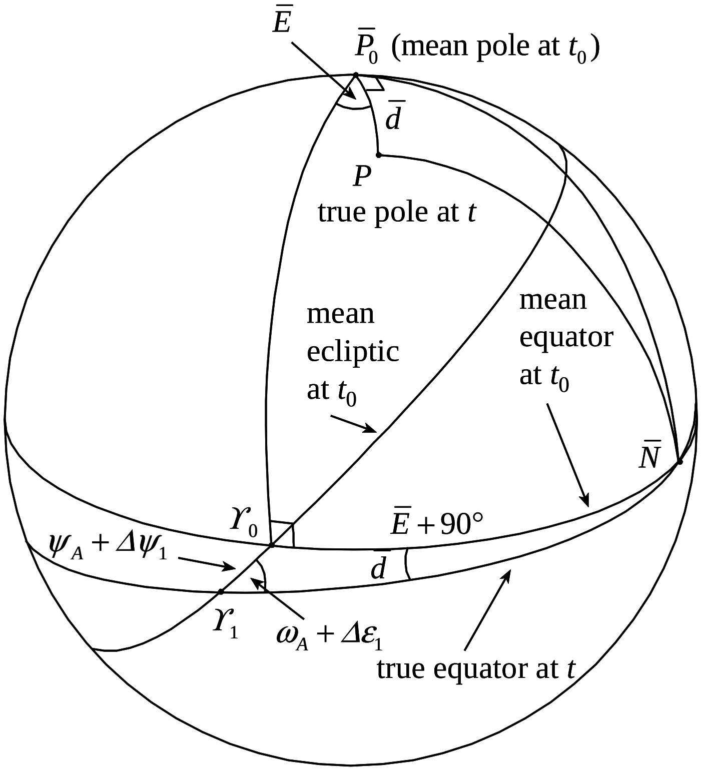


Figure 4.9: Coordinates of the true pole at *t* in the dynamic system of t_0.

The dynamic reference pole, $\bar{P}_0$, of the previous realization (FK5 catalogue) is offset from the kinematic pole of the ICRS, as shown in Figure 4.10, by small angles, ξ_0 in X and η_0 in Y. Also, a small rotation, $d\alpha_0$, separates the dynamic reference equinox from the origin of the ICRS. These offsets, called *frame bias*, are defined for the mean dynamic system in the ICRS, so that the transformation between $(\bar{X},\bar{Y},\bar{Z})$ and (X,Y,Z) is given by

$$\begin{pmatrix}\bar{X}\\ \bar{Y}\\ \bar{Z}\end{pmatrix} = R_1(-\eta_0)R_2(\xi_0)R_3(d\alpha_0)\begin{pmatrix}X\\ Y\\ Z\end{pmatrix}$$
$$= \begin{pmatrix}1 & d\alpha_0 & -\xi_0\\ -d\alpha_0 & 1 & -\eta_0\\ \xi_0 & \eta_0 & 1\end{pmatrix}\begin{pmatrix}X\\ Y\\ Z\end{pmatrix} \tag{4.53}$$

where the approximation, equation (1.9), was used. Or, setting $Z \approx 1$, and neglecting second-order terms,

$$X = \bar{X} + \xi_0 - d\alpha_0 \bar{Y}$$
$$Y = \bar{Y} + \eta_0 + d\alpha_0 \bar{X} \tag{4.54}$$

McCarthy and Petit (2003, Ch.5, p.9,12) give the following values for these offsets based on the IAU 2000 nutation model (they have not changed for the IAU 2006/2000A model);

$$\begin{aligned} \xi_0 &= -16.6170 \pm 0.01 \text{ mas} \\ \eta_0 &= -6.8192 \pm 0.01 \text{ mas} \\ d\alpha_0 &= -14.60 \pm 0.05 \text{ mas} \end{aligned} \tag{4.55}$$

The rotation, $d\alpha_0$, refers to the offset of the mean dynamic equinox of an ecliptic interpreted as being inertial (i.e., not rotating). In the past, the rotating ecliptic was used to define the dynamic equinox. The difference (due to a Coriolis term) between the two equinoxes is about 93.7 milliarcsec (Standish, 1981), so care in definition must be exercised when applying the transformation, equations (4.54), with values given by equations (4.55). Note that Figure 4.10 only serves to *define* the offsets according to equation (4.54), but does not show the actual numerical relationships (equations (4.55)) between the ICRS and the CEP(J2000.0) since the offsets are negative. Again, these offsets are already included in the expressions (4.49) and (4.50) for X and Y.

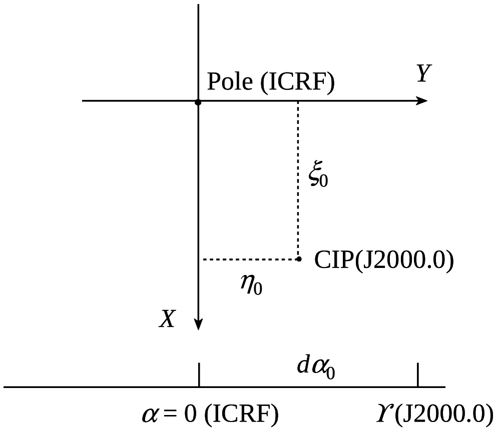


Figure 4.10: Definition of offset parameters of dynamic mean system in the ICRS.

In summary, the differences between the former precession/nutation transformation matrix, $\mathbf{NP}$, equations (4.17) and (4.35), and the new transformation matrix, $\mathbf{Q}^{\mathrm{T}}$, equation (4.40), is twofold: 1) $\mathbf{NP}$ refers to the equinox, while $\mathbf{Q}^{\mathrm{T}}$ refers to the non-rotating origin, which is used to defined the unadulterated Earth rotation angle (Section 5.2.1); and, 2) $\mathbf{NP}$ refers to the dynamic reference pole, while $\mathbf{Q}^{\mathrm{T}}$ refers to the kinematic reference pole, which gives rise to the frame bias.

The celestial pole offsets in longitude and obliquity, $(\delta\psi, \delta\varepsilon)$, that correct the IAU 2006/2000A precession-nutation model on the basis of VLBI observations are not included, however, and must be added. The corrections are published by IERS in terms of corrections to X and Y. The coordinates of the CEP thus are (Petit and Luzum, 2010, Ch.5, p.57)

$$X = X\left(\text{IAU 2006/2000A}\right) + \delta X, \qquad Y = Y\left(\text{IAU 2006/2000A}\right) + \delta Y, \tag{4.56}$$

where

$$\begin{aligned} \delta X &= \delta\psi \sin\varepsilon_A + \left(\psi_A \cos\varepsilon_A - \chi_A\right)\delta\varepsilon \\ \delta Y &= \delta\varepsilon - \left(\psi_A \cos\varepsilon_A - \chi_A\right)\delta\psi \sin\varepsilon_A \end{aligned} \tag{4.57}$$

4.1.4 Problems

1. a) Make a rough estimate of the present declination and right ascension of the vernal equinox in 120 B.C., the date when precession was discovered.

b) Determine the mean coordinates at J1950.0 of the vernal equinox of the celestial frame defined at J2000.0. Then determine the mean coordinates at J2000.0 of the vernal equinox of the celestial frame defined at J1950.0. In both cases use the precession expressions derived for the constants defined at the fundamental epoch J2000.0. Compare the precessional elements in each case and compare the resulting coordinates. Use 10-digit precision in your computations.

2. The coordinates of a star at J2000.0 are: $\alpha = 16\text{ hr } 56\text{ min } 12.892\text{ sec}$, $\delta = 82°12'39.03''$. Determine the accumulated precession of the star in right ascension during the year 2001 using the precession transformation formula.

3. Give a procedure (flow chart with clearly identified input, processing, and output) that transforms coordinates of a celestial object given in the celestial reference system of 1900 (1900 constants of precession) to its present *true* coordinates. Be explicit in describing the epochs for each component of the transformation and give the necessary equations.

4. Derive the following: equation (4.46) starting with equation (4.43); equation (4.48) starting with equation (4.40); and equation (4.52).

5. Show that

$$a = \frac{1}{2} + \frac{1}{8}\left(X^2 + Y^2\right) + \cdots, \tag{4.58}$$

where a is defined after equation (4.47).

6. Derive equations (4.57).

4.2 Observational Systematic Effects

The following sections deal with effects that need to be corrected in order to determine true coordinates of celestial objects from observed, or apparent coordinates. These effects are due more to the kinematics of the observer and the objects being observed than the dynamics of Earth's motion.

4.2.1 Proper Motion

Proper motion refers to the actual motion of celestial objects with respect to inertial space. As such their coordinates will be different at the time of observation than what they are in some fundamental reference frame that refers to an epoch, t_0. We consider only the motion of stars and not of planets, since the former were used historically to determine coordinates of points on the Earth (Section 2.3.5) and still today to orient satellite systems to the inertial frame via on-board star cameras. Proper motion, also known as *space motion* and *stellar motion*, can be decomposed into motion on the celestial sphere (tangential motion) and radial motion. Radial stellar motion would be irrelevant if the Earth had no orbital motion (see the effect of parallax in Section 4.2.3).

Accounting for proper motion is relatively simple and requires only that rates be given in right ascension, in declination, and in the radial direction (with respect to a particular celestial reference frame). If $\boldsymbol{r}(t_0)$ is the vector of coordinates of a star in a catalogue (celestial reference frame) for fundamental epoch, t_0, then the coordinate vector at the current epoch, t, is given by

$$\boldsymbol{r}(t)=\boldsymbol{r}(t_0)+(t-t_0)\dot{\boldsymbol{r}}(t_0), \tag{4.59}$$

where this linearization is sufficiently accurate because the proper motion, $\dot{\boldsymbol{r}}$, is very small (by astronomic standards). With

$$\boldsymbol{r}=\begin{pmatrix} r\cos\delta\cos\alpha \\ r\cos\delta\sin\alpha \\ r\sin\delta \end{pmatrix}, \tag{4.60}$$

where α and δ are right ascension and declination, as usual, and $r=|\boldsymbol{r}|$, we have

$$\dot{\boldsymbol{r}}=\begin{pmatrix} \dot{r}\cos\delta\cos\alpha-r\dot{\alpha}\cos\delta\sin\alpha-r\dot{\delta}\sin\delta\cos\alpha \\ \dot{r}\cos\delta\sin\alpha+r\dot{\alpha}\cos\delta\cos\alpha-r\dot{\delta}\sin\delta\sin\alpha \\ \dot{r}\sin\delta+r\dot{\delta}\cos\delta \end{pmatrix}. \tag{4.61}$$

The units of proper motion in right ascension and declination, $\dot{\alpha}$ and $\dot{\delta}$, typically are rad/century and for the radial velocity, $\dot{r}$, the units are AU/century, where 1 AU is one astronomical unit, the mean radius of Earth's orbit:

$$1\text{ AU} = 1.4959787066\times 10^{11}\text{ m;} \qquad 1\text{ km/s} = 21.095\text{ AU/century}\,. \tag{4.62}$$

The radial distance is given as (see Figure 4.11, where $r_S \equiv r$)

$$r = \frac{1\text{ AU}}{\sin\pi}\,, \tag{4.63}$$

where π is called the *parallax angle* (see Section 4.2.3). This is the angle subtended at the object by the semi-major axis of Earth's orbit. If this angle is unknown or insignificant (e.g., because the star is at too great a distance), then the radial velocity is not important. Also, if linear approximation is sufficient then one may correct the coordinates of the star for proper motion according to

$$\begin{aligned} \alpha(t) &= \alpha(t_0) + (t - t_0)\dot{\alpha}(t_0) \\ \delta(t) &= \delta(t_0) + (t - t_0)\dot{\delta}(t_0) \end{aligned} \tag{4.64}$$

For further implementation of proper motion corrections, see Section 4.3.3.

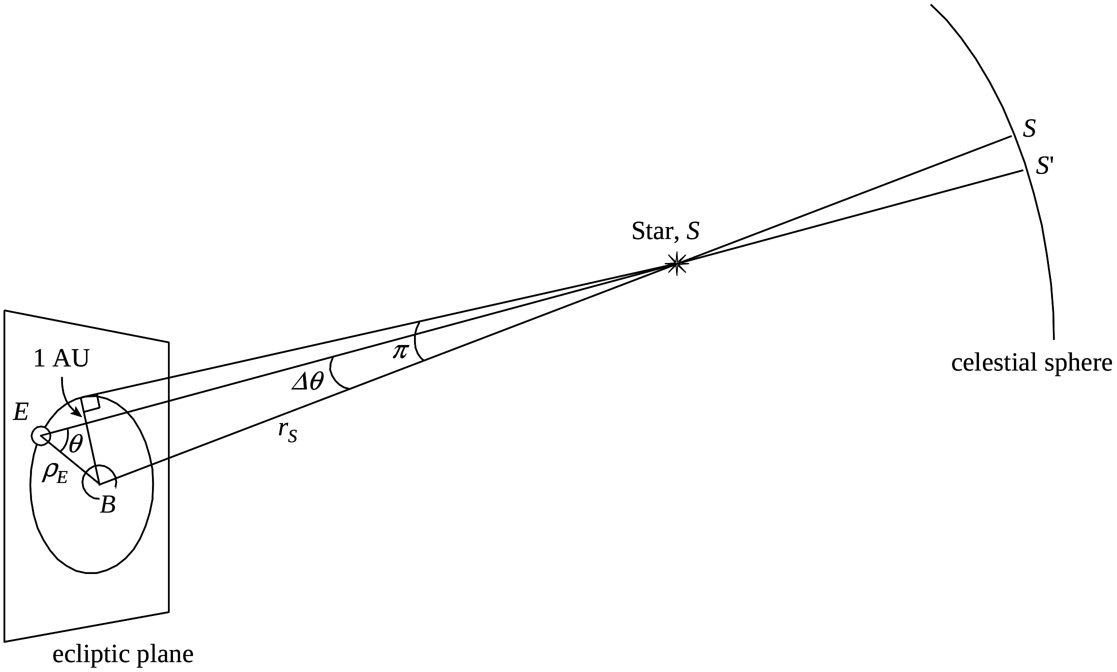


Figure 4.11: Geometry of star with respect to solar system. See also Figure 4.13 for the geometry on the celestial sphere.

4.2.2 Aberration

Aberration is a displacement of the apparent object from its true position on the celestial sphere due to the velocity of the observer and the finite speed of light. The classic analog is the apparent slanted direction of vertically falling rain as viewed from a moving vehicle; the faster the vehicle, the more slanted is the apparent direction of the falling rain. Likewise, the direction of incoming light from a star is distorted if the observer is moving at a non-zero angle with respect to the true direction (see Figure 4.12). In general, the apparent coordinates of a celestial object deviate from the true coordinates as a function of the observer's velocity with respect to the direction of the celestial object.

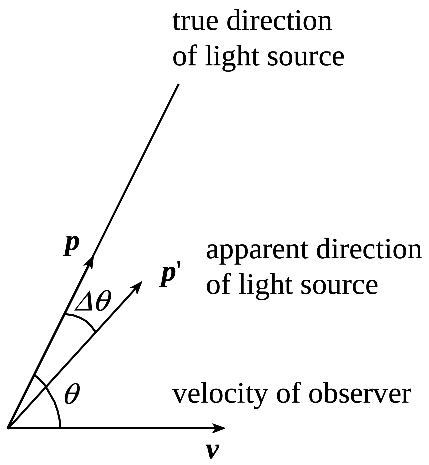


Figure 4.12: The effect of aberration.

Diurnal aberration is due to the observer's velocity associated with Earth's rotation; and, *annual aberration* is due to the observer's velocity associated with Earth's orbital motion (there is also *secular aberration* due to the velocity of the solar system, but this is not observable–it is a constant). These aberrations are grouped as *stellar aberrations*, as opposed to *planetary aberrations*, where the motions of both the observer and the celestial body are considered. We do not consider planetary aberration. Furthermore, aberration differs from the *light-time effect* that accounts for the distance the light must travel from the time it is emitted to the time it is actually observed (thus, again, the apparent coordinates of the object are not the same as the true coordinates). This effect must be considered for planets, and it is familiar to those who process GPS data, but for stars this makes little sense since most stars are hundreds and thousands of light-years distant.

We treat stellar aberration using Newtonian physics, and only mention the special relativistic effect. Accordingly, the direction of the source will appear to be displaced in the direction of the velocity of the observer (Figure 4.12). That is, suppose in a stationary frame the light is coming from the direction given by the unit vector, $\boldsymbol{p}$. Then, in the frame moving with velocity, $\boldsymbol{v}$, the light appears to originate from the direction defined by the unit vector, $\boldsymbol{p}'$, which is proportional to the vector sum of the two velocities, $\boldsymbol{v}$ and $c\boldsymbol{p}$:

$$\boldsymbol{p}' = \frac{\boldsymbol{v} + c\boldsymbol{p}}{\left|\boldsymbol{v} + c\boldsymbol{p}\right|}, \tag{4.65}$$

where c is the speed of light (in vacuum). Taking the cross-product on both sides with $\boldsymbol{p}$ and extracting the magnitudes, we obtain, with $\left|\boldsymbol{p}\times\boldsymbol{p}'\right| = \sin\Delta\theta$, $\left|\boldsymbol{p}\times\boldsymbol{v}\right| = v\sin\theta$, and $\left|\boldsymbol{p}\times\boldsymbol{p}\right| = 0$, the following:

$$\begin{aligned}\sin \mathit{\Delta}\theta &= \frac{v\sin\theta}{\left|\boldsymbol{v}+c\boldsymbol{p}\right|}\\ &= \frac{v\sin\theta}{\sqrt{v^2+c^2+2vc\cos\theta}}\\ &= \frac{v}{c}\sin\theta+\cdots\end{aligned}\tag{4.66}$$

where v is the magnitude of the observer's velocity, and higher powers of v/c are neglected. Accounting for the effects of special relativity, Seidelmann (1992, p.129) gives the second-order formula:

$$\sin\mathit{\Delta}\theta = \frac{v}{c}\sin\theta - \frac{1}{4}\left(\frac{v}{c}\right)^2 \sin 2\theta + \cdots .\tag{4.67}$$

Realizing that the aberration angle is relatively small, we use the approximate formula:

$$\mathit{\Delta}\theta = \frac{v}{c}\sin\theta .\tag{4.68}$$

With respect to Figure 4.13, let S denote the true position of the star on the celestial sphere with true coordinates, $\left(\delta_S,\alpha_S\right)$, and let S' denote the apparent position of the star due to aberration with corresponding aberration errors, $\mathit{\Delta}\delta$ and $\mathit{\Delta}\alpha$, in declination and right ascension. Note that S' is on the great circle arc, $\overset{\frown}{SF}$, where F denotes the point on the celestial sphere in the direction of the observer's velocity (that is, the aberration angle is in the plane defined by the velocity vectors of the observer and the incoming light). By definition:

$$\begin{aligned}\delta_S &= \delta_{S'} - \mathit{\Delta}\delta\\ \alpha_S &= \alpha_{S'} - \mathit{\Delta}\alpha\end{aligned}\tag{4.69}$$

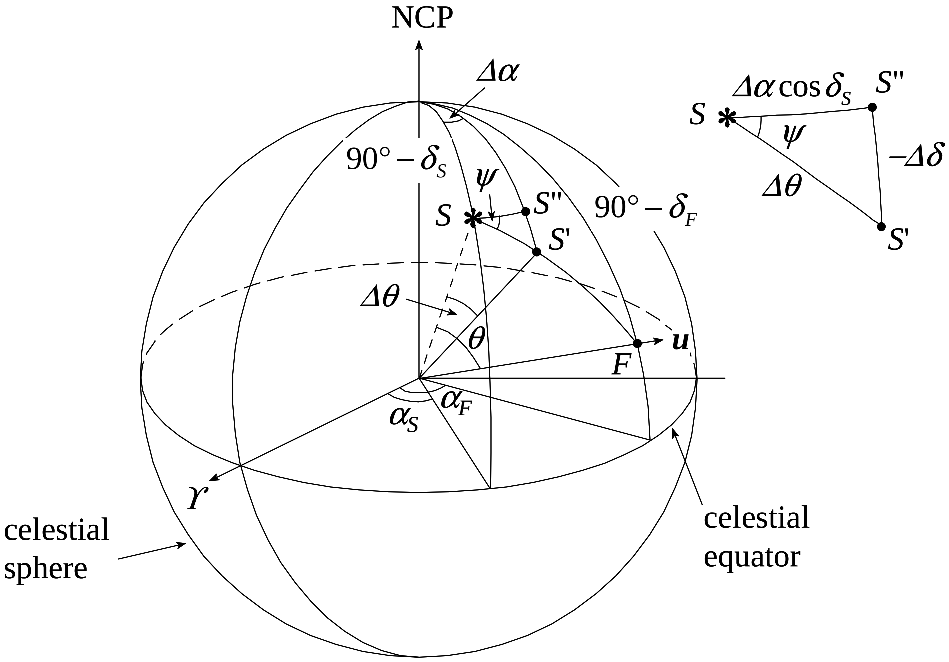


Figure 4.13: Geometry on the celestial sphere for aberration and parallax. For aberration, $\boldsymbol{u} = \boldsymbol{v} =$ velocity of the observer; for parallax, $\boldsymbol{u} = \boldsymbol{e}_B =$ direction of barycenter.

