

Defining the geoid by $W = W_0 + U_0$: Theory and practice of a modern height system

Roger Hipkin

Department of Geology & Geophysics, University of Edinburgh, West Mains Road, Edinburgh EH9 3JW, UK.

Abstract. This paper deals with two key questions: “How should we define and realise height?” and “What is the geoid?” The concept of height as a vertical distance arose from its contribution to determining absolute coordinates referred to the centre of the ellipsoid. Since satellite positioning, the concept is no longer necessary but scaling potential by gravity to get ‘height’ may still be conceptually convenient but should not be definitive. Now, a vertical reference system only needs a procedure whose output is the Earth’s potential on or above its surface and GPS geocentric coordinates are the input: the model $W(x,y,z)$ becomes the vertical reference frame. In this context, the roles of levelling, orthometric and normal heights, and the geoid are reviewed and I illustrate some general conclusions by an analysis of Stokes’ mass-condensation approach. Contrary to conventional analysis, Stokes’ method generate an estimate of the potential on or outside the Earth *that is hypothesis-free*. Apart from its use of the wrong tidal system, using a global model like EGM96 as a pre-processing tool is entirely consistent with Stokes-like local integration.

Defining the geoid to be identical to ‘mean sea level’ is no longer practically or physically sound: it confuses the crucial role geodesy can play in monitoring sea level rise, climate change and ocean circulation. Defining a vertical reference system by the potential at a monument cannot be global and, even if the relation with the sea surface is discarded, leads to inconsistencies because of tectonic activity. For a global and absolute vertical reference system, whose datum is defined by $W = W_0 \equiv U_0$, monumented datum points become only a ‘working hypothesis’ and cannot be definitive. I discuss the ideas of this paper as exemplified by the European Vertical Reference System and the difficulties of realisation by the European Vertical Reference Frame.

Keywords. Geoid, normal height, orthometric height, vertical reference system

1 Introduction: an historical critique of the impact of GPS on heighting.

Height as a geometrical distance. Half a century ago, geometrical reference systems were found by fitting an ellipsoid to the zero height surface constructed along long geodetic arcs. The position of the centre was *adopted* by *assigning* something equivalent to the deflection of the vertical at a datum point: Central Europe chose $\mathbf{x}_0 = 3''.36$ and $\mathbf{h}_0 = 1''.78$ at the Helmert Tower in Potsdam. Other values were assigned elsewhere: for example Meades Ranch, Kansas, for North America, or the triangulation tower at Pulkovo for the Soviet Union. Each defined a regional datum that was not connected with any other and neither its size, shape nor origin was based on physical Earth parameters. Chapters 8 and 9 of Heiskanen & Vening Meinesz (1958) provide an interesting commentary on the search, fifty years ago, for a global system for *geometrical* geodesy. It is parallel to our own present day attempt to convert the mosaic of regional height datums to a globally absolute vertical reference system.

In the pre-satellite world, height played a central role in constructing *geometrical* coordinates that were *absolute*. The radius vector \mathbf{r} from the centre of the reference ellipsoid to a point on the Earth’s surface was estimated as the sum of three other vectors: levelling gave the height \mathbf{H} of the external point; a Stokes-type integral of gravity gave the height \mathbf{N} of the levelling datum above the ellipsoid, and we computed the distance \mathbf{R} from the origin to a point on the ellipsoid’s surface.

$$\mathbf{r} = \mathbf{H} + \mathbf{N} + \mathbf{R} \quad (1)$$

In reality, levelling does not give a geometrical height at all: it gives a difference in potential. Similarly, Stokes-type integration does not really give the ‘height’ of the ‘geoid’ above the ellipsoid but, again, a difference in potential. Any procedure to transform levelling and the product of Stokes’ integration into some geometrical distance called ‘height’ involves purely conventional choices. In an

absolute sense, we should paraphrase the title of Smith's (1998) tutorial on EGM96 to become "*There is no such thing as **the** geoid.*" It is equally true that *there is no such thing as **the** orthometric height.* These are artificial and conventional constructs. It was only because we needed to determine the coordinate \mathbf{r} , where 'height' was required as a geometrical distance, that a transformation to convert 'height' to distance was necessary.