We have from the small triangle, $SS'S''$:

$$\cos\psi = \frac{\Delta\alpha\cos\delta_S}{\Delta\theta}, \tag{4.70}$$

and

$$\sin\psi = -\frac{\Delta\delta}{\Delta\theta}. \tag{4.71}$$

From triangle $S - NCP - F$, by the law of sines, we have

$$\sin\theta\cos\psi = \cos\delta_F \sin\left(\alpha_F - \alpha_S\right), \tag{4.72}$$

where the coordinates of F are $\left(\delta_F, \alpha_F\right)$. Substituting equation (4.70) into equation (4.68) and using equation (4.72) yields

$$\begin{aligned}\varDelta\alpha\cos\delta_S &= \frac{v}{c}\sin\theta\cos\psi \\ &= \frac{v}{c}\cos\delta_F\sin\left(\alpha_F-\alpha_S\right) \\ &= \frac{v}{c}\cos\delta_F\left(\sin\alpha_F\cos\alpha_S-\cos\alpha_F\sin\alpha_S\right)\end{aligned} \tag{4.73}$$

Now, the velocity, $\boldsymbol{v}$, of the observer, in the direction F on the celestial sphere, can be expressed as

$$\boldsymbol{v} = \begin{pmatrix}\dot{x}\\ \dot{y}\\ \dot{z}\end{pmatrix} = \begin{pmatrix} v\cos\delta_F\cos\alpha_F \\ v\cos\delta_F\sin\alpha_F \\ v\sin\delta_F\end{pmatrix}, \tag{4.74}$$

where $v=|\boldsymbol{v}|$. Hence, using equation (4.74) in equation (4.73), the effect of aberration on right ascension is given by

$$\varDelta\alpha = \left(\frac{\dot{y}}{c}\cos\alpha_S-\frac{\dot{x}}{c}\sin\alpha_S\right)\sec\delta_S\,. \tag{4.75}$$

For the declination, we find, again from the triangle, $S-NCP-F$, now by the law of cosines, that:

$$\sin\delta_F = \sin\delta_S\cos\theta-\cos\delta_S\sin\theta\sin\psi\,. \tag{4.76}$$

Also, with the unit vector defining the position of the star on the celestial sphere,

$$\boldsymbol{p} = \begin{pmatrix}\cos\delta_S\cos\alpha_S\\ \cos\delta_S\sin\alpha_S\\ \sin\delta_S\end{pmatrix}, \tag{4.77}$$

we have the scalar product, using equation (4.74):

$$\begin{aligned}\boldsymbol{p}\cdot\boldsymbol{v} &= v\cos\theta \\ &= \dot{x}\cos\delta_S\cos\alpha_S+\dot{y}\cos\delta_S\sin\alpha_S+\dot{z}\sin\delta_S\end{aligned} \tag{4.78}$$

We solve equation (4.78) for $\cos\theta$ and substitute this into equation (4.76), which is then solved for $\sin\theta\sin\psi$ to get

$$\sin\theta\sin\psi = \frac{\dot{x}}{v}\sin\delta_S\cos\alpha_S + \frac{\dot{y}}{v}\sin\delta_S\sin\alpha_S - \frac{\dot{z}}{v}\cos\delta_S \,. \tag{4.79}$$

From equations (4.71) and (4.68), we finally have

$$\Delta\delta = -\frac{\dot{x}}{c}\sin\delta_S\cos\alpha_S - \frac{\dot{y}}{c}\sin\delta_S\sin\alpha_S + \frac{\dot{z}}{c}\cos\delta_S \,. \tag{4.80}$$

For diurnal aberration, the observer (assumed stationary on the Earth's surface) has only eastward velocity with respect to the celestial sphere due to Earth's rotation rate, ω_e; it is given by (see Figure 4.14):

$$v = \omega_e\left(N+h\right)\cos\phi \,, \tag{4.81}$$

where N is the ellipsoid radius of curvature in the prime vertical and $\left(\phi, h\right)$ are the geodetic latitude and ellipsoid height of the observer (see Section 2.1.3.1). In this case (see Figure 4.15):

$$\begin{aligned} \dot{x} &= v\cos\left(\alpha_S + t_S - 270^\circ\right) \\ \dot{y} &= v\sin\left(\alpha_S + t_S - 270^\circ\right) \\ \dot{z} &= 0 \end{aligned} \tag{4.82}$$

where t_S is the hour angle of the star. Substituting equations (4.82) into equations (4.75) and (4.80), we find the diurnal aberration effects, respectively, in right ascension and declination to be:

$$\begin{aligned} \Delta\alpha &= \frac{v}{c}\cos t_S \sec\delta_S \\ \Delta\delta &= \frac{v}{c}\sin t_S \sec\delta_S \end{aligned} \tag{4.83}$$

In order to appreciate the magnitude of the effect of diurnal aberration, consider, using equation (4.81), that

$$\frac{v}{c} = \frac{a\omega_e}{c}\frac{N+h}{a}\cos\phi = 0.3200\frac{N+h}{a}\cos\phi \text{ [arcsec]}, \tag{4.84}$$

which is also called the “constant of diurnal aberration”. Diurnal aberration, thus, is always less than about 0.32 arcsec.

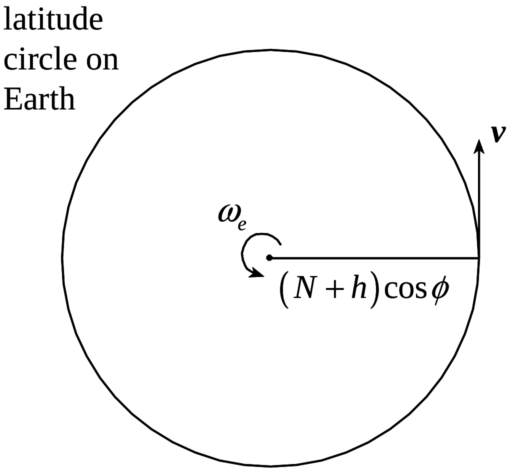


Figure 4.14: Velocity of terrestrial observer for diurnal aberration.

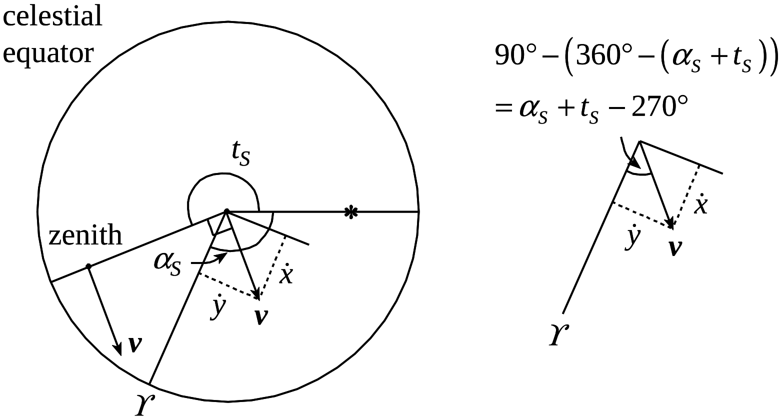


Figure 4.15: Celestial geometry for diurnal aberration.

Annual aberration, on the other hand, is two orders of magnitude larger! In this case, the velocity of the observer is due to Earth’s orbital motion and the velocity vector is in the ecliptic plane. The ”constant of annual aberration” is given by

$$\frac{v}{c} = \frac{2\pi \text{ AU/yr}}{3\times10^8 \text{ m/s}} \approx 10^{-4} = 20 \text{ arcsec}. \tag{4.85}$$

From this, one can determine (left to the reader) how accurately Earth's velocity must be known in order to compute the annual aberration to a given accuracy. Accurate velocity components are given in the Astronomical Almanac (Section B, p.44) in units of 10^{-9} AU/day in the barycentric system. Note that the second-order effect, given in equation (4.67), amounts to no more than:

$$\frac{1}{4}\left(\frac{v}{c}\right)^2 \approx 0.25\times 10^{-8} = 5\times 10^{-4} \text{ arcsec}. \tag{4.86}$$

We further note that, aside from the approximations in equations (4.75) and (4.80), other approximations could be considered in deriving the annual aberration formulas, e.g., taking Earth's orbit to be circular. In this case, corrections may be necessary to account for the actual non-constant speed along the elliptical orbit. Also, if the velocity coordinates are given in a heliocentric system, then the motion of the sun with respect to the barycentric system must be determined, as must the effect of the planets whose motion causes the heliocentric velocity of the Earth to differ from its barycentric velocity.

4.2.3 Parallax

Parallax is a displacement of the apparent object on the celestial sphere from its true position due to the shift in position of the observer. *Diurnal parallax* is due to the observer's change in position associated with Earth's rotation; *annual parallax* is due to the observer's change in position associated with Earth's orbital motion. For objects outside the solar system, the diurnal parallax can be neglected since the Earth's radius is much smaller than the distance even to the nearest stars. Therefore, we consider only the annual parallax. For quasars, which are the most distant objects in the universe, the parallax is zero.

Returning to Figure 4.11, the coordinates of E, denoted by the vector, $\left(x_B, y_B, z_B\right)^{\mathrm{T}}$, are given in the barycentric frame. The parallax angle, π, of a star is the maximum angle that the radius, ρ_E, of Earth's orbit (with respect to the barycenter) subtends at the star (usually, ρ_E is taken as the semi-major axis of Earth's elliptical orbit, or with sufficient accuracy, 1 AU). From the law of sines applied to the triangle, EBS, according to the figure:

$$\frac{\sin \Delta\theta}{\sin\theta} = \frac{\rho_E}{r_S} = \pi, \tag{4.87}$$

where r_S is the distance to the star. The effect of parallax, is therefore, approximately

$$\Delta\theta = \pi \sin\theta. \tag{4.88}$$

Clearly, this formula has a strong similarity to the aberration effect, equation (4.68); and, indeed, we can use the same Figure 4.13 as before, but now identify the point, F, with the direction from the observer to the barycenter of the celestial coordinate frame. Also, from Figure 4.11, the angle between F and S in Figure 4.13 is $\theta + \Delta\theta$ in the parallax case. But this is of no consequence since this angle enters only as an intermediate quantity in the derivations, not in the final result (moreover, equation (4.88) is approximate to first order in $\Delta\theta$); we will ignore this difference. The unit vector defining F is, therefore,

$$\boldsymbol{p} = \begin{pmatrix} -\dfrac{x_B}{\rho_E} \\ -\dfrac{y_B}{\rho_E} \\ -\dfrac{z_B}{\rho_E} \end{pmatrix} = \begin{pmatrix} \cos\delta_F \cos\alpha_F \\ \cos\delta_F \sin\alpha_F \\ \sin\delta_F \end{pmatrix}, \tag{4.89}$$

(note the negative signs in $\boldsymbol{p}$ are due to the geocentric view). From equations (4.70) and (4.88),

$$\Delta\alpha = \pi \sin\theta \cos\psi \sec\delta_S . \tag{4.90}$$

Substituting equations (4.72) and (4.89), we obtain the effect of annual parallax on right ascension:

$$\Delta\alpha = \pi\left(\frac{x_B}{\rho_E}\sin\alpha_S - \frac{y_B}{\rho_E}\cos\alpha_S \right)\sec\delta_S . \tag{4.91}$$

Similarly, from equations (4.71) and (4.88),

$$\Delta\delta = -\Delta\theta \sin\theta \sin\psi . \tag{4.92}$$

Using equation (4.79) with appropriate substitutions for the unit vector components, we find

$$\Delta\delta = \pi\left(\frac{x_B}{\rho_E}\cos\alpha_S \sin\delta_S + \frac{y_B}{\rho_E}\sin\alpha_S \sin\delta_S - \frac{z_B}{\rho_E}\cos\delta_S \right). \tag{4.93}$$

In using equations (4.91) and (4.93), we can approximate $\rho_E \approx 1\ \text{AU}$ and then the coordinate vector, $\left(x_B, y_B, z_B\right)^{\mathrm{T}}$, should have units of AU.

4.2.4 Refraction

As light (or any electromagnetic radiation) passes through the atmosphere, being a medium of non-zero mass density, its path deviates from a straight line due to the effect of *refraction*, thus causing the apparent direction of a visible object to depart from its true direction. We distinguish between *atmospheric refraction* that refers to light reflected from objects within the atmosphere, and *astronomic refraction* that refers to light coming from objects outside the atmosphere. Atmospheric refraction is important in terrestrial surveying applications, where targets within the atmosphere (e.g., on the ground) are sighted. We concern ourselves only with astronomic refraction of light. In either case, modeling the light path is difficult because refraction depends on the temperature, pressure, and water content (humidity) along the path.

For a spherically symmetric (i.e., spherically layered) atmosphere, Snell's law of refraction leads to (Smart, 1960, p.63):

$$nr \sin z = \text{constant} , \tag{4.94}$$

where n is the *index of refraction*, assumed to depend only on the radial distance, r, from Earth's center, and z is the angle, at any point, P, along the actual path, of the tangent to the light path with respect to r (Figure 4.16). It is assumed that the light ray originates at infinity, which is reasonable for all celestial objects in this application. With reference to Figure 4.16, z_S is the true *topocentric* zenith distance of the object, topocentric meaning that it refers to the location of the terrestrial observer. The topocentric apparent zenith distance is given by z_0; and, as the point, P, moves along the actual light path from the star to the observer, we have

$$0 \le z \le z_0 . \tag{4.95}$$

We define auxiliary angles, $\overline{z}_P$ and z_P, in Figure 4.16, and note that

$$\overline{z}_P = z_P + z . \tag{4.96}$$

The angle, $\overline{z}_P$, is the apparent zenith angle of the point, P, as it travels along the path. Therefore, the total *error* in the observed zenith angle due of refraction, defined by $\Delta z = z_0 - z_S$, is

$$\Delta z = \int_{z_s}^{z_0} d\overline{z}_P . \tag{4.97}$$

The error is generally negative, and the correction (being the negative of the error) is positive.

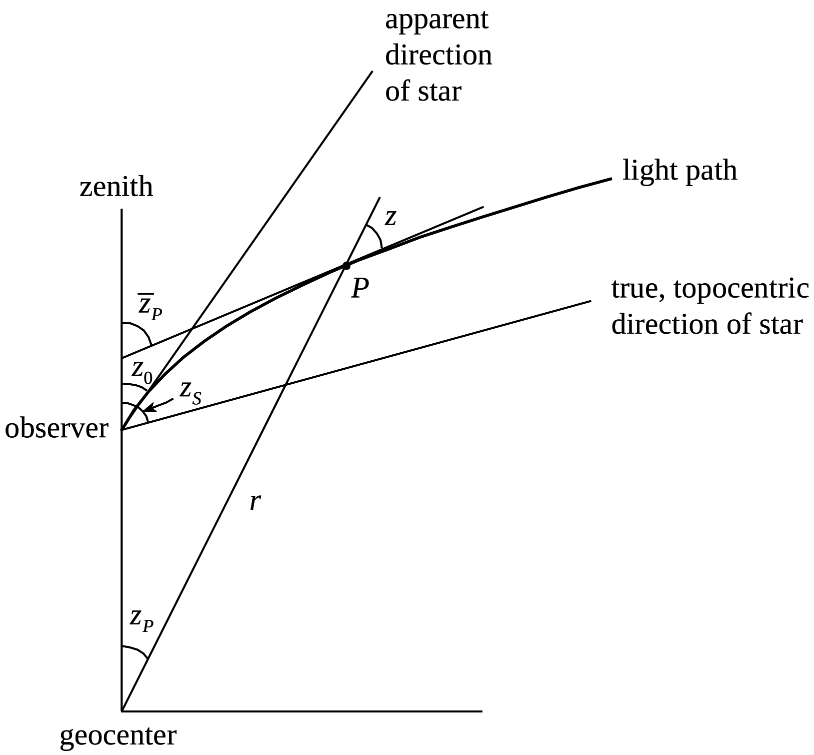


Figure 4.16: Geometry for astronomic refraction.

From equation (4.96), there is

$$d\bar{z}_P = dz_P + dz \ . \tag{4.98}$$

Taking differentials of equation (4.94), we have

$$d\left(nr\right)\sin z + nr\cos z\, dz = 0 \ , \tag{4.99}$$

which leads to

$$dz = -\tan z \frac{d\left(nr\right)}{nr} \ . \tag{4.100}$$

From Figure 4.17, which represents the differential displacement of the point, P , along the light path, we also have

$$\tan z = \frac{r dz_P}{dr} \quad \Rightarrow \quad dz_P = \frac{dr}{r} \tan z \, . \tag{4.101}$$

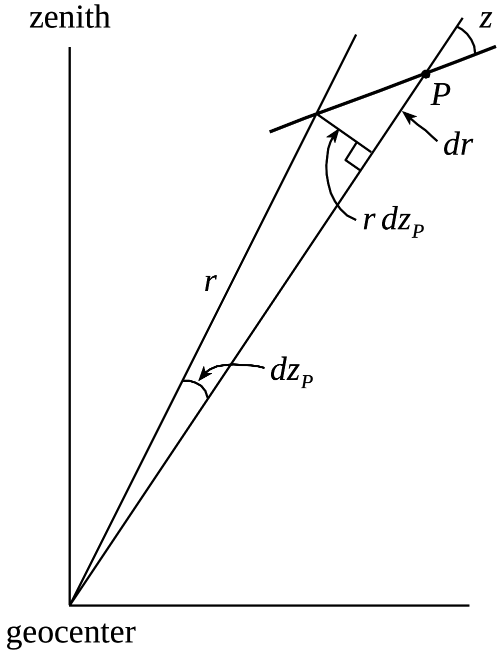


Figure 4.17: Differential change of *P* along light path.

Substituting equations (4.100) and (4.101) into equation (4.98), we find:

$$d\bar{z}_P = -\tan z \left(\frac{d\left(nr\right)}{nr} - \frac{dr}{r} \right). \tag{4.102}$$

This can be simplified using $d\left(nr\right) = rdn + ndr$, yielding

$$d\bar{z}_P = -\frac{dn}{n} \tan z \, . \tag{4.103}$$

Substituting equation (4.100) now gives

$$\begin{aligned} d\overline{z}_P &= \frac{dn}{n}\frac{nr}{d\left(nr\right)}dz \\ &= \frac{rdn}{ndr+rdn}dz \\ &= \frac{r\dfrac{dn}{dr}}{n+r\dfrac{dn}{dr}}dz \end{aligned} \tag{4.104}$$

Putting this change of integration variable from $\overline{z}_P$ to z into equation (4.97), we have

$$\Delta z = \int_0^{z_0} \frac{r\dfrac{dn}{dr}}{n+r\dfrac{dn}{dr}}dz\,, \tag{4.105}$$

where the limits of integration are obtained by noting that when $P \to \infty$ ($\overline{z}_P = z_s$), $z = 0$, and when *P* is at the observer, $z = z_0$. Again, note that equation (4.105) yields the refraction error; the correction is the negative of this.