The impact of artificial satellites. The 1967 Geodetic Reference System arose because artificial satellites enabled us to sample the whole Earth's gravity field properly and determine its mean properties in a truly representative way. Long geodetic arcs had just been biased local samples! Also, an artificial satellite, unlike the Moon, has a mass that is negligible compared with the Earth. As a result, Newton's mechanics put the focus of its elliptical orbit at the centre of mass of the Earth and direct geometry gives an absolute reference for the origin for the coordinate system. The other key component was the new possibility of rigour in using the Pizetti-Somigliana (P-S) model for a reference Earth (Heiskanen & Moritz, 1967). The P-S theory allows the ellipsoidal reference surface and distance coordinates to be matched absolutely to the real Earth. The model generates a lumped potential U - whose derivative gives the sum of gravity and centrifugal forces - called the reference potential. The *defining* property of the P-S model is that one (and only one!) of its equipotential surfaces is an *exact* ellipsoid, identified by $U = U_0$. Apart from its orientation, its gravitational and geometrical properties are then completely specified by four parameters: $\{GM, a, J_2, \mathbf{w}\}$.

Knowing the rotation rate \mathbf{w} with sufficient accuracy for applications to physical geodesy has always been trivial but, now, being able to measure the length of the semi-major axis and period of an artificial satellite orbit gave the mass of the Earth GM . The backwards motion of the plane of the orbit in space, or, alternatively, the forward motion around the orbit of the line of closest approach to the Earth, determined the Earth's gravitational flattening coefficient, J_2 . These three parameters are real physical properties of the Earth. In contrast, the equatorial radius of the ellipsoid, a , just defines which of the family of equipotential surfaces of the P-S Earth model we choose to call the reference ellipsoid: we are free to adopt its value by convention.

The new ability to define an absolute geodetic reference system remained essentially geometric: only geometry and kinematics are needed to measure the parameters of the model Earth. In principle, the process does not involve any knowledge of how the real Earth's gravity field W differs from that of the gravity field U describing the P-S model Earth. (In practice, it is necessary to solve for several thousand spherical harmonic coefficients describing W but, for the geometrical reference system, these are only free parameters whose function is to avoid biasing the estimates of the P-S model parameters GM and J_2 .)

GPS and ellipsoidal heights. Since the early 1980s, satellite positioning has added a completely novel feature: GPS provides the vector \mathbf{r} directly. Consequently, it has been supposed that, because levelling is error-prone and expensive, we should re-arrange equation (1) and use it to determine \mathbf{H}

$$\mathbf{H} = \mathbf{r} - \mathbf{N} - \mathbf{R} \quad (2)$$

However, the geodetic community has been very slow to understand that this new ability results in a much more philosophically radical revolution in geodesy, completely changing the relation between its geometrical and physical disciplines: *we no longer need conventional height to tell us where we are.* Now, the *only* application of height is to map level surfaces in space. What is the shape of the level surface at P and how can I describe it in terms of the coordinates given by GPS? Is the potential energy at point P more or less than point Q and, if so, by how much?

With this more limited role, the conventional or approximate transformations from potential to distance - transformations that were obligatory when we sought to use equation (1) to find \mathbf{r} - are no longer needed. GPS gives us the absolute geocentric coordinates $\{x_P, y_P, z_P\}$ for any point P on or above the Earth's surface. All we really need for a vertical reference system is a procedure for which $W_P = W(x_P, y_P, z_P)$ is the output when $\{x_P, y_P, z_P\}$ is the input. Using the alternative geometric coordinates $\{h, \mathbf{j}, \mathbf{I}\}$, equivalent to $\{x, y, z\}$, we can then trace the ellipsoidal height of the level surface through P by solving the equation $W(h, \mathbf{j}, \mathbf{I}) = W_P$ for $h(\mathbf{j}, \mathbf{I})$. Similarly, if we have their geometrical coordinates from GPS, potential energy differences between points P and Q come directly from $W_P - W_Q = W(x_P, y_P, z_P) - W(x_Q, y_Q, z_Q)$.

Later in the paper, I shall examine some practical consequences of this unorthodox picture in which we no longer need to think of height as a distance

measured in metres. First, it is useful to be reminded what distance really is in modern geometrical geodesy.

Geometrical geodesy and modern metrology supply us with *distance* as a *time-interval* scaled by the *velocity of light*. (Alternative techniques based on phase comparison or counting wavelengths are entirely equivalent.) The trajectory of distance is then controlled by the *rectilinear propagation* of light, so that differences in position are found along straight lines. Coordinates are now propagated by trilateration, with the satellite at one vertex. Triangulation, with its practical reference to a levelled surface and therefore the gravity field, is obsolescent for control networks.

Perhaps because of rectilinear propagation, the community seems to have adopted the trajectory of ellipsoidal height to be a straight line, tangential to the normal of the reference surface, even though other choices might seem more consistent with the theory of the P-S model Earth on which the reference system is based. For example, we could have chosen it along the normal plumb line or along the hyperbolic coordinate line of the Jacobi ellipsoidal coordinates in which the theory was developed.