To implement formula (4.105) requires a model for the index of refraction, and numerical methods to calculate it are indicated by Seidelmann (1992, p.141-143). The *errors* in the observed coordinates are obtained as follows. From equation (2.180), we have

$$\sin\delta_S = \cos A_S \cos\Phi \sin z_S + \sin\Phi \cos z_S\,, \tag{4.106}$$

where A_S is the azimuth of the star. Under the assumptions, $\Delta A_S = 0$ and $\Delta\Phi = 0$, this leads to

$$\Delta\delta = \frac{\Delta z}{\cos\delta_S}\left(\cos A_S \cos\Phi \cos z_S - \sin\Phi \sin z_S\right). \tag{4.107}$$

Again, from equation (2.180), it can be shown easily that

$$\tan t_S = \frac{\sin A_S}{\sin\Phi \cos A_S - \cos\Phi \cot z_S}\,. \tag{4.108}$$

With $\Delta A_S = 0$ and $\Delta\Phi = 0$, and noting that $\Delta t_S = -\Delta\alpha_S$, one readily can derive (left to the reader – use equation (2.180)!) that:

$$\Delta\alpha = -\frac{\sin t_S \cos \Phi}{\sin z_S \cos \delta_S} \Delta z \ . \tag{4.109}$$

4.2.5 Problems

1. Derive equations (4.94) and (4.109).

2. In VLBI (Very Long Baseline Interferometry), we analyze signals of a quasar (celestial object at an extremely large distance from the Earth) at two points on the Earth to determine the directions of the quasar at these two points, and thus to determine the *terrestrial* coordinate *differences*, $\Delta x, \Delta y, \Delta z$. The coordinates of the quasar are given in the ICRF. State which of the following effects would have to be considered for maximum accuracy in our coordinate determination in the ITRF (note that we are concerned only with coordinate *differences*):
precession, nutation, polar motion, proper motion, annual parallax, diurnal parallax, annual aberration, diurnal aberration, refraction. Justify your answer for *each* effect.

4.3 Relationship to the Terrestrial Frame

Previous Sections provided an understanding of the relationship between catalogued coordinates of celestial objects (i.e., in a celestial reference frame) and the coordinates as might be observed with respect to instantaneous celestial axes. It is now required to apply Earth rotation to obtain corresponding terrestrial coordinates. But the axes that define the terrestrial reference system differ from the axes described casually in Section 2.3. In fact, the spin axis and various other "natural" axes associated with Earth's rotation exhibit motion with respect to the Earth's crust due to the natural dynamics of the rotation; whereas, the axes of the terrestrial reference system are fixed to Earth's crust. Euler's equations describe the motion of the natural axes for a rigid body, but because the Earth is partially fluid and elastic, the motion of these axes is not accurately predictable. The reader is referred to Moritz and Mueller (1987) for theoretical and mathematical developments of the dynamics equations for rotating bodies; we restrict the discussion to a description of the effects on coordinates. However, a heuristic discussion of the different types of motion of the axes is also given here, leading ultimately to the definition of the *Celestial Intermediate Pole,* CIP (previously also called the *Celestial Ephemeris Pole,* CEP). The recent (turn of the century) changes in the fundamental conventions of the transformation between the celestial reference system and the CIP have also been extended to the transformation between the terrestrial reference system and the CIP; and these are described in Sections 4.3.1.1 and 4.3.2.1. The last sub-section then summarizes the entire transformation from celestial to terrestrial reference frames.

4.3.1 Polar Motion

The motion of an axis, like the instantaneous spin axis, of the Earth with respect to the body of the Earth is called *polar motion,* also *wobble* (Dehant and Mathews 2015). In terms of coordinates, the dynamic location of the axis is described as $\left(x_P, y_P\right)$ with respect to the terrestrial reference pole (IRP of the International Terrestrial Reference System). Figure 4.18 shows the polar motion coordinates for the CIP (see Section 4.3.2); they are functions of time (note the defined directions of x_P and y_P). Since they are small angles, they can be viewed as Cartesian coordinates near the reference pole, varying periodically around the pole with magnitude of the order of 6 m; but they are usually given as angles in units of arcsec.

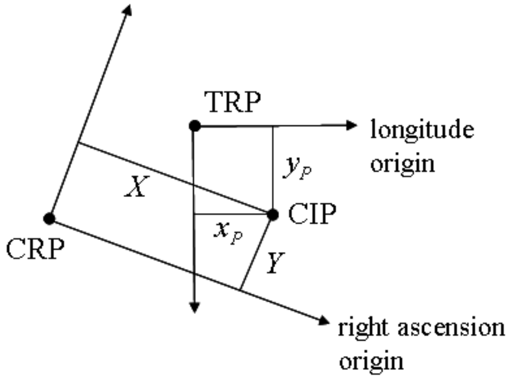


Figure 4.18: Polar motion coordinates, x_p, y_p, in relation to precession/nutation coordinates, X, Y. TRP = terrestrial reference pole; CRP = celestial reference pole, CIP = celestial intermediate pole.

The principal component of polar motion is the *Chandler wobble*. This is basically the free Eulerian motion which would have a period of about 304 days, based on the moments of inertia of the Earth, if the Earth were a rigid body. Due to the elastic yielding of the Earth, resulting in displacements of the maximum moment of inertia, this motion has a longer period of about 430 days. Around 1890 S.C. Chandler analyzed astronomic latitudes and longitudes (that are tied to Earth's spin axis after applying precession and nutation) and discovered this actual period. Shortly afterward Newcomb gave the dynamical explanation for the discrepancy relative to the Eulerian prediction (Mueller, 1969, p.80). The period of the main component of polar motion is called the *Chandler* period; its amplitude is about 0.2 arcsec. Other components of polar motion include the approximately annual signal due to the redistribution of masses by way of meteorological and geophysical processes, with amplitude of about $0.05-0.1\text{ arcsec}$, and the *nearly diurnal free wobble*, due to the slight misalignments of the rotation axes of the mantle and liquid outer core, with an amplitude of the order of μas in the Earth-fixed frame (also known as *free core nutation*, with magnitude of about 0.1 mas, when referred to the inertial frame). Finally, there is the so-called *polar wander*, which is the secular motion of the pole. During 1900 – 2000, Earth's spin axis wandered about 0.004 arcsec per year in the direction of the $80°\text{ W}$ meridian. Figure 4.19 shows the Chandler motion of the pole for the period 2000 to 2010, and also the general drift for the last 110 years.

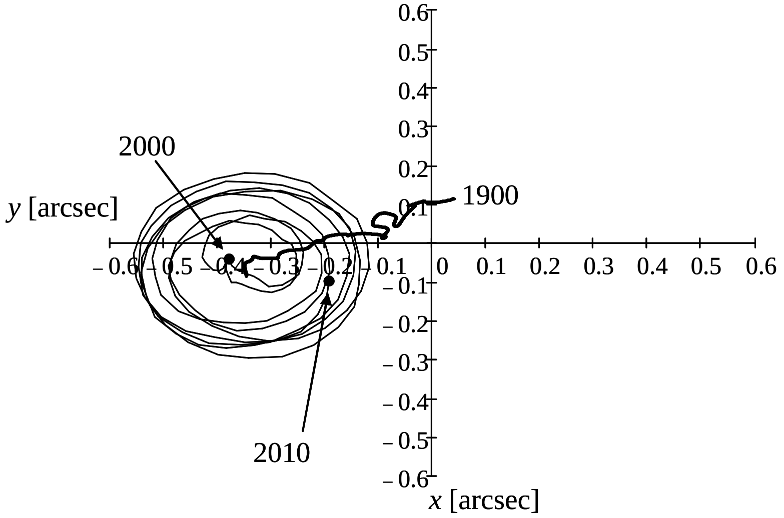


Figure 4.19: Polar motion from 2000 to 2010, and polar wander since 1900. Polar motion coordinates were obtained from IERS[3] and smoothed to obtain the trend.

The transformation of astronomic terrestrial coordinates and azimuth from the instantaneous pole (the CIP) to the terrestrial reference pole fixed on the Earth's crust (the IRP) is constructed with the aid of Figures 4.20 and 4.21. Let $\mathit{\Phi}_t$, $\mathit{\Lambda}_t$, A_t denote the apparent (observed) astronomic latitude, longitude, and azimuth at epoch, t, with respect to the CIP; and let $\mathit{\Phi}$, $\mathit{\Lambda}$, A denote the corresponding angles with respect to the terrestrial pole, such that

$$\begin{aligned}\mathit{\Delta\Phi} &= \mathit{\Phi} - \mathit{\Phi}_t \\ \mathit{\Delta\Lambda} &= \mathit{\Lambda} - \mathit{\Lambda}_t \\ \mathit{\Delta} A &= A - A_t\end{aligned} \tag{4.110}$$

represent the *corrections* to the apparent angles. In linear approximation, these corrections are the small angles shown in Figures 4.20 and 4.21.

[3] https://www.iers.org/IERS/EN/DataProducts/EarthOrientationData/eop.html

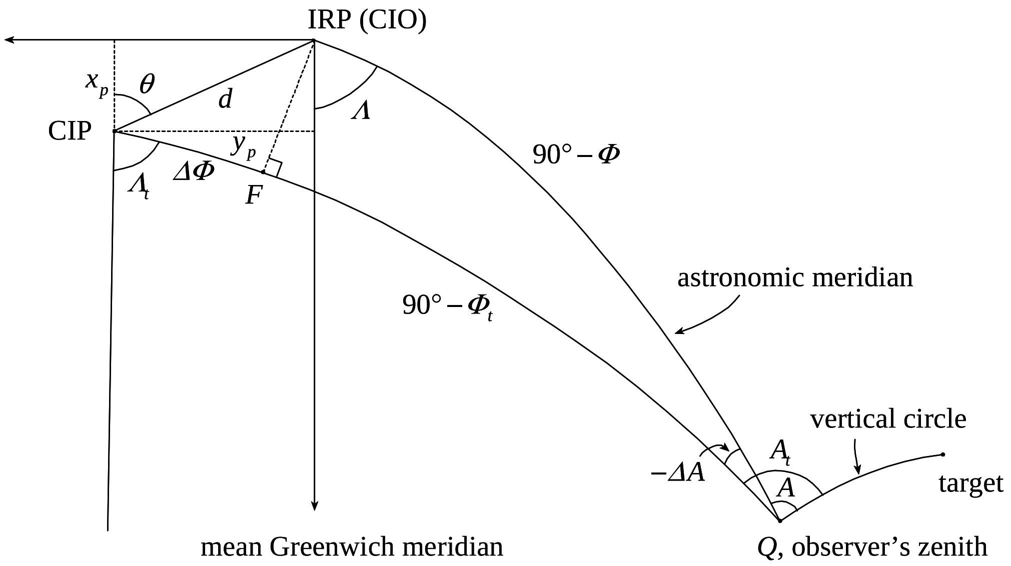


Figure 4.20: Relationship between apparent astronomical coordinates at current epoch, t, and corresponding coordinates with respect to the terrestrial reference frame.

We introduce the polar coordinates, d and θ, so that:

$$\begin{aligned} x_P &= d\cos\theta \\ y_P &= d\sin\theta \end{aligned} \tag{4.111}$$

Then, for the latitude, we have from the triangle, $CIP - IRP - F$:

$$\begin{aligned} \Delta\Phi &= d\cos\left(180° - \Lambda_t - \theta\right) \\ &= -d\cos\Lambda_t\cos\theta + d\sin\Lambda_t\sin\theta \\ &= y_P\sin\Lambda_t - x_P\cos\Lambda_t \end{aligned} \tag{4.112}$$

For the azimuth, using the law of sines on the spherical triangle, $CIP - IRP - Q$, we have:

$$\frac{\sin\left(-\Delta A\right)}{\sin d} = \frac{\sin\left(180° - \Lambda_t - \theta\right)}{\sin\left(90° - \Phi\right)}. \tag{4.113}$$

With the usual small angle approximations, this leads to

$$\begin{aligned} \Delta A &= -\frac{d}{\cos\Phi}\left(\sin\Lambda_t\cos\theta + \cos\Lambda_t\sin\theta\right) \\ &= -\left(x_P\sin\Lambda_t + y_P\cos\Lambda_t\right)\sec\Phi \end{aligned} \tag{4.114}$$

Finally, for the longitude we again apply the law of sines to the triangle, QRM , Figure 4.21, to obtain:

$$\frac{\sin\left(-\mathit{\Delta} A\right)}{\sin\left(-\mathit{\Delta}\mathit{\Lambda}\right)}=\frac{\sin 90^\circ}{\sin\mathit{\Phi}_t}. \tag{4.115}$$

From this and with equation (4.114), we have

$$\begin{aligned}\mathit{\Delta\Lambda}&=\sin\mathit{\Phi}_t\,\mathit{\Delta}A\\&=-\left(x_P\sin\mathit{\Lambda}_t+y_P\cos\mathit{\Lambda}_t\right)\tan\mathit{\Phi}\end{aligned} \tag{4.116}$$

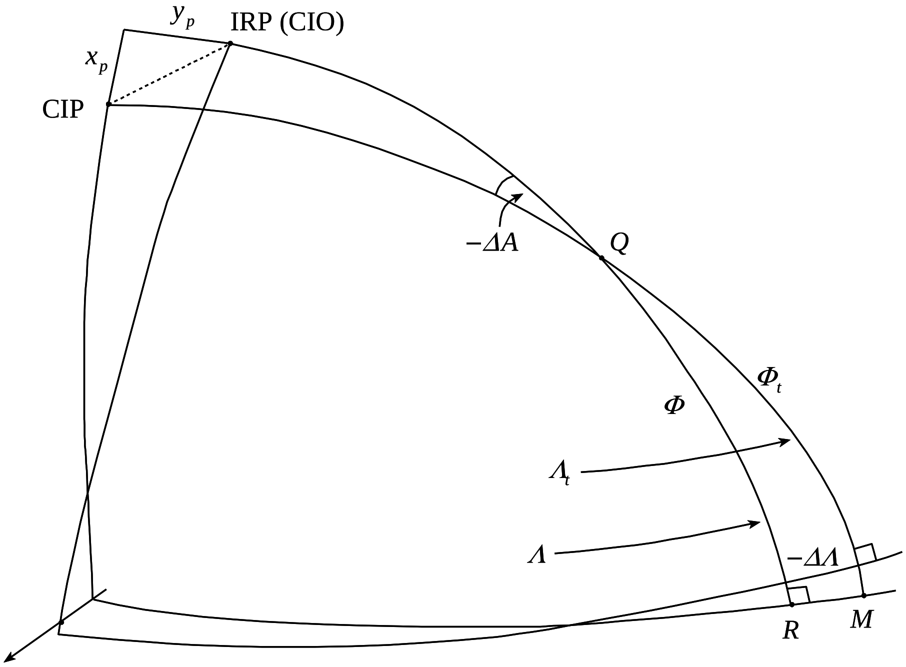


Figure 4.21: Relationship between the apparent longitude with respect to the CIP and the longitude with respect to IRP.

Relationships (4.112) and (4.116) can also be derived from

$$\begin{pmatrix}\cos\mathit{\Phi}_t\cos\mathit{\Lambda}_t\\ \cos\mathit{\Phi}_t\sin\mathit{\Lambda}_t\\ \sin\mathit{\Phi}_t\end{pmatrix}=\mathbf{R}_1\left(y_P\right)\mathbf{R}_2\left(x_P\right)\begin{pmatrix}\cos\mathit{\Phi}\cos\mathit{\Lambda}\\ \cos\mathit{\Phi}\sin\mathit{\Lambda}\\ \sin\mathit{\Phi}\end{pmatrix}, \tag{4.117}$$

where the vectors on either side represent unit vectors in the direction of the tangent to the local plumb line, but in different coordinate systems; and the rotation matrices are given by equations (1.4) and (1.5). Indeed, the last element of vector equation (4.117) easily gives $\Delta\Phi$ in equation (4.112) (neglecting second-order terms); while multiplying the first by $\sin\Lambda$ and the second by $\cos\Lambda$ and subtracting yields equation (4.116) (the student should fill in the details). The combined rotation matrix, in equation (4.117), for polar motion is also denoted by $\mathbf{W}$, representing the transformation from the terrestrial reference pole to the celestial intermediate pole:

$$\mathbf{W} = \mathbf{R}_1\left(y_P\right)\mathbf{R}_2\left(x_P\right). \tag{4.118}$$

The polar motion coordinates are tabulated by the IERS as part of the Earth Orientation Parameters (EOP) on the basis of observations, such as from VLBI and satellite ranging. Thus, $\mathbf{W}$ is a function of time, but there are no analytic formulas for polar motion as there are for precession and nutation.

4.3.1.1 New Conventions

As described in Section 4.1.3, the celestial coordinate system associated with the instantaneous pole (the CIP) possesses a newly defined origin point for right ascensions: a non-rotating origin (NRO), σ, called the *Celestial Intermediate Origin*, CIO (previously also called the *Celestial Ephemeris Origin*, CEO; and not to be confused with the conventional international origin – the pre-1980s name for the reference pole). The instantaneous pole can also be associated with an instantaneous terrestrial coordinate system, where likewise, according to resolutions adopted by the IAU (and IERS), the origin of longitudes is a non-rotating origin, called the *Terrestrial Intermediate Origin*, TIO (previously also called the *Terrestrial Ephemeris Origin*, TEO). It should be noted that neither the CIO nor the TIO represents an origin for coordinates of points in a *reference* system. They are origin points associated with an instantaneous coordinate system, moving with respect to the celestial sphere (the CIO) or with respect to the Earth’s crust (TIO), whence their previous designation, “ephemeris,” and now simply “intermediate”.

With this new definition of the instantaneous terrestrial coordinate system, the polar motion transformation, completely analogous to the precession-nutation matrix, $\mathbf{Q}^{\mathrm{T}}$, equation (4.40), is now given as

$$\mathbf{W} = \mathbf{R}_3\left(-s'\right)\mathbf{R}_3\left(-F\right)\mathbf{R}_2\left(g\right)\mathbf{R}_3\left(F\right), \tag{4.119}$$

where the instantaneous pole (CIP) has coordinates, $\left(g, F\right)$, in the terrestrial reference system. As shown in Figure 4.22, g is the co-latitude (with respect to the instantaneous equator) and F is the longitude (with respect to the TIO, ω); and we may write:

$$\begin{pmatrix} x_P \\ y_P \\ z_P \end{pmatrix} = \begin{pmatrix} \sin g \cos F \\ -\sin g \sin F \\ \cos g \end{pmatrix}, \tag{4.120}$$

where the adopted polar motion coordinates, x_P, y_P, are defined as before (Figure 4.20), with y_P along the 270° meridian.

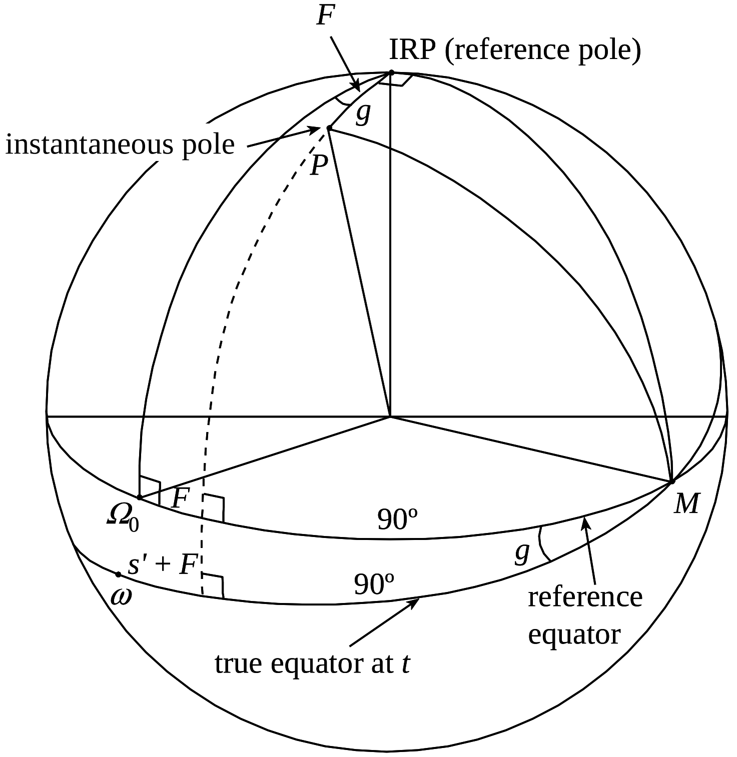


Figure 4.22: Coordinates of instantaneous pole in the terrestrial reference system.

With a completely analogous derivation as for the precession-nutation matrix, $\mathbf{Q}$, we find that

$$\mathbf{W} = \mathbf{R}_3\left(-s'\right)\begin{pmatrix} 1-a'x_P^2 & a'x_P y_P & -x_P \\ a'x_P y_P & 1-a'y_P^2 & y_P \\ x_P & -y_P & 1-a'\left(x_P^2+y_P^2\right) \end{pmatrix}, \tag{4.121}$$

where $a' = 1\big/\left(1+\cos g\right) = 1/2 + \left(x_P^2+y_P^2\right)\big/8$. Also, the parameter, s', defining the location of the TIO as a non-rotating origin on the instantaneous equator, is given (analogous to equation (4.46)) by

$$s' = s_0' + \int\limits_{t_0}^{t} \frac{x_P \dot{y}_P - y_P \dot{x}_P}{1+z_P}\,dt\,, \tag{4.122}$$

again, noting that y_P is positive along the $270°$ meridian. The constant, s_0', may be chosen to be zero (i.e., s' is zero at $t=t_0$).

It is easy to show that by neglecting terms of third and higher orders, the exact expression (4.121) is approximately equal to

$$\mathbf{W} = \mathbf{R}_3\left(-s'\right)\mathbf{R}_3\left(\frac{1}{2}x_P y_P\right)\mathbf{R}_1\left(y_P\right)\mathbf{R}_2\left(x_P\right). \tag{4.123}$$

Furthermore, s' is significant only because of the largest components of polar motion and an approximate model is given by (McCarthy and Petit 2003)

$$s' = -0.0015\left(\frac{a_c^2}{1.2}+a_a^2\right)\tau\ \text{[arcsec]}, \tag{4.124}$$

where a_c and a_a are the amplitudes, in arcsec, of the Chandler wobble $O\left(0.2\text{ arcsec}\right)$ and the annual wobble $O\left(0.05\text{ arcsec}\right)$. Hence, the magnitude of s' is of the order of 0.1 mas. The IERS Conventions 2003 and 2010 also neglect the second-order terms (being of order $0.2\ \mu\text{as}$) in equation (4.123) and give:

$$\mathbf{W} = \mathbf{R}_3\left(-s'\right)\mathbf{R}_1\left(y_P\right)\mathbf{R}_2\left(x_P\right), \tag{4.125}$$

which is the traditional transformation due to polar motion, equation (4.118), with the additional small rotation that exactly realizes the instantaneous zero meridian of the instantaneous pole and equator.

The polar motion coordinates should now also contain short-period terms in agreement with the new definition of the intermediate pole. Thus, according to the IERS Conventions 2010 (Petit and Luzum 2010), which describes these in detail,

$$\left(x_P, y_P\right) = \left(x, y\right)_{\text{IERS}} + \left(\mathit{\Delta} x, \mathit{\Delta} y\right)_{\text{tides}} + \left(\mathit{\Delta} x, \mathit{\Delta} y\right)_{\text{libration}}, \tag{4.126}$$

where $\left(x, y\right)_{\text{IERS}}$ are the polar motion coordinates published by the IERS, $\left(\mathit{\Delta} x, \mathit{\Delta} y\right)_{\text{tides}}$ are modeled tidal components in polar motion derived from tide models (mostly diurnal and sub-diurnal variations), and $\left(\mathit{\Delta} x, \mathit{\Delta} y\right)_{\text{libration}}$ are long-period polar motion effects corresponding to short-period (less than 2 days) nutations. The latter should be added according to the new definition of the intermediate pole that should contain no nutations with periods shorter than 2 days.

4.3.1.2 Problems

1. Derive equations (4.121) and (4.123).

2. a) From the web site:
https://www.iers.org/IERS/EN/DataProducts/EarthOrientationData/eop.html
extract the polar motion coordinates (Earth orientation parameters (EOP)) from 1846 to 2010 at 0.05 year (0.1 year) intervals.

b) Plot the polar motion for the intervals 1900.0 - 1905.95 and 2000.0 – 2005.95. Determine the period of the motion for each interval. Describe the method you used to determine the period (graphical, Fourier transform, least-squares, etc.).

c) Using the period determined (use an average of the two) in b) divide the whole series from 1846 to 2010 into intervals of one period each. For each such interval determine the average position of the CIP. Plot these mean positions and verify the polar wander of 0.004 arcsec per year in the direction of –80° longitude.

3.(advanced) From the data obtained in 1a) determine the Fourier spectrum in each coordinate and identify the Chandler and annual components (to use a Fourier transform algorithm, such as FFT, interpolate the data to a resolution of 0.05 year, where necessary). For each polar motion coordinate, plot these components separately in the time domain, as well as the residual of the motion (i.e., the difference between the actual motion and the Chandler plus annual components). Discuss your results in terms of relative magnitudes. What beat-frequency is recognizable in a plot of the total motion in the time domain?

4.3.2 Celestial Ephemeris Pole

This section describes the previously defined Celestial Ephemeris Pole (CEP), as the precursor to the newly defined Celestial Intermediate Pole (CIP). Both are the same at a certain level of precision, where the CIP is a refinement on the CEP owing to the increased resolution afforded by new VLBI observations. In order to understand how the CEP was chosen as the defining axis for which nutation (and precession and polar motion) are computed, it is necessary to consider briefly the dynamics and kinematics of Earth rotation. The development here refers to the theory given in much greater detail by Moritz and Mueller (1987). We consider the following axes for the Earth:

1. *Instantaneous rotation axis,* R. It is the direction of the instantaneous rotation vector, $\boldsymbol{\omega}_e$.

2. *Figure axis,* F. It is the *principal axis of inertia* that corresponds to the *moment of inertia* with the maximum value. These terms are explained as follows. Every body has an associated inertia tensor, $\mathbf{I}$, which is the analogue of (inertial) mass. (A *tensor* is a generalization of a vector, in our case, to second order; that is, a vector is really a first-order tensor.) The tensor may be represented as a 3×3 matrix of elements, I_{jk}, that are related to the second-order moments of the mass distribution of a body with respect to the coordinate axes. Specifically, the *moments of inertia,* I_{jj}, occupy the diagonal of the matrix and are given by

$$I_{jj} = \int\limits_{\text{mass}} \left(r^2 - x_j^2\right) dm, \quad j = 1, 2, 3, \tag{4.127}$$

where $r^2 = x_1^2 + x_2^2 + x_3^2$; and the *products of inertia,* I_{jk}, are the off-diagonal elements expressed as

$$I_{jk} = -\int\limits_{\text{mass}} x_j x_k \, dm, \quad j \neq k . \tag{4.128}$$

Thus, the inertia tensor is given by

$$\mathbf{I} = \begin{pmatrix} I_{11} & I_{12} & I_{13} \\ I_{21} & I_{22} & I_{23} \\ I_{31} & I_{32} & I_{33} \end{pmatrix}. \tag{4.129}$$

The products of inertia vanish if the coordinate axes coincide with the *principal axes of inertia* for the body. This happens with a suitable rotation of the coordinate system (with origin assumed to be at the center of mass) that *diagonalizes* the inertia tensor (this can always be assumed possible). Heuristically, these principal axes represent the axes of symmetry in the mass distribution of the body.

3. *Angular momentum axis*, H. It is defined by the direction of the angular momentum vector, $\boldsymbol{H}$, as a result of rotation. We have, by definition,

$$\boldsymbol{H} = \mathbf{I}\boldsymbol{\omega}_e . \tag{4.130}$$

This shows that the angular momentum vector, $\boldsymbol{H}$, and the angular velocity vector, $\boldsymbol{\omega}_e$, generally are not parallel. Equation (4.130) is the analogue to linear momentum, $\boldsymbol{p}$, being proportional (hence always parallel) to linear velocity, $\boldsymbol{v}$ ($\boldsymbol{p} = m\boldsymbol{v}$, where m is the total mass of the body).

For rigid bodies, *Euler's equation* describes the dynamics of the angular momentum vector in a *body-fixed frame* (coordinate axes fixed to the body):

$$\boldsymbol{L}^b = \dot{\boldsymbol{H}}^b + \boldsymbol{\omega}_e \times \boldsymbol{H}^b , \tag{4.131}$$

where $\boldsymbol{L}^b$ is the vector of external torques applied to the body (in our case, e.g., luni-solar gravitational attraction acting on the Earth). The superscript, b, in equation (4.131) designates that the coordinates of each vector are in a body-fixed frame. In the inertial frame (which does not rotate), equation (4.131) specializes to

$$\boldsymbol{L}^i = \dot{\boldsymbol{H}}^i . \tag{4.132}$$

Again, the superscript, i, designates that the coordinates of the vector are in the inertial frame. If $\boldsymbol{L}^i = \mathbf{0}$, then no torques are applied, and this expresses the *law of conservation of angular momentum*: the angular momentum of a body is constant in the absence of applied torques. That is, $\dot{\boldsymbol{H}}^i = \mathbf{0}$ clearly implies that $\boldsymbol{H}$ remains fixed in inertial space.

In general, equation (4.131) is a differential equation for $\boldsymbol{H}^b$ with respect to time. Its solution shows that both $\boldsymbol{H}^b$ and $\boldsymbol{\omega}_e$ (through equation (4.130)) exhibit motion with respect to the body, even if $\boldsymbol{L}^b = \mathbf{0}$. This is *polar motion*. Also, if $\boldsymbol{L}^b \neq \mathbf{0}$, then $\boldsymbol{H}^b$ changes direction with respect to an inertial frame. Indeed, in the presence of external torques, all axes change with respect to the inertial frame – we have already studied this as *precession and nutation*. Comprehensively, we define the following:

<u>Polar Motion</u>: the motion of the Earth's axis (R, F, or H) with respect to the body of the Earth.

Nutation: the motion of the Earth's axis (R , F , or H) with respect to the inertial frame.

Both polar motion and nutation can be viewed as either motion in the absence of torques (*free* motion) or motion in the presence of torques (*forced* motion). Thus, there are four possible types of motion for each of the three axes. However, for one axis we can rule out one type of motion. For a rotating body not influenced by external torques ($\boldsymbol{L} = \boldsymbol{0}$), the angular momentum axis, H , has no nutation (as shown above, it maintains a constant direction in the inertial frame). Therefore, H *has no free nutation*. On the other hand, the direction of the angular momentum axis in space is influenced by external torques, and so H exhibits forced nutations.

We thus have the following types of motion:

i) forced polar motion of R , F , or H ;
ii) free polar motion of R , F , or H ;
iii) forced nutation of R , F , or H ;
iv) free nutation of R or F .

We also note that for a rigid body, F has no polar motion (free or forced) since it is an axis defined by the mass distribution of the body, and therefore, fixed within the body. On the other hand, the Earth is not a rigid body, which implies that F is not fixed to the crust of the Earth – it follows the principal axis of symmetry of the mass distribution as the latter changes in time (e.g., due to tidal forces). In summary, the consideration of nutation and polar motion involves:

a) three axes; R , F , and H (and one more fixed to the Earth, the IRP; we call it O);
b) rigid and non-rigid Earth models;
c) free and forced motions.

From a study of the mechanics of body motion applied to the Earth, it can be shown that (for an elastic Earth model; see Figure 4.23):

a) the axes R_0 , F_0 , and H_0 , corresponding to *free polar motion*, all lie in the same plane; similarly the axes, R , F , and H , corresponding to the (actual) forced motion also must lie in one plane;
b) forced polar motion exhibits nearly diurnal (24-hr period) motion, with amplitudes of $\sim 60\text{ cm}$ for R , $\sim 40\text{ cm}$ for H , and ~ 60 *meters* for F ;
c) free nutation exhibits primarily nearly diurnal motion.

On the other hand (again, see Figure 4.23):

d) free polar motion is mostly long-periodic (Chandler period, $\sim 430\text{ days}$), with dominant amplitudes of $\sim 6\text{ m}$ for R_0 and H_0 , and $\sim 2\text{ m}$ for F_0 ;
e) forced nutation is mostly long-periodic (18.6 yr , semi-annual, semi-monthly, etc.).

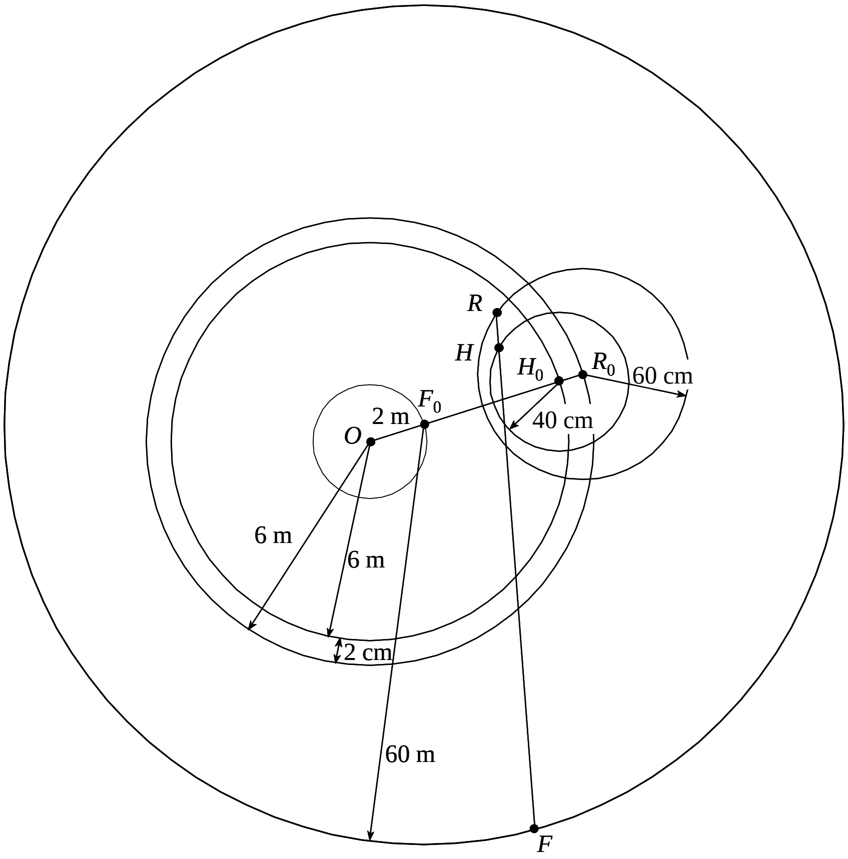


Figure 4.23: Free (zero-subscripted) and forced polar motions of axes for an elastic Earth. (Not to scale; indicated amplitudes are approximate.)

Free motion (polar motion and nutation) cannot be modeled by simple dynamics, and can only be determined empirically on the basis of observations. It is rather irregular. *Forced motion*, being due to torques from well known external sources, can be predicted quite accurately from luni-solar (and planetary) ephemerides.

If the Earth were a *rigid* body, then the F -axis would be fixed to the Earth ($F = F_0 = O$ in this case) and could serve as the reference for polar motion of the H - and R -axes. However, for a non-rigid Earth, in particular, for an elastic Earth, the F -axis deviates substantially from a fixed point on the Earth with a *daily* polar motion of amplitude $\sim 60\text{ m}$. Thus, F cannot serve as reference axis either for polar motion or for nutation.

In Figure 4.23, the point O is a fixed point on the Earth's surface, representing the mean polar motion (for the elastic Earth), and formally is called the *mean Tisserand figure axis*. It can be shown that free polar motion affects the nutations of the O - and R -axes, while the nutation of the H -axis is unaffected by free polar motion. This is because the motion of the angular momentum axis in the inertial frame is determined dynamically from the luni-solar torques (equation (4.132)) and not by the internal constitution of the Earth. This makes H a good candidate for the reference axis for nutations, since its (forced) nutation is unaffected by difficult-to-model free polar motion, and it has no free nutation.

However, it still has forced polar motion (diurnal and erratic). Therefore, the IAU in 1979 adopted H_0 as the CEP (the celestial ephemeris pole), since H_0 has no *forced* polar motion (by definition); and it, like H, has no free nutation. Thus H_0 has no nearly diurnal motions according to b) and c) above – it is rather stable with respect to the Earth and space. Note that H_0 still has free polar motion and forced nutation. On the other hand, as mentioned above, the (forced) nutation of H_0 does not depend on free polar motion. And since the O-axis (being fixed to the Earth's crust) also has no polar motion (i.e., by definition), its forced nutation, like that of H_0, does not depend on free polar motion. Therefore, both the O-axis and the H_0-axis have the same forced nutations. All these properties of H_0 make it the most suitable candidate for the CEP.

However, In particular, higher frequency components of polar motion (periods shorter than a two days) could be observed with higher resolution VLBI and nutation models expanded to included improved models of the Earth's interior. The definition of the CEP evolved from a physical quantity or model such as illustrated above to one defined strictly on the basis of frequency content, as elaborated in the sect section.

4.3.2.1 Celestial Intermediate Pole

The purpose of the CEP was to serve as the intermediate pole in the transformation between the celestial and terrestrial reference systems. That is, the motion of the CEP relative to the terrestrial reference pole is described by polar motion, while precession and nutation refer to its motion relative to the celestial reference pole. As such, the realization of the CEP depends on the models developed for precession and nutation and it also depends on observations of polar motion. However, modern observation techniques, such as VLBI (for an introduction to VLBI, see Seeber 2003), are now able to determine motions of the instantaneous pole with temporal resolution as high as a few hours. Also, the modern theories of nutation and polar motion now include diurnal and shorter-period motions (particularly the variations due to tidal components). Therefore, the conceptual definition of the CEP, being limited in frequency content, proved to be inadequate (Capitaine 2002). These developments made it necessary to revise the intermediate pole. Rather than defining the intermediate pole in terms of some particular physical model, such as the angular momentum axis, it is defined precisely in terms of realized frequency components of motion.

The new intermediate pole is called, to further emphasize its specific function, the *Celestial Intermediate Pole* (CIP). It separates the transformation between the terrestrial reference system and the celestial reference system into two parts (precession/nutation and polar motion) according to frequency content. With a resolution adopted by the IAU, the motions of precession and nutation of the CIP with respect to the celestial sphere are defined to have only periods greater than 2 days (frequencies within the band, $\left(-0.5\text{ cpsd}, 0.5\text{ cpsd}\right)$, where cpsd = cycles per sidereal day, and where a sidereal day corresponds to one rotation of the Earth relative to the celestial sphere; see Section 5.1); see Figure 4.24. These are the motions produced mainly by external torques on the Earth. Also included are polar motions in the so-called *retrograde diurnal band* (negative frequencies in the band, $\left(-1.5\text{ cpsd}, -0.5\text{ cpsd}\right)$), since it can be shown that they are equivalent to nutations with periods larger than 2 days. The terrestrial motions of the CIP, on the other hand, are defined to be those with frequencies outside the retrograde diurnal band. Note that the retrograde diurnal band is shifted from the celestial-motion frequencies by the effect of Earth rotation (1 cpsd). Thus, polar motion includes both long-period and short-period motions (outside this diurnal band) in the terrestrial system, as well as all short-period motions in the celestial system, as illustrated in Figure 4.24. In that sense, the CIP is merely an extension of the CEP in allowing higher frequency nutation components to be included, but as equivalent polar motions. These higher-frequency motions have minimal impact for most users, having at most a few tens of micro-arcsec in amplitude (for the nutations) and up to a few hundred micro-arcsec for tidally induced diurnal and semi-diurnal polar motions. The reader is referred to the IERS Conventions 2010 (Petit and Luzum 2010) and the IERS Technical Note 29 (Capitaine 2002) for further summaries, details, and references.