The effects of variation of the speed of light in different media and the bending of rays when the refractive index changes mean that the subject is not free of corrections but the principles that distance is found from time scaled by light speed and has a trajectory defined by rectilinear propagation remain fundamental. In the following section I look at the very different sort of distance provided by physical geodesy: it is essentially potential energy scaled by gravity. Choices about what trajectory it follows and what scaling to use are less clear cut.

2 ‘Orthometric’ heights and the geoid.

Criteria for a precise height system. Being able to determine a ‘true measure’ of height (this is what ‘orthometric’ means) demands definitions of what properties height must have. At least two conditions are *necessary*:

- (i) Height must be single-valued
- (ii) A surface of constant height must also be a level (equipotential) surface.

A third condition would be *convenient* but is much less fundamental and might be sacrificed if necessary:

- (iii) Height should be a geometrical distance and therefore measured in metres.

It will emerge that all three of these conditions are not satisfied simultaneously for any of the height systems currently in use.

Consider measurements of gravity $\mathbf{g} = -\nabla W$ and distance increments $d\ell$ measured along a path from P to Q on or above the Earth’s surface.

$$\int_P^Q \mathbf{g} \cdot d\ell = - \int_P^Q \nabla W \cdot d\ell = W_P - W_Q \quad (3)$$

Stokes’ Theorem requires

$$\oint \mathbf{F} \cdot d\ell = 0$$

for any vector \mathbf{F} derived from a scalar potential by $\mathbf{F} = \nabla \psi$, so the result of equation (3) is independent of path: W is single-valued.

If the point P is chosen as datum and assigned the potential $W = W_0$, then the ‘geopotential number’ c_Q

$$c_Q = W'_0 - W_Q \quad (4)$$

has all the *necessary* properties of a height system: it identifies whether two points lie on the same level surface and, if not, what forces will result. However, it is not measured in metres.

I shall now deal with two special cases of equation (3): the first deals with levelling; the second with the ‘geoid’.

Levelling. Choose the path between P and Q as alternating straight line segments. The first kind of segment is $Ds = Ds_f - Ds_b$ and lies along the line of sight from the levelled telescope to the fore- and back-sight staves; the second kind is Dh and lies along the staff from the present fore-sight reading to the next back-site reading. Because the path length Ds is tangential to the equipotential surface at the level, to an excellent approximation $\mathbf{g} \cdot d\ell = 0$ - making reasonable assumptions about gravity field curvature and torsion suggests $\sim 10^{-12}$. The increment $d\ell$ is measured along the staff held vertically, so $\mathbf{g} \cdot d\ell = |\mathbf{g}| Dh$. The need to project the measured increment Dh onto the vertical at the levelling telescope makes it systematically too large and systematically different in the east-west direction from in the north-south direction; however, the resulting scale error is only $\sim 10^{-10}$. Thus the integral reduces to the usual summation given for geopotential levelling

$$W'_0 - W_Q = \sum_i g_i Dh_i \quad (5)$$

Although practical errors are often significant, this standard result shows that, apart from the datum constant W_0 , levelling and gravity combine to provide a very precise height system indeed.

However, this system only satisfies the *necessary* conditions (i) and (ii): it does not give height as a geometrical distance, condition (iii).

Obviously, if $\sum_i g_i Dh_i$ provides a single-valued unique measure of height, the more obvious sum $\sum_i Dh_i$ cannot do so. Knowledge of the gravity field is essential for a height system.

Normal heights. In the next section, I consider a path of integration for equation (3) that goes inside the Earth and so passes from the real world into a land of myth and make-believe. However, first, Molodensky's normal height must be rescued from any such taint. The normal height is simply a hypothesis-free transformation of the geopotential number into a distance measured in metres. A normal height scales the potential with the model version of gravity generated by the P-S Earth.

$$H^n_Q = \frac{c_Q}{\bar{g}_Q} \quad (6)$$

(For historical reasons, the mean value of normal gravity

$$\bar{g}_Q = \frac{1}{H_Q^n} \int_0^{H_Q^n} g(h) dh \quad (7)$$

is used in the definition.) Normal gravity is a simple analytical function of position depending only on parameters already incorporated in the geometrical reference field.

No information is lost in constructing normal heights from geopotential numbers and no additional information is needed. Any hypothesis about a subterranean surface from which the normal height is measured is not needed for the definition and so does not detract from the evaluation or use of normal heights.