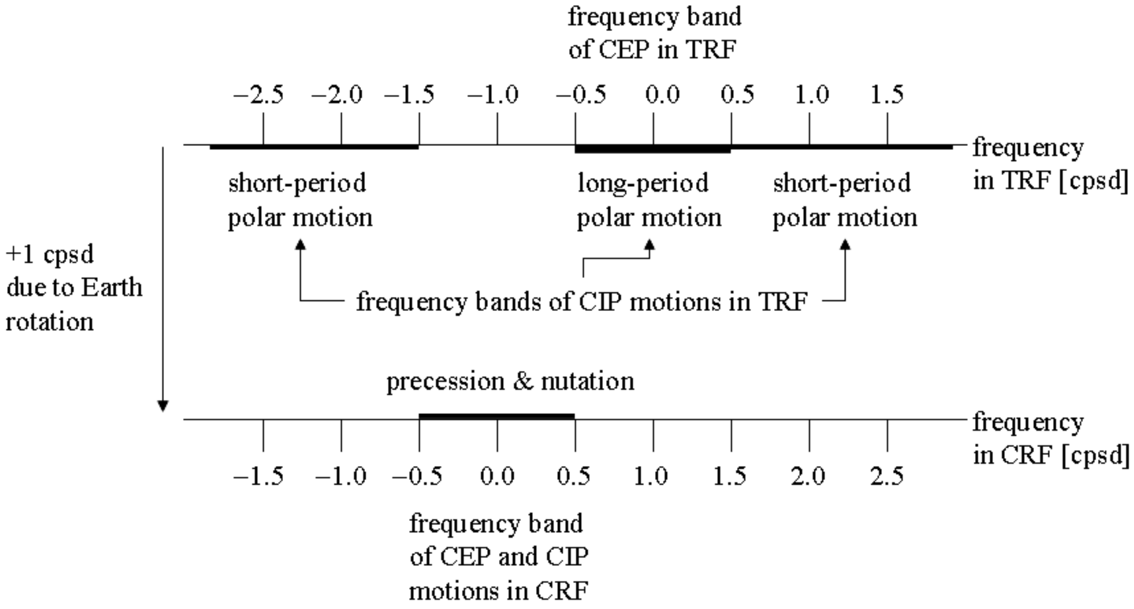


Figure 4.24: Spectral domains of the CEP and CIP motions in the terrestrial and celestial reference frames (after Dehant and Mathews 2015, p.87).

4.3.3 Transformations

We are interested in transforming the coordinates of a celestial object as given in a Celestial Reference Frame to the terrestrial reference system. The transformation, of course, is reversible; but this direction of the transformation is most applicable in geodesy, since we want to use the given coordinates of celestial objects in our observation models (e.g., to determine the coordinates for terrestrial stations). The given celestial frame coordinates are mean coordinates referring to some fundamental epoch and the coordinate transformations account for precession up to the epoch of date, nutation at the epoch of date, Earth rotation, and polar motion. In addition, various systematic effects due to proper motion of the object, aberration, parallax, and refraction must be applied as needed. The new conventions are similar with slight variations that account for differences in the origin for right ascension and formulation of the precession/nutation model. The transformation is formulated in terms of an algorithm for geocentric observers either "equinox-based", referring to the traditional methods, or "CIO-based", referring to the new conventions.

4.3.3.1 Apparent Place Algorithm

The object of this procedure is to formulate a transformation to compute the apparent geocentric coordinates of a star, given its mean position as listed in a catalogue. Apparent coordinates are those that would be observed in a geocentric, intermediate (instantaneous) celestial frame with annual aberration and parallax effects removed. Additional corrections for diurnal aberration (and diurnal parallax for objects in the solar system) are applied to obtain coordinates in a topocentric, intermediate celestial frame (Section 4.3.3.2). Applying polar motion and Earth rotation then brings the coordinates into the terrestrial reference frame. Finally, refraction needs to be considered when modeling observed coordinates.

The Apparent Place Algorithm follows the procedure described in the Astronomical Almanac[4]. The coordinates of a star are given in some catalogue that is a realization of the Celestial Reference System and includes also information on the velocity of the star (among other parameters). The coordinates and velocity, using the notation of the Hipparcos Catalogue, are valid at an epoch, t_0^{H}:

i) $\alpha_0,\ \delta_0,\ \pi$: catalogue celestial coordinates and parallax angle of the star;

ii) $\dot{\alpha} = \mu_{\alpha*} / \cos\delta_0$, $\dot{\delta}_0 = \mu_\delta$, $\dot{r}_0 = v$: velocities of proper motion. (4.133)

[4] The Astronomical Almanac, issued annually by the Nautical Almanac Office of the U.S. Naval Observatory, Washington, D.C.

It is noted that t_0^{H} for the Hipparcos catalogue is $t_0^H = 1991.25$, but the coordinates are in the ICRS (the Hipparcos catalogue in most cases does not give radial velocity, as it is considered of no consequence[5]. If it is available, e.g., from Doppler shift measurements, then it should be used). The algorithm proceeds by first determining the geocentric coordinates of the star at the epoch of the observation, t, still referred to the catalogue system. Usually, we have some time system in which we operate, e.g., Universal Time (Chapter 5). The star catalogues and celestial reference systems are established with respect to Barycentric Dynamic Time (TDB). Technically, one should distinguish between Terrestrial (Dynamic) Time (TT) and TDB, but practically the difference is less than 2 ms and can be ignored. We will define the relationship between TT and Universal Time, and among other time scales in Chapter 5. For now, assume that the time of observation, t, is in the scale of dynamic time, TT, in terms of Julian day numbers, e.g., $t = 2455984.5\text{ JD}$, which corresponds to 0^{hr} (midnight, civil time in UT) at Greenwich at the start of 27 February 2012. The fundamental epoch, t_0, for the precession and nutation series is, as usual, J2000.0, or $t_0 = 2451545.0\text{ JD}$. The Julian day number for t can be obtained from the Julian calendar (Astronomical Almanac, Section K); then we compute the appropriate fractions of a Julian century using

$$\tau = \frac{t - t_0}{36525} = \frac{t - 2451545.0}{36525}, \quad \tau^H = \frac{t - t_0}{36525} - \frac{t_0^H - 2000.0}{100}. \tag{4.134}$$

To continue with the determination of geocentric coordinates of the star at the time of observation, we require the location and velocity of the Earth at the time of observation in the barycentric system of reference. We may also need the barycentric coordinates of the sun for light-deflection corrections. The Jet Propulsion Laboratory publishes the standard ephemerides for bodies of the solar system, called DE405[6]. The Astronomical Almanac, Section B, lists some of these coordinates, as well, specifically, the vectors:

$\boldsymbol{E}_B(t)$: barycentric coordinates of Earth at time, t, in the ICRS.

$\dot{\boldsymbol{E}}_B(t)$: barycentric velocity of Earth at time, t, in the ICRS.

We need only 3 and 5 digits of accuracy, respectively, to obtain milliarcsec accuracy in the star's coordinates. The barycentric coordinates of the sun in the ICRS, $\boldsymbol{S}_B(t)$, are provided by DE405, and the heliocentric coordinates of the Earth are then

$$\boldsymbol{E}_H(t) = \boldsymbol{E}_B(t) - \boldsymbol{S}_B(t); \tag{4.135}$$

[5] https://en.wikipedia.org/wiki/Hipparcos

[6] http://ssd.jpl.nasa.gov/?ephemerides#planets

Both $\boldsymbol{S}_B(t)$ and $\boldsymbol{E}_H(t)$ are needed to compute the general relativistic light-deflection correction.

The catalogued position of the star may be represented by the vector in the barycentric system in units of A.U.:

$$\boldsymbol{r}_B\left(t_0^H\right) = r_0\left(\cos\delta_0\cos\alpha_0 \quad \cos\delta_0\sin\alpha_0 \quad \sin\delta_0\right)^{\mathrm{T}}, \tag{4.136}$$

with corresponding unit vector (direction),

$$\boldsymbol{p}_B\left(t_0^H\right) = \boldsymbol{r}_B\left(t_0^H\right)\big/r_0 . \tag{4.137}$$

From equation (4.59), the coordinate vector of the star at time, t, due to proper motion is given by

$$\boldsymbol{r}_B(t) = \boldsymbol{r}_B\left(t_0^H\right) + \tau^H\dot{\boldsymbol{r}}_B\left(t_0^H\right), \tag{4.138}$$

where from equations (4.61), (4.63), and (4.133)

$$\dot{\boldsymbol{r}}_B\left(t_0^H\right) = r_0\begin{pmatrix} v\pi\cos\delta_0\cos\alpha_0 - \mu_{\alpha*}\sin\alpha_0 - \mu_\delta\sin\delta_0\cos\alpha_0 \\ v\pi\cos\delta_0\sin\alpha_0 + \mu_{\alpha*}\cos\alpha_0 - \mu_\delta\sin\delta_0\sin\alpha_0 \\ v\pi\sin\delta_0 + \mu_\delta\cos\delta_0 \end{pmatrix} = r_0\boldsymbol{m}\left(t_0^H\right), \tag{4.139}$$

which defines the vector, $\boldsymbol{m}\left(t_0^H\right)$. Note that

$$\boldsymbol{m}\left(t_0^H\right) = \frac{\dot{\boldsymbol{r}}_B\left(t_0^H\right)}{r_0} = v\pi\boldsymbol{p}_B\left(t_0^H\right) + \begin{pmatrix} -\mu_{\alpha*}\sin\alpha_0 - \mu_\delta\sin\delta_0\cos\alpha_0 \\ \mu_{\alpha*}\cos\alpha_0 - \mu_\delta\sin\delta_0\sin\alpha_0 \\ \mu_\delta\cos\delta_0 \end{pmatrix}. \tag{4.140}$$

It is important to ensure that all terms in equation (4.139) have the same units ([rad/cent] in this case, since τ^H, equation (4.134), is a fraction of a century). Also, note that the correction for proper motion is performed in the celestial reference system (the ICRS); that is, the vector, $\boldsymbol{r}_B(t)$, does not indicate mean coordinates at the epoch of date, because precession has not yet been applied. If the radial distance and velocity have no effect in a linear approximation of the angular proper motion, then the unit vector, $\boldsymbol{p}_B(t) = \boldsymbol{r}_B(t)/r_0$, can be approximated using equation (4.64),

$$\boldsymbol{p}_B\left(t\right)\approx\begin{pmatrix}\cos\left(\delta_0+\tau^H\dot{\delta}_0\right)\cos\left(\alpha_0+\tau^H\dot{\alpha}_0\right)\\\cos\left(\delta_0+\tau^H\dot{\delta}_0\right)\sin\left(\alpha_0+\tau^H\dot{\alpha}_0\right)\\\sin\left(\delta_0+\tau^H\dot{\delta}_0\right)\end{pmatrix},\tag{4.141}$$

where $\dot{\alpha}_0$ may need to be derived from the catalogue data, if given through equation (4.133).

The corrections of the other effects continue to be based on information described in the reference coordinate system (ICRS). We proceed by transforming from the barycentric to the geocentric system, which corrects for parallax (see Figure 4.11):

$$\boldsymbol{r}_G\left(t\right)=\boldsymbol{r}_B\left(t\right)-\boldsymbol{E}_B\left(t\right).\tag{4.142}$$

Now, substituting equations (4.138), (4.140), and (4.63), we have

$$\begin{aligned}\boldsymbol{U}_G\left(t\right)&=\frac{\boldsymbol{r}_G\left(t\right)}{r_0}=\boldsymbol{p}_B\left(t_0^H\right)+\tau^H\,\frac{\dot{\boldsymbol{r}}_B\left(t_0^H\right)}{r_0}-\frac{1}{r_0}\boldsymbol{E}_B\left(t\right)\\&=\left(1+\tau^H v\pi\right)\boldsymbol{p}_B\left(t_0^H\right)+\tau^H\begin{pmatrix}-\mu_{\alpha*}\sin\alpha_0-\mu_\delta\sin\delta_0\cos\alpha_0\\\mu_{\alpha*}\cos\alpha_0-\mu_\delta\sin\delta_0\sin\alpha_0\\\mu_\delta\cos\delta_0\end{pmatrix}-\pi\boldsymbol{E}_B\left(t\right)\end{aligned}\tag{4.143}$$

where the components of $\boldsymbol{E}_B\left(t\right)$ are given in terms of AU. We define

$$\boldsymbol{p}_G\left(t\right)=\frac{\boldsymbol{U}_G\left(t\right)}{\left|\boldsymbol{U}_G\left(t\right)\right|}\tag{4.144}$$

to be the unit vector corresponding to $\boldsymbol{U}_G\left(t\right)$. These coordinates still refer to the celestial reference system, but now with the effects of annual parallax and proper motion applied. Using just the angles, we can augment equation (4.141) to get a first-order approximation,

$$\boldsymbol{p}_G\left(t\right)\approx\begin{pmatrix}\cos\left(\delta_0+\tau\,\dot{\delta}_0+\Delta\delta\right)\cos\left(\alpha_0+\tau\,\dot{\alpha}_0+\Delta\alpha\right)\\\cos\left(\delta_0+\tau\,\dot{\delta}_0+\Delta\delta\right)\sin\left(\alpha_0+\tau\,\dot{\alpha}_0+\Delta\alpha\right)\\\sin\left(\delta_0+\tau\,\dot{\delta}_0+\Delta\delta\right)\end{pmatrix},\tag{4.145}$$

where $\Delta\alpha$ and $\Delta\delta$ account for annual parallax and are given, respectively, by equations (4.91) and (4.93).

One can now apply corrections for gravitational light-deflection and aberration according to specific models. The light-deflection model utilizes $\boldsymbol{E}_H(t)$ and $\boldsymbol{S}_B(t)$ and the reader is referred to (Seidelmann, 1992, p.149). We neglect this part as it only affects stars viewed near the sun. The annual aberration can be included using vectors, according to equation (4.65), where the aberrated coordinates are given in the form of a unit vector by

$$\boldsymbol{p}_G'(t) = \frac{\boldsymbol{p}_G(t) + \dot{\boldsymbol{E}}_B(t)/c}{\left|\boldsymbol{p}_G(t) + \dot{\boldsymbol{E}}_B(t)/c\right|}, \tag{4.146}$$

and, if $\dot{\boldsymbol{E}}_B(t)$ is given in units of [AU/day], then the speed of light should be expressed accordingly: $c = 173.1446 \text{ AU/day}$. The formula given in the Astronomical Almanac (Section B) includes special relativistic effects,

$$\boldsymbol{p}_G'(t) = \frac{\left(\sqrt{1-V(t)^2}\,\boldsymbol{p}_G(t) + \left(1 + \dfrac{\boldsymbol{p}_G(t)\cdot\boldsymbol{V}(t)}{1+\sqrt{1-V(t)^2}}\right)\boldsymbol{V}(t)\right)}{1+\boldsymbol{p}_G(t)\cdot\boldsymbol{V}(t)}, \tag{4.147}$$

where $\boldsymbol{V}(t) = \dot{\boldsymbol{E}}_B(t)/c$, $V(t) = \left|\boldsymbol{V}(t)\right|$. Alternatively, to first-order approximation, one can simply augment the angular coordinates in equation (4.145) with the changes due to aberration given by equations (4.75) and (4.80). In any case, the result yields coordinates at the epoch of date that are geocentric and aberrated by Earth’s velocity, but still referring to the celestial reference system (ICRS).

Finally, we apply precession and nutation to bring the coordinates from the ICRS to the apparent coordinates in the intermediate (instantaneous) celestial frame. One may apply the traditional transformations, as in equation (4.37) (called the *equinox method*). However, the small offset (frame bias) between the dynamical system and the new definition of the celestial reference system should then be included. Thus,

$$\boldsymbol{p}_{NP}(t) = \mathbf{N}(t)\,\mathbf{P}(t,t_0)\,\mathbf{B}\;\boldsymbol{p}_G'(t), \tag{4.148}$$

where $\mathbf{P}$ and $\mathbf{N}$ are given, respectively, by equations (4.17) and (4.35), and from equation (4.53),

$$\mathbf{B} = \mathbf{R}_1(-\eta_0)\,\mathbf{R}_2(\xi_0)\,\mathbf{R}_3(d\alpha_0). \tag{4.149}$$

Since $\boldsymbol{p}_G'(t)$ is a unit vector, so is $\boldsymbol{p}_{NP}(t)$; and, its components contain the apparent coordinates of the star, with $\boldsymbol{p}(t) = \begin{pmatrix} p_x & p_y & p_z \end{pmatrix}^{\mathrm{T}}$:

$$\alpha = \tan^{-1}\frac{p_y}{p_x}, \qquad \delta = \tan^{-1}\frac{p_z}{\sqrt{p_x^2 + p_y^2}}. \tag{4.150}$$

Here α refer to the right ascension with origin at the equinox.

Using the new conventions (Section 4.1.3), the alternative transformation procedure (called the *CIO method*) substitutes equation (4.39) for equation (4.148), where $\mathbf{Q}$ is given by equation (4.48) with *X*, *Y*, *s*, and *a* shown in equations (4.49), (4.50), (4.51), and (4.58), respectively:

$$\boldsymbol{p}_Q(t) = \mathbf{Q}^{\mathrm{T}}\boldsymbol{p}_G'(t). \tag{4.151}$$

The two methods (equinox and CIO methods) do *not* give the same result (even if the same precession/nutation model is used) since one refers the right ascension to the equinox of date and the other to the non-rotating origin, the CIO. The declination in the intermediate system, however, is the same with both methods. The corresponding apparent celestial coordinates are given by equation (4.150).

To bring the coordinates of the star to the Terrestrial Reference Frame requires a transformation that accounts for Earth's rotation rate and for polar motion. We have for the equinox method,

$$\boldsymbol{p}_T(t) = \mathbf{W}^{\mathrm{T}}(t)\mathbf{R}_3\left(GAST(t)\right)\boldsymbol{p}_{NP}(t), \tag{4.152}$$

where $GAST(t)$ is Greenwich Apparent Sidereal Time (Section 2.3.4; also Section 5.1, equation 5.32), and $\mathbf{W}$ is the polar motion matrix, given by equation (4.118). The coordinates, $\boldsymbol{p}_T(t)$, are the apparent coordinates of the star at time, t, in the Terrestrial Reference Frame. With the new conventions, the $GAST$ in the transformation (4.152) is replaced by a time angle that refers to the CIO,

$$\boldsymbol{p}_T(t) = \mathbf{W}^{\mathrm{T}}(t)\mathbf{R}_3\left(\theta(t)\right)\boldsymbol{p}_Q(t). \tag{4.153}$$

The angle, θ, is the Earth rotation angle, defined in Section 5.2.1. The polar motion matrix, $\mathbf{W}$, is the same as before, but the extra rotation, s', may be included for higher accuracy (equation (4.125)).

4.3.3.2 Topocentric Place Algorithm

Topocentric coordinates of stars are obtained by applying diurnal aberration using the terrestrial position coordinates of the observer. Diurnal parallax can be ignored for stars, as noted earlier. Furthermore, the topocentric coordinates and the velocity of the observer need only be approximate without consideration of polar motion. We first find the observer's geocentric position in the inertial frame:

$$\boldsymbol{g}(t) = \mathbf{R}_3(-GAST)\boldsymbol{r}, \tag{4.154}$$

where $\boldsymbol{r}$ is the terrestrial position vector of the (stationary) observer (Earth-fixed frame). $\boldsymbol{g}(t)$ gives "true" coordinates at the time of observation. We find the velocity, $\dot{\boldsymbol{g}}(t)$, according to

$$\dot{\boldsymbol{g}}(t) = \mathbf{R}_3(-GAST)\begin{pmatrix} 0 & -\omega_e & 0 \\ \omega_e & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}\boldsymbol{r}, \tag{4.155}$$

since $GAST = \omega_e(t - t_1)$, where t_1 is the time when the terrestrial and celestial reference systems are parallel, and apply nutation and precession to obtain the geocentric velocity in the mean coordinate system of the fundamental epoch, t_0:

$$\dot{\boldsymbol{G}}(t) = \mathbf{P}^{\mathrm{T}}(t, t_0)\mathbf{N}^{\mathrm{T}}(t)\dot{\boldsymbol{g}}(t). \tag{4.156}$$

This neglects a small Coriolis term which occurs when taking time-derivatives in a rotating (true) system. Now the velocity of the observer, due to Earth's rotation and orbital velocity, in the barycentric reference system is given by

$$\dot{\boldsymbol{O}}_B(t) = \dot{\boldsymbol{E}}_B(t) + \dot{\boldsymbol{G}}(t), \tag{4.157}$$

which would be used in equation (4.146) or (4.147) instead of $\dot{\boldsymbol{E}}_B(t)$. The result, equation (4.148) or (4.151), is then the *topocentric* place of the star. A similar procedure may be applied using the new conventions.

A complete set of computational tools is available from the U.S. Naval Observatory on its internet site: http://aa.usno.navy.mil/software/novas/novas_info.php. These are FORTRAN, C, or Python programs that compute the various transformations discussed above with the older, as well as the new conventions. Details may be found, e.g., in (Kaplan et al. 2011).

4.3.3.3 Problems

1. Given the mean celestial coordinates of a star: $\alpha = 195°$, $\delta = 23°$ in the ICRS (assume that the coordinates refer to $t_0 = J2000.0$), determine the apparent coordinates of the star for noon, 4 July 2020, in Greenwich. Apply the IAU2006 precession model and 2000 nutation model (18.6 year, semi-annual, and fortnightly terms, only), parallax, aberration, and space motion. Also apply the frame bias. Use the Julian day calendar available in the Astronomical Almanac and the following information:

$$\dot{\alpha}_0 = -0.003598723 \text{ rad/cent} ,$$
$$\dot{\delta}_0 = +0.000337430 \text{ rad/cent} ,$$
$$\dot{r}_0 = -22.2 \text{ km/s} ,$$

$$\pi = 3.6458 \times 10^{-6} \text{ rad} ,$$

$$\boldsymbol{E}_B(t) = \begin{pmatrix} 0.200776901 & -0.911150265 & -0.394806169 \end{pmatrix}^{\mathrm{T}} \text{ AU} ,$$
$$\dot{\boldsymbol{E}}_B(t) = \begin{pmatrix} 16551216 & 3183909 & 1380187 \end{pmatrix}^{\mathrm{T}} \times 10^{-9} \text{ AU/day} .$$

Chapter 5

Time

A system of time is a *system* just like any other reference system (see Section 1.2), except that it is one-dimensional. The definition of a time system involves some kind of theory associated with changing phenomena. If the universe in its entirety were completely static, there would be no time as we understand it, and the only reason we can perceive time is that things change. We have relatively easy access to *units* of time because many of the changes that we observe are periodic. If the changing phenomenon varies uniformly, then the associated time *scale* is uniform. Clearly, if we wish to define a time system then it should have a uniform time scale; however, very few observed dynamical systems have rigorously uniform time units. In the past, Earth's rotation provided the most suitable and evident phenomenon to represent the time scale, with the unit being a (solar) day. It has been recognized for a long time, however, that Earth's rotation is not uniform (it is varying at many different scales (daily, bi-weekly, monthly, etc., and even slowing down over geologic time scales (Lambeck 1988, p.607)). In addition to scale or units, we need to define an origin for our time system; that is, a zero-point, or an epoch, at which a value of time is specified. Finally, whatever system of time we define, it should be accessible and, thereby, realizable, giving us a time *frame*.

Prior to 1960, a second of time was *defined* as $1/86400$ of a mean solar day. Today (since 1960), the time scale is defined by the natural oscillation of the cesium atom and all time systems can be referred or transformed to this scale. Specifically, the SI (*Système International*) second is defined as:

1 SI sec = 9,192,631,770 oscillations of the cesium-133 atom between two hyperfine levels of the ground state of this atom. (5.1)

This definition has been refined to specify that the atom should be at rest (0° K) and at mean sea level, thus independent of ambient radiation effects and relativistic gravitational changes. Corrections are applied to actual measurements to comply with these requirements. The value of the SI second was set to the previously (in 1956) adopted value of a second of *ephemeris time* (Section 5.3), defined as $1/31556925.9747$ of a mean tropical (solar) year, being computed for the epoch, 1 January 1900, on the basis of Newcomb's theory of motion of the Earth around the Sun (Seidelmann 1992).

Although the SI second now defines the fundamental time unit (*atomic time*), one still distinguishes between systems of time that have different origins and even different scales depending on the application. *Dynamic time* is the independent variable in the most complete theory of the dynamics of the solar system. It is uniform by definition. *Mean solar time*, or *universal time*, is the time scale based on Earth's rotation with respect to the Sun and is used for general civilian time keeping. Finally, *sidereal time* is defined by Earth's rotation with respect to the celestial sphere. We already encountered sidereal time when discussing astronomic coordinates (Section 2.3) and dynamic time when discussing precession and nutation (Section 4.1). These are presented again with a view toward transformation between all time systems.

5.1 Sidereal Time

Sidereal time, generally, is the hour angle of the vernal equinox; it represents the rotation of the Earth with respect to the celestial sphere and reflects the actual rotation rate of the Earth, plus effects due to precession and nutation of the equinox. Because of the nutation, we distinguish between *apparent sidereal time* (AST), which is the hour angle of the true current vernal equinox, and *mean sidereal time* (MST), which is the hour angle of the mean vernal equinox (also at the current time).

The fundamental unit in the sidereal time system is the *mean sidereal day*, which equals the interval between two consecutive transits of the mean vernal equinox across the same meridian (corrected for polar motion). Also,

$$\text{1 sidereal day} = \text{24 sidereal hours} = \text{86400 sidereal seconds.} \tag{5.2}$$

The apparent sidereal time is not used as a time scale because of its non-uniformity, but it is used as an epoch in astronomical observations. The relationship between mean and apparent sidereal time derives from nutation. Referring to Figure 4.6, we have

$$AST = MST + \Delta\psi \cos \varepsilon \,, \tag{5.3}$$

where the last term is called the "equation of the equinoxes" and is the right ascension of the mean equinox with respect to the true equinox and equator. Since the maximum-amplitude term in the series for the nutation in longitude is approximately $\left|\Delta\psi\right| \approx 17.2 \text{ arcsec}$, the magnitude of the equation of the equinoxes is $17.2\cos\left(23.44°\right) \text{ arcsec} = 1.05 \text{ s}$, using the conversion, $15° = 1 \text{ hr}$.

We specialize our definitions of sidereal time according to the astronomic meridian to which it refers, as follows: *local sidereal time* (LST) (mean, $LMST$, and apparent, $LAST$) and *Greenwich sidereal time* (GST) (mean, $GMST$, and apparent, $GAST$), where

$$GST = LST - \Lambda_t, \tag{5.4}$$

and the longitude, Λ_t, refers to the CIP, not the IRP. Clearly the equation of the equinoxes applies equally to GST and LST. Due to precession (in right ascension), 24 hours of sidereal time do not correspond exactly to one rotation of the Earth with respect to inertial space. The *rate* of general precession if the equinox in right ascension is approximately (using equation (4.13) with equations (4.21), (4.22), and (4.27)):

$$m = 4612.16 + 2.783\tau \text{ [arcsec/cent]}, \tag{5.5}$$

where τ is in Julian centuries (equation (4.28)). The amount for one day is

$$m\frac{1}{36525} = 0.126 \text{ arcsec/day} = 0.0084 \text{ s/day} = 6.11\times10^{-7} \text{ rad/day} = 7.07\times10^{-12} \text{ rad/s}. \tag{5.6}$$

It is the reason for introducing the non-rotating origin (Section 4.1.3).

5.2 Universal Time

Universal time is the time scale used for general civilian time keeping and is based (only approximately, since 1961) on the diurnal motion of the sun. However, the sun, as viewed by a terrestrial observer, moves neither on the celestial equator, nor (exactly) on the ecliptic because the motion is not uniform due to planetary gravitational effects on Earth's orbital motion. Therefore, the hour angle of the sun is not varying uniformly. For these reasons and the need for a uniform time scale, a so-called *fictitious*, or *mean sun* is introduced, and the corresponding time for the motion of the mean sun is known as *mean solar time* (MT). The basic unit of universal time is the *mean solar day*, being the time interval between two consecutive transits of the mean sun across the meridian. Analogous to sidereal time, equation (5.2),

1 mean solar day = 24 mean solar hours = 86400 mean solar seconds. (5.7)

Universal time (UT) is defined as mean solar time on the Greenwich meridian.

If t_M is the hour angle of the mean (or fictitious) sun with respect to the local meridian, then in terms of an *epoch* (an accumulated angle), mean solar time is given by

$$MT = t_M + 180° , \tag{5.8}$$

where we have purposely written the units in terms of angles on the celestial equator to denote an epoch. The angle, $180°$, is added because when it is noon (the mean sun is on the local meridian and $t_M = 0°$), the mean solar time epoch is 12 hours, or 180 degrees. Again, in terms of an angle, the universal time epoch in Greenwich is

$$UT = t_M^G + 180° . \tag{5.9}$$

The relationship between the universal time and mean sidereal time scales can be established once the right ascension of the mean sun, α_M, is determined. Always in terms of angles (epochs), we have from equations (2.182) and (5.9),

$$\begin{aligned} GMST &= \alpha_M + t_M^G \\ &= \alpha_M + UT - 180° \end{aligned} \tag{5.10}$$

The right ascension of the mean sun is determined on the basis of an empirical expression (based on observations), first obtained by Newcomb. Using modern adopted constants, the Astronomical Almanac (2015, p.B8) gives the following expression

$$\begin{aligned} GMST = \theta\left(d_{UT}\right) &+ 0.014506'' + 4612.156534''\tau + 1.3915817''\tau^2 \\ &- 0.00000044''\tau^3 - 0.000029956''\tau^4 - 3.68''\times 10^{-8}\tau^5 \end{aligned} \tag{5.11}$$

where τ is the usual fraction of a Julian century, equation (4.28), and $\theta\left(d_{UT}\right)$ is the Earth rotation angle (see Section 5.2.1), given by

$$\theta\left(d_{UT}\right) = 360°\left(0.7790572732640 + 0.00273781191135448 d_{UT} + d_{UT} \bmod 1\right), \tag{5.12}$$

where d_{UT} is the number of UT days since 1.5 January 2000 and $d_{UT} \bmod 1$ is the fraction of a day's interval past midday (12^{h} UT), i.e., $d_{UT} \bmod 1 = UT^{\mathrm{h}} / 24^{\mathrm{h}} - 0.5$. Defining $\tau_{UT} = d_{UT}/36525$ as the fraction of a Julian century of mean solar days, equation (5.12) then is

$$\theta\left(d_{UT}\right) = 100.46061837504 + 35999.48882240006\tau_{UT} + UT \text{ deg} ; \tag{5.13}$$

and equation (5.11) becomes

$$\begin{aligned} GMST - UT = {} & 100.4606224044844^\circ + 36000.769976992837^\circ \tau_{UT} \\ & +1.3915817'' \tau_{UT}^2 - 0.00000044'' \tau_{UT}^3 - 0.000029956'' \tau_{UT}^4 \\ & -3.68'' \times 10^{-8} \tau_{UT}^5 \end{aligned} \qquad (5.14)$$

where the distinction between τ and τ_{UT} is neglected (introducing errors less than a microsecond, as noted in the Astronomic Almanac, 2015, p.B8).

The universal time scale relative to the mean sidereal time scale is obtained by taking the derivative of equation (5.14) with respect to τ_{UT}. We have

$$\begin{aligned} \frac{d\left(GMST - UT\right)}{d\tau_{UT}} = {} & 36000.769976992837 \text{ [deg/cent]} \\ & +\left(0.000773101\tau_{UT} - 3.7\times 10^{-10}\, \tau_{UT}^2 + \cdots\right) \text{ [deg/cent]} \end{aligned} \qquad (5.15)$$

Hence, the number of degrees on the celestial equator between the epochs $GMST$ and UT after one mean solar day ($d\tau_{UT} = 1/36525$ cent) is

$$d\left(GMST - UT\right) = \left(\begin{aligned} & 36000.769976992837^\circ \\ & +\left(0.000773101\tau_{UT} - 3.7\times 10^{-10}\, \tau_{UT}^2 + \cdots\right) \text{ [deg]} \end{aligned} \right) \Big/ 36525\,; \qquad (5.16)$$

or, one mean solar day is a sidereal day (360° or 86400 sidereal seconds) plus the excess being the right-hand side, above, in degrees or sidereal seconds (see also Figures 5.1 and 5.2):

$$1^{\mathrm{d}}\left(MT\right) = 86400^{\mathrm{s}} + 236.5553674053^{\mathrm{s}} + \left(5.079924\times 10^{-6}\, \tau - 2.4\times 10^{-12}\, \tau^2\right) \text{ [s]}. \qquad (5.17)$$

From this we find

$$\begin{aligned} \frac{1^{\mathrm{d}}\left(MT\right)}{1^{\mathrm{d}}\left(MST\right)} & = \frac{86636.5553674053^{\mathrm{s}} + \left(5.079924\times 10^{-6}\, \tau_{UT} - 2.4\times 10^{-12}\, \tau_{UT}^2\right)\text{[s]}}{86400^{\mathrm{s}}} \\ & = 1.0027379093449686 + 5.8795\times 10^{-11}\, \tau - 2.8\times 10^{-17}\, \tau^2 \end{aligned} \qquad (5.18)$$

Neglecting the small secular terms:

$$1\text{ mean solar day} = 24^{\text{h}}03^{\text{m}}56.55537^{\text{s}}\text{ in sidereal time}$$

$$1\text{ mean sidereal day} = 23^{\text{h}}56^{\text{m}}04.09053^{\text{s}}\text{ in solar time} \tag{5.19}$$

A mean solar day is longer than a sidereal day because in order for the sun to return to the observer's meridian, the Earth must rotate an additional amount since it has advanced in its orbit and the sun is now in a different position on the celestial sphere (see Figure 5.1).

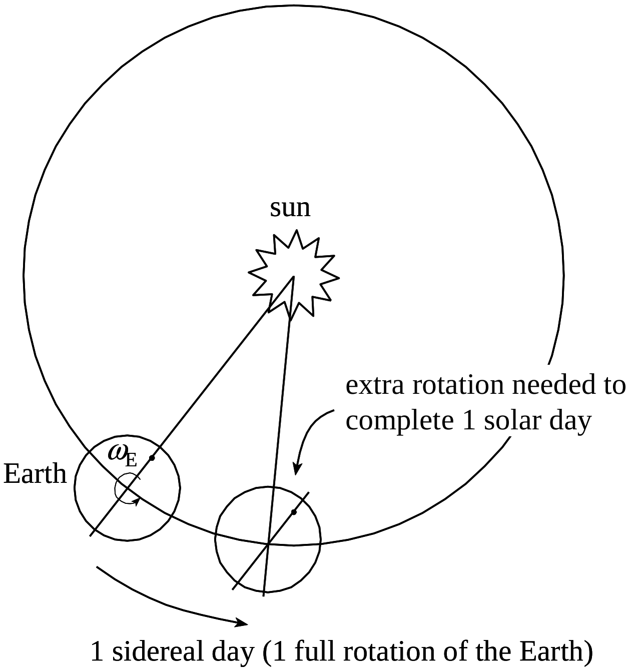

sun
extra rotation needed to complete 1 solar day
ω_E
Earth
1 sidereal day (1 full rotation of the Earth)

Figure 5.1: Geometry of sidereal and solar days.

It is noted that UT and ST are not uniform because of irregularities in Earth's rotation rate. The most important effect, however, in determining UT from observations is due to polar motion; that is, the meridian with respect to which the transit measurements are made refers to the IRP (fixed meridian on the Earth's surface), while UT should refer to the instantaneous rotation axis. Thus, one distinguishes between the epochs:

$UT0$: universal time determined from observations with respect to the fixed meridian (the IRP);

$UT1$: universal time determined with respect to the meridian attached to the CIP.

From Figure 4.21 we have

$$\Lambda_{\mathrm{CIP}} = \Lambda_{\mathrm{IRP}} - \Delta\Lambda , \tag{5.20}$$

where $\Delta\Lambda$ is the polar motion in longitude. Hence, as shown in Figure 5.2, the IRP meridian will pass a point on the celestial sphere before the CIP meridian (assuming, without loss in generality, that $\Delta\Lambda > 0$). Therefore, the $GMST$ epoch with respect to the IRP comes before the $GMST$ epoch with respect to the CIP:

$$GMST_{\mathrm{CIP}} = GMST_{\mathrm{IRP}} + \Delta\Lambda . \tag{5.21}$$

Thus, from equation (5.14),

$$\begin{aligned} UT1 &= GMST_{CIP} - \ldots = GMST_{IRP} + \Delta\Lambda - \ldots \\ &= UT0 + \Delta\Lambda \end{aligned} \tag{5.22}$$

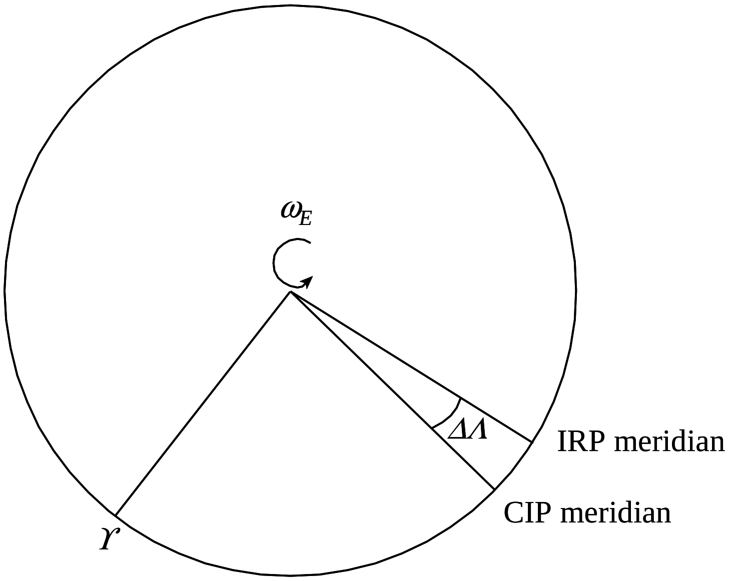


Figure 5.2: Geometry for the relationship between $UT0$ and $UT1$.

$UT1$ is still affected by irregularities in Earth's rotation rate (length of day variations), which can be removed to some extent (seasonal variations), thus yielding

$$UT2 = UT1 + \text{corrections for seasonal variations} . \tag{5.23}$$

Presently, $UT2$ is the best approximation of UT to a uniform time (although it is still affected by small secular variations). However, $UT1$ is used to define the orientation of the Greenwich mean astronomical meridian through its relationship to longitude, and $UT1$ has principal application when observations are referred to a certain epoch since it represents the true rotation of the Earth.

In terms of the SI second, the mean solar day is given by

$$1^d\left(MT\right)=86400-\frac{\Delta\tau}{n}\ [\mathrm{s}]\,, \tag{5.24}$$

where $\Delta\tau$, in seconds, is the difference over a period of n days between $UT1$ and dynamic time (see Section 5.3):

$$\Delta\tau=UT1-TDT\ . \tag{5.25}$$

The time-derivative of $\Delta\tau$ is also called the *length-of-day variation*. From observations over the centuries it has been found that the secular (average) variation in the length of a day (rate of Earth rotation) currently is of the order of 1.4 ms per century (Lambeck, 1988, p.607).