Normal heights satisfy conditions (i) and (iii) of the ideal height system but a surface of constant height is not level: suppose that the points P and Q have the same geopotential number but are at different latitude:

$$c_P = c_Q \text{ but } \bar{g}_P \neq \bar{g}_Q \text{ so } H_P^n \neq H_Q^n.$$

However, the surface $H_Q^n = H_P^n$ lies a known and easily calculable height above the level surface through P

$$dH_Q^n = \left[\frac{\bar{g}_P(H_P^n) - \bar{g}_Q(H_Q^n)}{g_Q(H_Q^n)} \right] H_Q^n \quad (8)$$

(Note that the point value rather than the mean value of g is used in the denominator.)

This result illustrates a more general feature: none of the definitions of height in current use satisfies all three conditions simultaneously (see footnote about Helmert's orthometric heights on page 172 of Heiskanen & Moritz (1967); I consider dynamic heights as obsolete).

The 'geoid' and 'orthometric' heights. We do not *know* what the density inside the Earth is – all inversions are non-unique – so we can never be certain about what subterranean gravity really is, or where equipotential surface are, or how far apart they are. If we choose an integration path for equation (3) that goes inside the Earth, we must *adopt* some model for the gravitational potential there, say $W = W^*$. The 'geoid' then becomes the surface $W^* = W^*_0$. The only purpose for such a procedure is to be able to predict the difference in the real potential W between points *on or outside the Earth's surface*. Inventing a suitable W^* is just an artifice used in this procedure, for which the only requirement is that it gives the correct value for W externally. There are many varieties of W^* in the literature and, of those that predict W correctly, none is more 'right' than any other. Here, I explore equation (3) for an interior path to illustrate some general features, without being too specific about the choice of W^* .

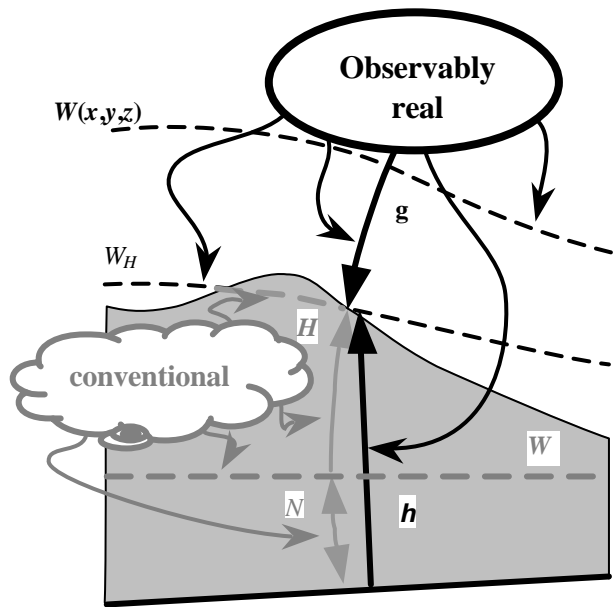


Fig.1: The real world and the grey underworld.

I follow a classical route of splitting the *external* potential W into three parts:

$$W = U + T^* + dT \quad (9)$$

Internally, I put

$$W^* = U + T^* \quad (10)$$

so U and T^* are described by the same expressions inside and outside the Earth but dT represents some property of the external field that we choose not to reproduce internally. Thus, dT changes discontinuously on passing through the topographic surface. (This description is consistent with the procedures of Stokes and Molodensky, but not those of Poincare, Prey and Helmert, for which equation (10) also needs an additional term dT_{int} , so the increment of equation (9) has to be re-labelled as dT_{ext} . However the following analysis needs only rather straightforward modification for them.)

We now evaluate equation (3) again between the surface points P and Q but for a path that lies inside the Earth, dealing with the particular case where P is the datum point when $W_P = W_0$. Passing down through the surface from P , the potential is decreased discontinuously by the value of dT at P :

$$W^* = W'_0 - dT_P \quad (11)$$

Now we start the integration path and follow the equipotential of W^* as far as the point Q' , lying on the \mathbf{g}^* plumb-line beneath Q . For this part of the path, $\mathbf{g}^* \cdot d\ell = 0$, so W^* remains the same. The interior integration is completed by

$$-\int_P^Q \mathbf{g}^* \cdot d\ell = W_Q - W_{Q'} = -\left| \bar{g}^* \right| H_Q^* \quad (12)$$

Adding the effect of passing through the topographic surface to Q gives the geopotential number there as

$$c_Q = W'_0 - W_Q = \left| \bar{g}^* \right| H_Q^* + dT_0 - dT_Q \quad (13)$$

Although H_Q^* is a 'true' (that is orthometric) height in a geometrical sense, calculable corrections $dT_0 - dT_Q$ are needed and there is nothing *unique about this truth*: H_Q^* depends entirely on our arbitrary and hitherto undefined choice of W^* and therefore $|\bar{g}^*|$. The same lack of uniqueness applies to the form of the interior equipotential surface that we might wish to call the 'geoid'.

Stokes' mass-condensation. So far, not much has been said about how to choose the interior potential W^* . The problem would be simple if we could use exactly the same mathematical formulae to describe W inside and outside the Earth (a procedure called analytical continuation) but this is not possible. In

practice the problem is not really that our description is based on LaPlace's Equation, which does not hold inside masses – we commonly use it successfully above the Earth's solid surface but inside the atmosphere. It is that the Earth's surface can be topographically rough and there is a large density contrast between air and rock. Together, these generate a gravity field with significant short wavelength power and this destabilises analytical continuation *downwards*. 'Destabilisation' means that a small change in the description at the surface – a new measurement of gravity or an error in an existing value – makes a big change in the estimate below the surface.

Stokes' solution was to change the Earth by compressing a geometrical model for the topographic masses to a surface density on the zero-height surface (the 'geoid'). Condensation is a two stage process – first removing the masses altogether than putting the same mass back on the geoid. The key point of this procedure (which he fully appreciated – see §33, Stokes, 1849) is that modelling and removing the topographic masses leaves the (complete Bouguer) gravity anomaly very much smoother so that it can be reliably interpolated. With this smoothness, downward continuation from the surface to the geoid involves a very small (for him negligible) correction. Once Bouguer anomalies are determined at zero height, exactly the same mass is put back as a surface density on the *underside* of the geoid. With gravity \mathbf{g} now determined on the 'geoid', a Stokes-like integral transforms it to the potential on the ellipsoid,

$$T^*(0, \mathbf{j}, I) \quad [\equiv T^*(0)].$$

We now see the significance of the terms T^* and dT . By condensing the topography, Stokes makes the region between the surface and the geoid mass-free. T^* is an analytical continuation of the external potential inside the Earth, having first removed the small component dT that would have made this procedure invalid. The correction dT is the difference in potential of the topographic masses *in situ* and condensed on the geoid, evaluated on or above the Earth's surface.

There is again a geometrical interpretation of Stokes' method. Consider the height change N^* and the change in the reference potential U between the 'geoid' and the ellipsoid. Making the approximation

$$-g(0) \approx \frac{U(N^*) - U(0)}{N^*} \quad (14)$$

gives

$$N^* \approx \frac{T^*(0) + U_0 - W'_0 + dT_0}{g(0)} \quad (15)$$

Thus the geoid is simply a way of visualising the way the anomalous potential T^* varies over the ellipsoid.

The geoid is a convenient stage (but not a necessary one) in the process of estimating W outside the Earth. Now that Stokes' integral gives us an analytical function describing T^* on the ellipsoid, the formal procedure of upward continuation by the ellipsoidal height h_p gives

$$\begin{aligned} T^*(h_p) &= T^*(0) + T^*(h_p) - T^*(0) \\ &= T^*(0) + DT^*(h_p) \end{aligned} \quad (16)$$

(Finding DT^* by upward continuation is much better than a Taylor series because $T^*(0)$ is evaluated immediately next to the condensed topography and so contains the short wavelength features of topographic roughness: a reasonable result with a Taylor series would require high order terms.) In contrast, the reference potential is smooth and only two or possibly three terms of a Taylor series suffice

$$\begin{aligned} U(h_p) &= U(0) + \frac{\partial U}{\partial h} h_p + \frac{1}{2} \frac{\partial^2 U}{\partial h^2} h_p^2 \dots \\ &= U_0 - \bar{g}(h_p) h_p \end{aligned} \quad (17)$$

We now have the relation between the normal height as given by levelling and the ellipsoidal height given by GPS:

$$\begin{aligned} H_p^n &= \frac{W'_0 - W(h_p)}{\bar{g}(h_p)} \\ &= h_p - N^* + \frac{\bar{g}(h_p) - g(0)}{\bar{g}(h_p)} N^* \\ &\quad + \frac{dT_0 - dT_p - DT^*(h_p)}{\bar{g}(h_p)} \end{aligned} \quad (18)$$

This is a rather general result: the sort of height generated by levelling is approximately the difference in the ellipsoidal height and the 'geoid' height but corrections are needed to make the result exact. The fact that here I have developed it in terms of a Stokes' geoid and normal heights was deliberate – most procedures and their components are equivalent. The folklore that a Stokes geoid must combine with a Helmert or Poincaré-Prey 'orthonormal' height or a normal height with a Molodensky 'quasi-geoid' is just that: in all cases, there are small additional corrections and it is

arbitrary how we choose to partition them between 'height' and 'geoid'.