5.2.1 Earth Rotation Angle

With the definitions of the Celestial Intermediate Origin (CIO) and the Terrestrial Intermediate Origin (TIO), both being non-rotating origins on the instantaneous equator, we are able to define $UT1$ more succinctly. The angle between the CIO and the TIO (Figure 5.3) is known as the *Earth Rotation Angle,* θ. Since neither the CIO nor the TIO, by definition, have angular rate along the instantaneous equator due to precession/nutation and polar motion, the time associated with Earth's rotation rate, that is, $UT1$, is defined simply as being proportional to θ:

$$\theta\left(d_{UT}\right)=2\pi\left(\psi_0+\psi_1 d_{UT}\right), \tag{5.26}$$

where ψ_0 and ψ_1 are constants (with units of [cycle] and [cycle per day], respectively), and

$$d_{UT}=\text{ Julian } UT1 \text{ date } -t_0\,, \tag{5.27}$$

and the Julian $UT1$ date is the Julian day number interpreted as UT (mean solar time) scale. The fundamental epoch, t_0, is, as usual, the Julian day number, 2451545.0, associated with 1.5 January 2000 in Greenwich. In practice, the Julian $UT1$ day number is obtained from

$$UT1=UTC+\left(UT1-UTC\right), \tag{5.28}$$

where UTC is Coordinated Universal Time (an atomic time scale, see Section 5.4), and the difference, $UT1-UTC$, is either observed or provided by the IERS. The constants, ψ_0 and ψ_1, are derived below theory and models and are given by the Astronomic Almanac (2015, p.B8; see also equation (5.12):

$$\theta\left(d_{UT}\right)=2\pi\left(0.7790572732640+1.00273781191135448 d_{UT}\right). \tag{5.29}$$

The constant, $2\pi\psi_1=\omega_E$, is Earth's mean rotation rate in units of [rad/day], or in units of [rad/s],

$$\omega_E=7.29211514670698\times 10^{-5}\ \text{rad/s}. \tag{5.30}$$

If the new transformation, equation (4.39), with matrix, $\mathbf{Q}$, is used to account for precession and nutation, then the Earth Rotation Angle, θ, should be used instead of the Greenwich Apparent Sidereal Time ($GAST$), in the transformation between the Celestial and Terrestrial Reference Systems. The total transformation under the old conventions from the Celestial Reference System to the Terrestrial Reference System was given by equation (4.37) to account for precession and nutation, and by equation (4.152) to account for polar motion and Earth rotation, where we omit the observational effects, for the moment:

$$\boldsymbol{u}_{\mathrm{TRS}}(t)=\mathbf{W}^{\mathrm{T}}(t)\mathbf{R}_3(GAST)\mathbf{N}(t)\mathbf{P}(t,t_0)\mathbf{B}\,\boldsymbol{u}_{\mathrm{CRS}}(t_0), \tag{5.31}$$

where $\boldsymbol{u}$ is a unit vector on the celestial sphere, and $\mathbf{B}$ is included to account for the frame bias (Section 4.1.3; see also equation (4.148)). The new transformation, based on IAU resolutions adopted in 2000 and the new IERS 2003 Conventions, is

$$\boldsymbol{u}_{\mathrm{TRS}}(t)=\mathbf{W}^{\mathrm{T}}(t)\mathbf{R}_3(\theta)\mathbf{Q}^{\mathrm{T}}(t)\boldsymbol{u}_{\mathrm{CRS}}, \tag{5.32}$$

where the polar motion transformation, $\mathbf{W}$, is given by equation (4.125), and the precession-nutation transformation, $\mathbf{Q}$, is given by equation (4.47). The Greenwich Sidereal Time (GST) now is no longer explicitly involved in the transformation, but we can demonstrate the essential equivalence of the old and new methods of transformation through the relationship between the Earth Rotation Angle, θ, and GST.

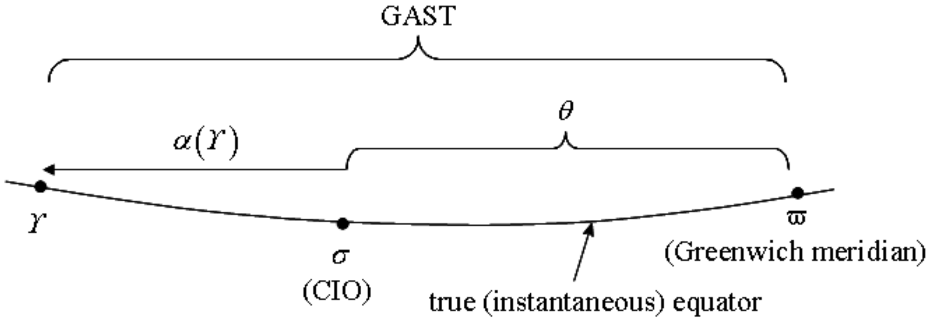


Figure 5.3: Relationship between $GAST$ and Earth Rotation Angle, θ. All right ascensions, whether from Υ or σ, are positive eastward. Thus, the indicated $\alpha(\Upsilon)$ is negative in this case.

From Figure 5.3, it is clear that if $GAST$ is the hour angle, at the TIO (ϖ), of the true vernal equinox at the epoch of date, t, then

$$GAST = \theta - \alpha(\Upsilon), \tag{5.33}$$

where $\alpha(\Upsilon) \le 0$ is the right ascension of the true equinox, Υ, at t relative to the CIO (indeed, due to precession and nutation since 2000, the right ascension of Υ relative to the CIO has accumulated to about $\alpha(\Upsilon) = -12.25'$ by the end of 2015). The angle, $\alpha(\Upsilon)$, is also called the "equation of origins" (analogous to the equation of the equinoxes, (5.3)). The old precession and nutation transformations, $\mathbf{P}$ and $\mathbf{N}$ (with the frame bias, $\mathbf{B}$), bring the reference 1-axis (reference equinox) to the true equinox of date. Therefore, a further rotation about the CIP by $-\alpha(\Upsilon)$ brings the 1–axis to the CIO, σ; and, we have

$$\mathbf{R}_3\left(-\alpha(\Upsilon)\right)\mathbf{NPB} = \mathbf{Q}^{\mathrm{T}}, \tag{5.34}$$

since the CIO is the point to which the transformation, $\mathbf{Q}^{\mathrm{T}}$, brings the 1-axis due to precession and nutation. Combining equations (5.33) and (5.34),

$$\mathbf{R}_3(\theta)\mathbf{Q}^{\mathrm{T}} = \mathbf{R}_3(GAST)\mathbf{NPB}, \tag{5.35}$$

showing that equations (5.31) and (5.32) are equivalent.

5.3 Dynamic Time

As already discussed in Chapter 4, the *dynamic time* (DT) scale is represented by the independent variable in the equations of motion of bodies in the solar system. In theory it is the most uniform time scale known since it governs all dynamics of our local universe according to the best theory (the theory of general relativity) that has been developed to date. Prior to 1977, the "dynamical" time was called *ephemeris time* (ET). ET was based on the time variable in the theory of motion of the sun relative to the Earth – Newcomb's ephemeris of the sun. This theory suffered from the omission of relativistic theory, the dependence on adopted astronomical constants that, in fact, show a time dependency (such as the "constant" of aberration). It also omitted the effects of planets on the motion.

In 1976 and 1979, the IAU adopted a new dynamic time scale based on the time variable in a relativistic theory of motion of all the bodies in the solar system. The two systems, ET and DT, were constrained to be consistent at their boundary (a particular epoch); specifically

$$DT = ET \text{ at 1977 January 1.0003725 } (1^{\text{d}}00^{\text{h}}00^{\text{m}}32.184^{\text{s}},\ \text{exactly})\,. \tag{5.36}$$

The extra fraction in this epoch was included since this would make the point of continuity between the systems exactly 1977 January 1.0 in atomic time, TAI (Section 5.4). This is the origin point of modern dynamic time. The unit for dynamic time is the SI second, or, also a Julian day of 86400 SI seconds.

With respect to the theory of general relativity, the dynamic time scale refers to a coordinate system and thus represent a *coordinate time*. Common choices include the barycentric reference system (origin at the center of mass of the solar system) or the geocentric reference system. The corresponding time scales are thus designated as Barycentric Coordinate Time (TCB) and Geocentric Coordinate Time (TCG). Note that acronyms for time systems generally follow the corresponding French names, e.g., "Temps-coordonne barycentrique" for Barycentric Coordinate Time. Dynamic time defined in this way is the fourth coordinate and transforms according to the theory of general relativity as the fourth coordinate from one point in space-time to another.

On the other hand, dynamic time has also been defined as a proper time, the time associated with the frame of the observer that a uniformly running clock would keep and that describes observed motions in that frame. We have:

TDT : *Terrestrial dynamic time* is the dynamic time scale of geocentric ephemerides of bodies in the solar system. It is *defined* to be uniform and the continuation of ET (which made no distinction between geocentric and barycentric coordinate systems). It is also identical, by resolution, to the time scale of terrestrial atomic physics.

TDB: *Barycentric dynamic time* is the time scale of barycentric ephemerides of bodies in the solar system. The difference between TDB and TDT is due to relativistic effects caused mainly by the eccentricity of Earth's orbit, producing periodic variations.

In 1991, as part of a clarification in the usage of these time scales in the context of general relativity, the IAU adopted a change in the name of TDT to Terrestrial Time (TT). TT is a *proper time*, meaning that it refers to intervals of time corresponding to events as measured by an observer in the same frame (world-line) as occupied by the event. By definition TT refers to proper time at the geoid (approximately mean sea level); it has the same origin defined by equation (5.36); and, its scale is defined by the SI second. However, in 2000 the IAU further recommended, due to uncertainties in the realization of the geoid, that TT be redefined as differing from TCG by a constant, specified rate. Its relation to a proper time then more precisely depends on the location and velocity of the observer's clock in the ambient gravitational field. For mathematical connections to the coordinate times, TCB and TCG, and to TDB reader is directed to Seidelmann (1992), McCarthy (1996, Petit and Luzum (2010, Chapter 10), and the Astronomical Almanac (Section B6). In calculations of Earth orientation (Chapter 4), the difference between TT and TDB is usually neglected.

5.4 Atomic Time

Atomic time refers to the time scale realized by the oscillations in energy states of the cesium-133 atom, as defined in equation (5.1). The SI second, thus, is the unit that defines the scale; this is also the time standard for *International Atomic Time* (TAI, for the French *Temps Atomique International*) which was officially introduced in January 1972. TAI is realized by the BIPM (Bureau International des Poids et Mesures) which combines data from over 400 high-precision atomic clocks around the world in order to maintain the SI-second scale as closely as possible. The TAI scale is published and accessible as a correction to each time-center clock. In the U.S., the official atomic time clocks are maintained by the U.S. Naval Observatory (USNO) in Washington, D.C., and by the National Institute of Standards and Technology (NIST) in Boulder, Colorado. Within each such center several cesium beam clocks are running simultaneously and averaged. Other participating centers include observatories in Paris, Greenwich, Moscow, Tokyo, Ottawa, Wettzell, Beijing, and Sydney, among over 70 others. The comparison and amalgamation of the clocks of participating centers around the world are accomplished by LORAN-C, satellite transfers (GPS playing the major role), and actual clock visits. Worldwide synchronization is about 100 ns (Leick, 1995, p.34). Since atomic time is computed from many clocks it is also known as a *paper clock* or a *statistical clock.*

Due to the exquisite precision of the atomic clocks, general relativistic effects due to the spatially varying gravitational potential must be considered. Therefore, the SI second is defined on the “geoid in rotation”, meaning also that TAI is defined for an Earth frame and not in a barycentric system.

Atomic time was not realized until 1955; and, from 1958 through 1968, the BIH maintained the atomic time scale. The origin, or zero-point, for atomic time has been chosen officially as $0^{\mathrm{h}}0^{\mathrm{m}}0^{\mathrm{s}}$, January 1, 1958. Also, it was determined and subsequently defined that on $0^{\mathrm{h}}0^{\mathrm{m}}0^{\mathrm{s}}$, January 1, 1977 ($TAI$), the ephemeris time epoch was $0^{\mathrm{h}}0^{\mathrm{m}}32.184^{\mathrm{s}}$, January 1, 1977 ($ET$). Thus, with the evolving definitions of dynamic time:

$$ET - TAI = TDT - TAI = TT - TAI = 32.184^{\mathrm{s}}. \tag{5.37}$$

So far, no difference in scale has been detected between TAI and TT, but their origins are offset by 32.184^{s}.

All civil clocks in the world now are set with respect to an atomic time standard. Atomic time is much more uniform than solar time, and yet we still would like civil time to correspond to solar time. Hence, a new atomic time scale was defined that keeps up with universal time in discrete steps. This atomic time scale is called *Universal Coordinated Time* (UTC). It is adjusted recurrently to stay close to universal time. UTC was established in 1961 by the BIH and is now maintained by the BIPM. Initially, UTC was adjusted so that

$$\left|UT2 - UTC\right| < 0.1\ \mathrm{s}, \tag{5.38}$$

which required that UTC be modeled according to

$$TAI - UTC = b + s\left(t - t_0\right), \tag{5.39}$$

where b is a step adjustment and s a frequency offset. As of 1972, the requirement for the correspondence of UTC with universal time was loosened to

$$\left|UT1 - UTC\right| < 0.9\ \mathrm{s}, \tag{5.40}$$

with $b = 1\ \mathrm{s}$ and $s = 0$. The step adjustment, b, is called a *leap second* and is introduced either 1 July or 1 January of any particular year. The last leap second (as of July 2016) was introduced on 1 July 2015, and currently (2016), $TAI - UTC = 36.0\ \mathrm{s}$. The next is anticipated for 1 January 2017.

The lengthening of a day by about 1:4 ms per century as measured by Earth’s slowing rate of rotation, due to tidal friction, implies that the $UT1$ clock continues to run more and more behind the TAI clock. It has been determined that the mean solar day today is actually about 86;400:0027 SI seconds long, since the SI second was originally identified with the ET second based on the motion of the mean sun at Newcomb’s time in the nineteenth century. Indeed, 86400 SI seconds exactly equaled a mean solar day in 1820, or 1.95 centuries ago. This disparity between the scales of the defined SI second and the current mean solar day has an accumulative effect that adds, on the average, about $1.4\ \text{ms/day/century} \times 1.95\ \text{century}$, or about 1 s to $UT1$ during the course of a year; hence, the introduction of the leap seconds. The difference,

$$DUT1 = UT1 - UTC\,, \tag{5.41}$$

is broadcast along with UTC so that users can determine $UT1$. There is current debate (Nelson et al. 2001) about the need to maintain a small difference between UTC and $UT1$ considering the technical inconveniences (if not outright difficulties) this imposes on the many civilian telecommunications systems and other networks that rely on a precise time scale.

GPS time ($GPST$) is also an atomic time scale, consistent with TAI to within $1\,\mu\text{s}$. Its zero point is $\text{January 6.0, 1980} = \text{JD2444244.5}$, and $GPST = UTC$ at that epoch only, since $GPST$ is not adjusted by leap seconds to keep up with universal time. Thus, we always have

$$GPST = TAI - 19.0\ \text{s}\,. \tag{5.42}$$

Other Global Navigation Satellite Systems (GNSS) follow their own conventions, though most are similar to GPS. For example, both Europe's Galileo (Zanello et al. 2007, Stehlin et al. 2006) and China's BeiDou (Han et al. 2011) global satellite navigation systems are offset by a constant with respect to TAI, with no leap second adjustments,

$$\text{Galileo System Time} = TAI - 19.0\ \text{s}\,, \tag{5.43}$$

$$\text{BeiDou Time} = TAI - 33.0\ \text{s}\,.$$

where the origin for the Galileo system time is the same as for $GPST$ and that of BeiDou Time is 1.0 January 2006 UTC. Fixed offsets are also adopted by the Japanese (QZSS) and Indian (IRNSS) regional satellite navigation systems. The time system for the Russian system, GLONASS, on the other hand, is tied to the scale of UTC, although in Moscow's time zone. That is,

$$\text{GLONASS System Time} = UTC + 3\ \text{hr}\,. \tag{5.44}$$

Besides incorporating leap seconds, GLONASS System Time is always three hours ahead of UTC because of the time zone difference between Greenwich and Moscow.

The relationships among the various atomic time scales are illustrated along with dynamic time in Figure 5.4. It should be realized that the time differences between the various satellite systems and TAI or UTC, as given by equations (5.42) through (5.44) describe only the nominal integer-second offsets, but omit the small fractional offsets (tens, up to a hundred or more ns) that occur in the actual realization of these time scales. These small differences are determined relative to laboratory master atomic clocks and published, for example, by the BIPM (see also Lewandowski and Aria 2011).

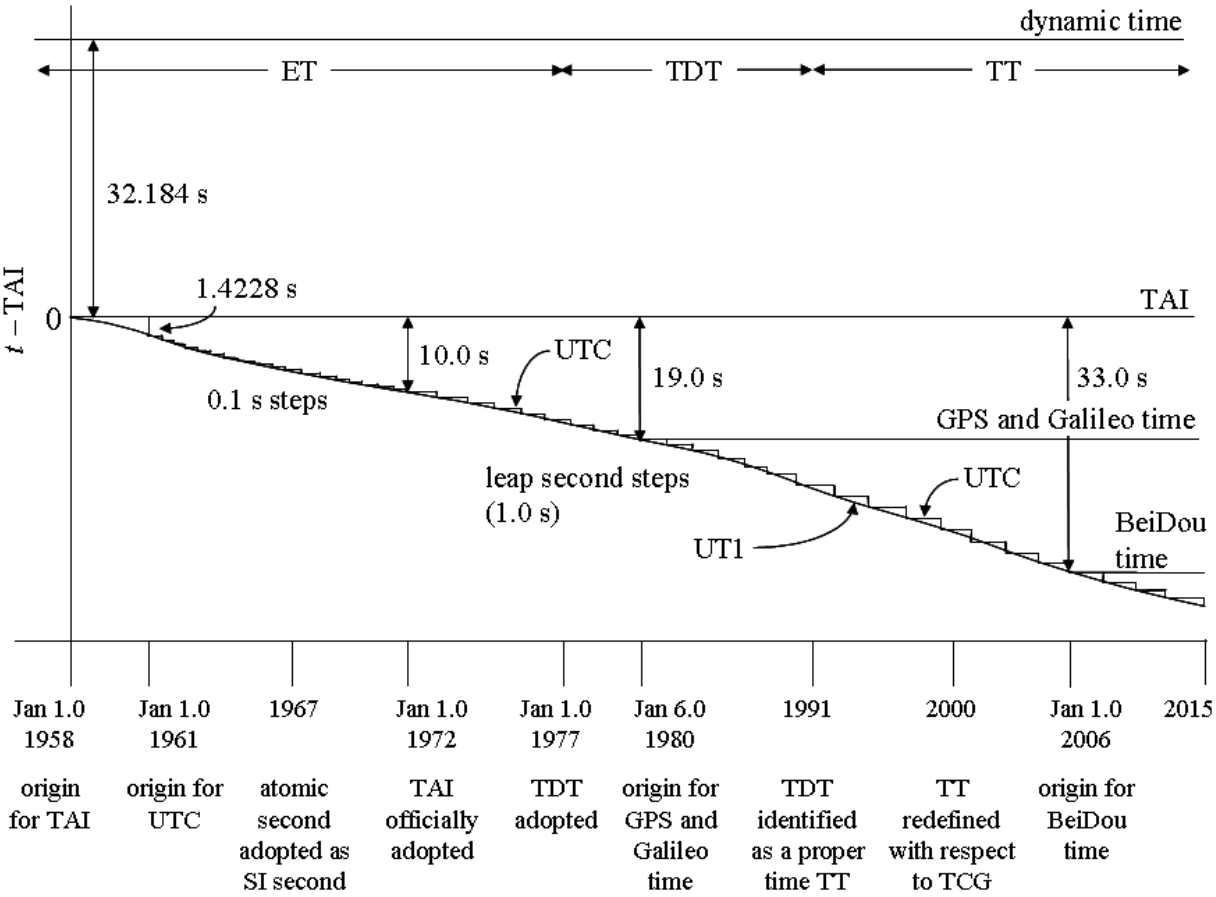


Figure 5.4: Relationships between atomic time scales and dynamic time (indicated leap seconds are schematic only). Until 1960, the second was defined by mean Earth rotation. Acronyms are explained fully in the text.