The mythology that a Stokes geoid lies embedded in the topographic masses or that we need to know what real equipotential surfaces are inside the Earth is part of the accumulated baggage collected by a subject that is 150 years old. It is not commonly understood that, in principle, *Stokes' method is hypothesis free* and introduces no conceptual error in its estimate of potential outside the Earth. Neither the assigned density nor some aspects of the shape of the model need match the real Earth. Provided that the *gravity effects* of the topography, both *in situ* and condensed, are computed and removed from observations consistently and use the same terrain model as the computation which restores the difference in *potential* between the *in situ* and condensed topography, the result is independent of the model. If any model is perfectly removed and perfectly restored and the Earth is unchanged at the end, there can be no error.

However, the key issue with Stokes' condensation method is whether removing a model for the geometrical shape of the topography that has some fixed, conventionally assigned density is enough to make downward continuation of the Bouguer anomaly stable. Noting that variations in both rock density and topographic height are bounded, I discussed (Hipkin, 1988) stability criteria and introduced the concept of 'normal density' – still a constant density but one chosen to optimise stability. In practice, it is very rare indeed than the conventional value of 2.67 kg m^{-3} is inadequate.

Because of his historical priority and the philosophically satisfactory nature of the result, I suggested (Hipkin, 1988) that Stokes' 'geoid' be called just "the geoid" and we abandon the esoteric jargon "Stokes' condensation anomaly co-geoid". Newton's concept involving a surface only accessible by digging wells and canals, which lies behind the Poincaré-Prey reduction as well as Helmert's approximation to it, seems outdated and impractical. While I am not aware of any geodesist reacting to this proposal, the other issue on how to define the geoid is what value to give W_0 : this has attracted much attention and I discuss it in section 4.

However, in the next section I stay with the problem of how to compute $W(x,y,z)$. Apart from removing and restoring the P-S model Earth and a model for the topographic masses, all modern work also removes and restores a global gravity model. The next section deals with its role in the process.

3 The global gravity model.

A global gravity model, for example EGM96, is now almost universally used as part of the process of finding the Earth's potential from local observations of gravity anomalies. Like Stokes' treatment for the topographic masses, the global model is first represented as gravity g_{GM} and removed from observations, then represented as potential T_{GM} and restored. The most important but often overlooked feature of using a global model is that the residual gravity $g^* - g_{GM}$ is small. The consequence is that the rather difficult stage of numerical integration involved in Stokes transformation from gravity to potential accounts for only small corrections. It may therefore be rather grossly approximated by computationally simpler algorithms: a spherical approximation is always sufficient and even a 'flat-Earth' planar FFT is often good enough.

Self-consistency. In principle, a remove-restore procedure with the global model will be error-free provided that the model free-air anomaly is exactly consistent with the restored potential. In reality, the most common algorithms (essentially based on equation 2.155 of Heiskanen & Moritz, 1967) are *not* self-consistent and the errors amount to several decimetres (Hipkin, 2002b). In place of the standard formula for computing the free-air anomaly from spherical harmonic coefficients of the potential, I derived a formula evaluated as the difference between observed gravity on the geoid and normal gravity on the ellipsoid, $|g(N)| - |g(0)|$. This expression is not an ellipsoidal approximation, developed in powers of the Earth's eccentricity, but an exact closed formula for the part of the gravity difference that is linear in T . The 'free-air anomaly on the geoid' is given by equation (51) of Hipkin (2002b). However, despite a simplified description in their publications, several users *do* use calculations that are correctly self-consistent: for example, EGG97 actually uses an unpublished procedure derived by H.-G. Wenzel (Denker, pers. comm., 2001).

Here, I look at two other problems: the first, that of evaluating EGM96 within the topography and the second a mismatch of the tidal corrections between gravity observations and the global gravity model.

EGM96 inside the topography. Smith (1998) discusses the validity of using the global model beneath the Earth's surface. I believe his discussion arises from the misconceptions over the meaning of the geoid and orthometric heights. The global model

is a description of the Earth's gravity field using harmonic functions – a description derived assuming a mass-free environment. However, we are not attempting to find the *real* gravity field inside the Earth – this is unknowable. The criterion for validity depends only on whether analytical continuation is stable. For a global gravity model, analytical continuation into the topographic masses can *never* become unstable because the description is inherently band-limited. It uses a *finite* series of harmonics whose minimum equivalent wavelength – $(40000/n)$ km for $n = 360$ – is much greater than the range of topographic height. Thus, provided that the EGM96 potential is only considered as a convenient intermediate stage towards estimating W outside the Earth, no problem exists. The computation of Hipkin's (2002b) equation (51) is evaluated 'in the free air' and is therefore entirely consistent with the Stokes' anomaly $|g^*(N)| - |g(0)|$, provided the latter is part of a process that restores the dT before evaluating W on or above the Earth's surface.

Tidal systems. This subject seems to have caused more confusion than almost any other over the past 40 years of physical geodesy. In general, unless the model is part of a rigorous remove-restore process, when the result is independent of the model, no part of the gravity field should be modelled and removed unless the model is so well determined that no possibility of error arises. For strictly periodic effects, a long-term average might be sufficient, even if the mechanism is imperfectly known, so it is only the so-called 'permanent tide' that causes a problem. Here, the community concerned with the geometric reference system seem to be having a re-run of the mistakes made by the gravity community twenty years ago. There was the 'do nothing' approach of the Honkasalo correction in 1964; the 1979 IAG resolution in Canberra to eliminate the whole of the permanent tide and then its reversal by the 1983 IAG resolution in Hamberg. The discussion is not helped by the terminology referring to 'tide-free', 'zero-tide' and 'mean-tide' systems, where the adjectives are far from self-explanatory.

On the Earth's surface a measurement of gravity is influenced in three ways by the permanent tide. First, there is the direct tide-generating force. For the gravity meter, this is the main part of the Earth tide. It is a potential due to *external* masses (the Sun and Moon), varying like r^2 not r^{-3} , and so must be removed. It is known exactly. For a satellite, the direct effect is not seen as part of an Earth-tide at all, for which the sources are *inside* the orbit, but an

external force resulting in a simple three-body disturbing function.

Secondly there are direct and indirect effects of a redistribution of the Earth's mass in response to tidal forces. The indirect effect is a consequence of displacing the Earth's surface (where the measurement is made) through the normal gravity field and has no equivalent for the satellite. The direct effect is commonly described by a Love number k . For a solid, perfectly elastic body, the potential dV produced by the tidal redistribution of the Earth's mass is a linear multiple of the potential V describing the tide-generating forces.

$$dV = kV \quad (19)$$

Within the seismic frequency band, say 10^3 Hz to 10^6 Hz, extended to 10^6 Hz by Earth-tide observations, the solid Earth fits Love's elastic theory and gives a well determined value of $k \sim 0.30$. However, for very long periods, the 'solid' Earth behaves more like a perfect fluid than an elastic solid: the main part of the Equatorial bulge is seen as a fluid response to centrifugal force. The secular response of a perfect fluid would give $k_{\text{fluid}} \sim 0.96$. However, we know that some small part of the bulge is not a hydrostatic response to rotation but due to geodynamic effects like mantle convection. Although the hydrostatic approximation is not bad, the flattening of an equivalent fluid is not exactly equal to the observed one: $f_{\text{hydrostatic}} \approx 1/299.627$ (Nakiboglu, 1982) compared with $f_{\text{observed}} \approx 1/298.254\dots$. Apart from a different value for k , there is also a time delay – the *viscous* response of the Earth to post-glacial rebound and to its decreasing spin rate makes its shape a function of a past forces, not current ones. Observations aiming to determine the effective value of k_{secular} are further perturbed.

The point of this discussion is that *we do not know* what value to give k for the permanent tide. The seismic value $k \approx 0.30$ is definitely wrong, but the fluid alternative is uncertain. We should therefore not remove the part of the permanent tide due to redistributed Earth mass at all. (In any case, removing it would not pass Popper's test for what science is: no experiment can be conceived that would allow us to distinguish between those components of the Earth's mass that result as a response to the permanent tide from those with another cause!)

The gravity community has followed the 1983 IAG Resolution whereas EGM96 and major parts of the geometrical community have not. In a solution

for W , the global model needs to be corrected for this error before being combined with surface gravity data.

4 The W_0 problem: which surface is the geoid?

Assigning a value to the constant W_0 means we choose which of the family of equipotential surfaces is assigned zero height. Adding an arbitrary constant to potential energy has no physical consequences whatsoever – only force defined by the gradient of potential is observable – so it might be supposed that an argument over the value of W_0 is meaningless. However, we already have a universal convention for defining the arbitrary constant – we choose the gravitational potential to be zero at infinite distance – so the choice of W_0 *does* correspond to a unique way of defining which surface has zero height.