Note in Figure 5.4 that the time scales of TAI and TDT (or, TT) are the same (1 SI second is the same in both), but they are offset by a constant that is fixed for all time. Also, the time scale for UTC is 1 SI second, but occasionally it is offset by 1 s. The time scale for $UT1$ is very close to 1 SI second; that is, the difference between the $UT1$ and the TAI "clocks" is only about 36 s over about 60 years (compare this to the difference between mean solar time and mean sidereal time of 4 minutes per day!). The history of $TAI-UTC$ (only schematically shown in Figure 5.4) can be obtained from the USNO[1]. Note, however, that this does *not* mean that the

[1] ftp://maia.usno.navy.mil/ser7/tai-utc.dat

Earth has slowed down at the *rate* of 36 s in sixty years. The continual slowing of the $UT1$ clock relative to the TAI clock represents the *accumulative* effect of Earth's decreasing rate of rotation (a deceleration), which is only about 1.4 msec per day per century.

5.4.1 Determination of Atomic Time

Atomic time is currently the most precise and accessible of the uniform scales of time. It is determined using *frequency standards*, or atomic clocks, that are based on atomic energy oscillations. The standard for comparison is based on the oscillations of the cesium atom, but other atomic clocks are used with different characteristics in stability and performance. For any signal generator, considered as a clock, we assume a nearly perfect sinusoidal signal voltage,

$$V(t) = \left(V_0 + \delta V(t)\right)\sin\phi(t), \tag{5.45}$$

where $\delta V(t)$ is the error in amplitude, which is of no consequence, and $\phi(t)$ is the phase of the signal. The change in phase with respect to time is a measure of time. The phase is given by

$$\phi(t) = \omega t + \delta\phi(t), \tag{5.46}$$

where ω is the ideal (radian) frequency of the generator (i.e., ω is constant), and $\delta\phi(t)$ represents the phase error; or, its time derivative, $\delta\dot{\phi}(t)$, is the frequency error. Note that in terms of cycles per second, the frequency is

$$f = \frac{\omega}{2\pi}. \tag{5.47}$$

Thus, let

$$y(t) = \frac{1}{\omega}\delta\dot{\phi}(t) = \frac{1}{2\pi f}\delta\dot{\phi}(t) \tag{5.48}$$

be the *relative frequency error*.

Now, the average of the relative frequency error over some interval, $\tau = t_{k+1} - t_k$, is given by

$$\overline{y}_k = \frac{1}{\tau}\int\limits_{t_k}^{t_{k+1}} y(t)\,dt = \frac{1}{2\pi f\tau}\left(\delta\phi(t_{k+1}) - \delta\phi(t_k)\right). \tag{5.49}$$

The stability of the clock, or its performance, is characterized by the sample variance of the first N differences of contiguous averages, $\bar{y}_k$, with respect to the interval, τ:

$$\sigma_y^2(\tau) = \frac{1}{N}\sum_{k=1}^{N}\frac{1}{2}\left(\bar{y}_{k+1} - \bar{y}_k\right)^2 . \tag{5.50}$$

This is known as the *Allan variance*, and σ_y represents the *fractional frequency stability* of the oscillator. Substituting equation (5.46) into equation (5.49) yields

$$\bar{y}_k = \frac{1}{\omega\tau}\left(\phi(t_{k+1}) - \phi(t_k) - \omega\tau\right). \tag{5.51}$$

Putting this into equation (5.50) gives

$$\sigma_y^2(\tau) = \frac{1}{2N(\omega\tau)^2}\sum_{k=1}^{N}\left(\phi(t_{k+2}) - 2\phi(t_{k+1}) + \phi(t_k)\right)^2 , \tag{5.52}$$

which is a form that can be used to compute the Allan variance from the indicated phase, $\phi(t)$, of the oscillator.

Most atomic clocks exhibit a stability as a function of τ, characterized generally by $\sigma_y(\tau)$ decreasing as τ increases from near zero to an interval of the order of a second. Then, $\sigma_y(\tau)$ reaches a minimum over some range of averaging times; this is called the “flicker floor” region and yields the figure of merit in terms of stability. For longer averaging times, after this minimum, $\sigma_y(\tau)$ again rises. Table 5.1 is constructed from the discussion by Seidelmann (1992, p.60-61); and, Figure 5.5 qualitatively depicts the behavior of the square root of the Allan variance of different types of clocks as a function of averaging time (from Vig (1992); see also Kamas and Howe 1979). Recently, NIST reported a fractional frequency stability of 10^{-18} for a ytterbium atomic clock (Hinkley et al. 2013)[2].

[2] see also http://www.nist.gov/pml/div688/clock-082213.cfm

Table 5.1: Fractional frequency stabilities for various atomic (and other) clocks.

Clock		stability (min σ_y)
quartz oscillator		$\geq 10^{-13}$
cesium beam	laboratory	1.5×10^{-14}
	commercial	2×10^{-12}
		3×10^{-14}
	Block II GPS	$O\left(10^{-14}\right)$
rubidium	laboratory	$\geq 10^{-13}$
	GPS	2×10^{-13}
hydrogen maser		2×10^{-15}

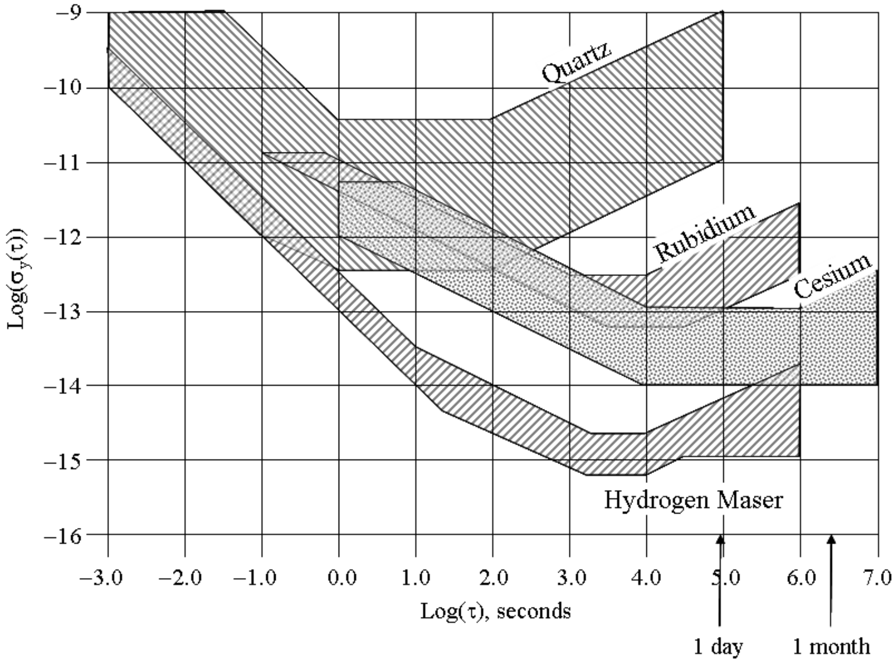


Figure 5.5: Fractional frequency stability for various clocks. Transcribed from Vig (1992) (http://www.oscilent.com/esupport/TechSupport/ReviewPapers/IntroQuartz/vigcomp.htm)

References

Amos, M. (2010): New Zealand Vertical Datum 2009 – a geoid based height system for the unification of disparate local vertical datums. Presented at FIG Congress 2010, Sydney, Australia, 11-16 April 2010.

Altamimi, Z., Boucher, C., Sillard, P. (2002a): New trends for the realization of the International Terrestrial Reference System. *Adv. Space Res.*, 30(2):175-184.

Altamimi, Z., Sillard, P., Boucher, C. (2002b): ITRF2000: A new release of the International Terrestrial Reference Frame for earth science applications. *J. Geophys. Res.*, 107(B10):2214, doi:10.1029/2001JB000561.

Arfken, G. (1970): *Mathematical Methods for Physics*. Academic Press, New York.

Astronomical Almanac, issued annually by the Nautical Almanac Office of the U.S. Naval Observatory, Washington, D.C.

Bomford, G. (1971): *Geodesy*, 3rd edition. Oxford University Press.

Borkowski, K.M. (1989): Accurate algorithms to transform geocentric to geodetic coordinates. *Bulletin Géodésique*, 63:50-56.

Capitaine, N. (1990): The celestial pole coordinates. *Celes. Mech. Dyn. Astr.*, 48:127-143.

Capitaine, N. (2002): Comparison of “old” and “new” concepts: the celestial intermediate pole and Earth orientation parameters. In: IERS Technical Note No. 29, Capitaine, N., et al. (eds.), Verlag des Bundesamts für Kartographie und Geodäsie, Frankfurt am Main. Available on-line: http://www.iers.org/iers/publications/tn/tn29/.

Capitaine, N., Guinot, B., Souchay, J. (1986): A non-rotating origin on the instantaneous equator - definition, properties, and use. *Celestial Mechanics*, 39:283-307.

Capitaine, N., Wallace, P.T., Chaprone, J. (2003): Expressions for IAU 2000 precession quantities. *Astronomy and Astrophysics*, 412(2):567-586.

Capitaine, N., Wallace, P.T., Chapront, J. (2005): Improvement of the IAU 2000 precession model. *Astronomy and Astrophysics*, 432(1):355 – 367.

Craymer, M., Ferland, R. Snay, R.A. (2000): Realization and unification of NAD83 in Canada and the U.S. via the ITRF. In: Rummel, R., H. Drewes, W. Bosch, H. Hornik (eds.), *Towards an Integrated Global Geodetic Observing System (IGGOS)*. IAG Symposia, vol.120, pp.118-21, Springer-Verlag, Berlin.

Dehant, V., Mathews, P.M. (2015): *Precession, Nutation, and Wobble of the Earth.*, Cambridge University Press, Cambridge, U.K.

DMA (1987): Supplement to Department of Defense World Geodetic System 1984 Technical Report, Part I. DMA TR 8350.2-A, Defense Mapping Agency, Washington, D.C.

Ehlert, D. (1993): Methoden der ellipsoidischen Dreiecksberechnung. Report no.292, Institut für Angewandte Geodäsie, Frankfurt a. Main, Deutsche Geodätische Kommission.

Ewing, C.E., Mitchell, M.M. (1970): *Introduction to Geodesy*. Elsevier Publishing Co., Inc., New York.

Feissel, M., Mignard, F. (1998): The adoption of ICRS on 1 January 1998: Meaning and consequences. *Astron. Astrophys.*, 331:L33-L36.

Fischer, I. (1975): Another look at Eratosthenes' and Posidonius' determinations of the Earth's circumference. *Quarterly Journal of the Royal Astronomical Society*, 16:152-167.

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.

Guinot, B. (2000): History of the Bureau International de l'Heure. In: Dick, S., McCarthy, D., Luzum, B. (eds.) (2000): *Polar Motion: Historical and Scientific Problems, ASP Conference Series*, 208, pp.175-184.

Han, C., Yang, Y., Cai, Z. (2011): BeiDou navigation satellite system and its time scales, *Metrologia*, 48(4):S213-S218.

Heiskanen, W.A., Moritz, H. (1967): *Physical Geodesy*. Freeman and Co., San Francisco.

Hinkley, N., Sherman, J.A., Phillips, N.B., Schioppo, M., Lemke, N.D., Beloy, K., Pizzocaro, M., Oates, C.W., Ludlow, A.D. (2013): An atomic clock with 10^{-18} instability. *Science*, 341(6151):1215-1218.

IAG (1992): Geodesist's Handbook. *Bulletin Géodésique*, 66(2):132-133.

Jordan, W. (1962): *Handbook of Geodesy*, vol.3, part 2. English translation of Handbuch der Vermessungskunde (1941), by Martha W. Carta, Corps of Engineers, United States Army, Army Map Service.

Kamas, G., Howe, S. (1979): Time and frequency users' manual. NBS Special Publication 559, National Bureau of Standards, Boulder, Colorado.

Kaplan, G., Bartlett, J., Monet, A., Bangert, J., Puatua, W. (2011): User's Guide to NOVAS Version F3.1, Naval Observatory Vector Astrometry Software. U.S. Naval Observatory, Washington, D.C. http://www.usno.navy.mil/USNO/astronomical-applications/software-products/novas/novas-fortran/NOVAS_F3.1_Guide.pdf.

Kinoshita, H. (1977): Theory of the rotation of the rigid Earth. *Celestial Mechanics*, 15(3):277 – 326.

Lambeck, K. (1988): *Geophysical Geodesy*. Clarendon Press, Oxford.

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

Lewandowski, W., Aria, E.F. (2011): GNSS times and UTC. *Metrologia*, 48:S219-S224.

Lieske, J.H., Lederle, T., Fricke, W., Morando, B. (1977): Expressions for the Precession quantities based upon the IAU (1976) system of astronomical constants. *Astron. Astrophys.*, 58:1-16.

Mathews, P.M., Herring, T.A., Buffett, B.A. (2002): Modeling of nutation-precession: New nutation series for nonrigid Earth, and insights into the Earth's interior. *J. Geophys. Res.*, 107(B4), doi:10.1029/2001JB000390.

McCarthy, D.D. (ed.) (1992): IERS Conventions (1992). IERS Tech. Note 13, Observatoire de Paris, Paris.

McCarthy, D.D. (ed.) (1996): IERS Conventions (1996). IERS Tech. Note 21, Observatoire de Paris, Paris.

McCarthy, D.D., Petit, G. (2003): IERS Conventions 2003. IERS Technical Note 32, U.S. Naval Observatory, Bureau International des Poids et Mesures.

McConnell, A.J. (1957): *Applications of Tensor Analysis*. Dover Publ. Inc., New York.

Merrigan, M.J., Swift, E.R., Wong, R.F., Saffel, J.T. (2002): A refinement to the World geodetic System 1984 reference frame. Proceedings of the 15th International Technical Meeting of the Satellite Division of The Institute of Navigation (ION GPS 2002), September 24 - 27, 2002, Portland, Oregon, pp.1519-1529.

Moritz, H. (1978): The definition of a geodetic datum. Proceedings of the Second International Symposium on Problems Related to the Redefinition of North American Geodetic Networks, 24-28 April 1978, Arlington, VA, pp.63-75, National Geodetic Survey, NOAA.

Moritz, H., Mueller, I.I. (1987): *Earth Rotation, Theory and Observation*. Ungar Publ. Co., New York.

Mueller, I.I. (1969): *Spherical and Practical Astronomy as Applied to Geodesy*. Frederick Ungar Publishing Co., New York.

NASA (1978): Directory of Station Locations, 5th ed., Computer Sciences Corp., Silver Spring, MD.

NGS (1986): Geodetic Glossary. National Geodetic Survey, National Oceanic and Atmospheric Administration (NOAA), Rockville, MD.

NGS (2008): The National Geodetic Survey Ten-Year Plan, Mission, Vision and Strategy, 2008-2018. http://www.ngs.noaa.gov/INFO/NGS10yearplan.pdf.

NGS (2010): Proceedings of the 2010 Federal Geospatial Summit on Improving the National Spatial Reference System. http://www.ngs.noaa.gov/2010Summit/2010FederalGeospatialSummitProceedings.pdf

Nelson, R.A., et al. (2001): The leap second – its history and possible future. *Metrologia*, 38:509-529.

NIMA (1997): Department of Defense World Geodetic System 1984, Its Definition and Relationships with Local Geodetic Systems. Technical report TR8350.2, third edition, National Imagery and Mapping Agency, Washington, D.C.

Petit, G., Luzum, B. (2010): IERS Conventions (2010). IERS Technical Note No.36, Verlag des Bundesamts für Kartographie und Geodäsie, Frankfurt am Main.

Pick, M., Picha, J., Vyskocil, V. (1973): *Theory of the Earth’s Gravity Field*. Elsevier Scientific Publ. Co., Amsterdam.

Rapp, R.H. (1991): Geometric geodesy, Part I. Lecture Notes; Department of Geodetic Science and Surveying, Ohio State University. http://hdl.handle.net/1811/24333

Rapp, R.H. (1992): Geometric Geodesy, Part II. Lecture Notes; Department of Geodetic Science and Surveying, Ohio State University. http://hdl.handle.net/1811/24409

Schwarz, C.R. (ed.) (1989): North American Datum 1983. NOAA Professional Paper NOS 2, national Geodetic Information Center, National Oceanic and Atmospheric Administration, Rockville, Maryland.

Schwarz, C.R., Wade, E.B. (1990): The North American Datum of 1983: Project methodology and execution. *Bulletin Gé odé sique*, 64:28-62.

Seeber, G. (2003): *Satellite Geodesy, Foundations, Methods, and Applications*. Walter DeGruyter, Berlin.

Seidelmann, P.K. (ed.) (1992): *Explanatory Supplement to the Astronomical Almanac*. Univ. Science Books, Mill Valley, CA.

Smart, W.M. (1977): *Textbook on Spherical Astronomy*. Cambridge University Press, Cambridge.

Snay, R.A. (2003): Introducing two spatial reference frames for regions of the Pacific Ocean. *Surv. Land Inf. Sci.*, 63(1):5–12.

Soler, T, Snay, R.A. (2000): Modern terrestrial reference systems, Part 2: The evolution of NAD 83. *The Professional Surveyor Magazine* 20(2):16-18.

Soler, T., Snay, R.A. (2004): Transforming Positions and Velocities between the International Terrestrial Reference Frame of 2000 and North American Datum of 1983. *Journal of Surveying Engineering*, 130(2):49-55. doi: 10.1061/(ASCE)0733-9453(2004)130:2(49).

Standish, E.M. (1981): Two differing definitions of the dynamical equinox and the mean obliquity. *Astron. Astrophys*., 101:L17-L18.

Stehlin, X., Wang, Q., Jeanneret, F., Rochat, P., Detoma, E. (2006): Galileo System Time Physical Generation. Proc. 38th Annual PTTI Meeting, 7-9 Dec. 2006 Washington, DC, 395-406.

Thomas, P.D. (1970): Spheroidal geodesics, reference systems and local geometry. U.S. Naval Oceanographic Office, SP-138, Washington, DC.

Torge, W. (1991): *Geodesy*, 2nd edition. W. deGruyter, Berlin.

Torge, W. (2001): *Geodesy*, 3rd edition. W. deGruyter, Berlin.

Vig, J.R. (1992): Introduction to Quartz Frequency Standards. Army Research Laboratory. http://www.oscilent.com/esupport/TechSupport/ReviewPapers/IntroQuartz/vigtoc.htm

Vinti, J.P. (1998): *Orbital and Celestial Mechanics*. AIAA, Reston, VA.

Wahr, J.M. (1985): Deformation induced by polar motion. *Journal of Geophysical Research*, 90(B11):9363 – 9368.

Whitaker, R. (2004): The Mapmaker's Wife: A True Tale of Love, Murder, and Survival in the Amazon. Random House, Inc., New York.

Wong, R.F., Rollins, C.M., Minter, C.F. (2012): Recent Updates to the WGS 84 Reference Frame. *Proceedings of the 25th International Technical Meeting of The Satellite Division of the Institute of Navigation (ION GNSS 2012),* Nashville, TN, September 2012, pp.1164-1172.

Woolard, E.W. (1953): Theory of the rotation of the Earth around its center of mass. Nautical Almanac Office, U.S. Naval Observatory, Washington, D.C.

Zanello, R., Mascarello, M., Galleani, L., Tavella, P., Detoma, E., Bellotti, A. (2007): The Galileo Precise Timing Facility. Proc. IEEE FCS 2007 and 21st EFTF, Geneva 29 May - 1 June 2007, 458 – 462.