W_0 is the potential at a datum monument. So far, I have used the dash was to indicate the potential at an arbitrary datum monument; this then corresponds to the W_0 introduced in Molodensky's theory of height. Even if our model for $W(x,y,z)$ was arbitrarily accurate, so that the chosen value of W_0 could be used world-wide and not just in those regions where a levelling connections was possible, there would still be a problem. To some extent, everywhere is affected by tectonic activity, so the real height of the datum could change: there is nothing physically absolute with this approach.

At the present time, our model for $W(x,y,z)$ is not good enough to give us the real potential at the datum point exactly. If we assign a specific but slightly incorrect value to W_0 , and then apply it globally, we may be violating the condition $W_g(\infty) \rightarrow 0$.

Alternatively, the value might have only a regional application, with regions that are not physically connected having their own datum; this is essential the present situation. For our current vertical reference systems, we are in a position analogous to the 50 year old problem of having separate reference ellipsoids defined by Meades Ranch, Potsdam, Pulkovo, Dehra Dun,

W_0 is the potential of mean sea level. The surface of a fluid in hydrostatic equilibrium is a gravitational equipotential surface, so the nineteenth century approach to establishing a global vertical datum supposed that mean sea level could bridge regions not connectable by levelling. The 'geoid' was formalised into the equipotential best fitting mean

sea level and, for more than a century, the concepts of mean sea level, the geoid and the levelling datum were used synonymously. Nowadays, when observations are much more precise, their differences are distinguishable and present practice leads to confusion.

It is now essential that we no longer associate mean sea level with any aspect of defining the geoid. First, the mean surface of the sea is not an equipotential – wind-stress results in permanent sea surface topography with a range of about two metres; secondly, thermal expansion of the oceans due to climate change (it is now believed that most of the 1-2 mm/year global sea level rise is due to thermal expansion) would result in the mean sea surface rising with little corresponding displacement of the equipotential surface.

Two of the key scientific contributions that geodesy has to make in the coming decades are: (i) to determine the global rate of sea level rise because of its relation to climate change, and (ii) to determine sea surface topography, and thence ocean circulation, though the combination of a gravimetric geoid and satellite sea surface altimetry. If we insist on having a reference system chosen such that the mean height of the sea is defined to be zero, the logic of geodetic contributions to oceanography and climate change becomes muddled.

At the present time, models for ocean circulation give more reliable estimates for the height of the mean sea surface above the equipotential (sea surface topography) than do space geodetic methods. Although the new gravity field satellite missions might well reverse this position, we can use current oceanographic results to understand how best to use geodesy in the future. For example, on a global scale, the OCCAM model (Webb *et al*, 1997) has a resolution of $1/4^\circ$. It is driven by three-day averaged surface winds, themselves the product of assimilating meteorological observations into a global atmospheric circulation model. The evolution of the ocean circulation is then verified by hydrological observations and, by assimilating them, corrected as it evolves. On a more regional scale, a shelf model like POLCOMS (Holt *et al*, 2001) can be driven by three-hour averaged winds and achieve a spatial resolution of $1/6^\circ - 1/9^\circ$. Although tide gauges have long shown that *monthly mean* sea surface heights can have persistent excursions of several decimetres due to non-random wind stress, a high resolution shelf model can confirm several centimetres of difference from one year to another in the *annual*

mean sea surface topography. This is no longer potentially a local effect of the tide gauge harbour but an whole sea phenomenon. Thus, it is not even clear how the mean sea surface at one location can be translated into a reliable height datum, never mind an extension of the process world-wide!

W_0 is equal to the reference potential on the ellipsoid, that is $W_0 = U_0$. Defining the geoid by $W = W_0 \equiv U_0$ binds the geometrical and physical reference systems together inseparably. I have made this choice for Edinburgh geoid computations for nearly 20 years, but mostly implicitly because to me it seemed self-evidently correct. Now that I have had four years to think about the problem in a more informed way (I am indebted to the excellent tutorial on EGM96 (Smith, 1988) for exposing me to some different choices), I remain convinced that my earlier view really is 'self-evidently correct'!

The great advantage of defining the geoid by $W = W_0 \equiv U_0$ is that we have a vertical reference system that is truly global, 'absolute' (in the restricted sense discussed below) and one that needs no further free parameters: the Geodetic Reference System still has only four parameters, instead of the five that would result from adding W_0 to $\{GM, a, J_2, \mathbf{w}\}$. A potential disadvantage is the datum point for the vertical reference system becomes *virtual*: unless we have infinitely perfect knowledge of the Earth's gravity field, we cannot create a physical monument and be certain that its potential is $W = W_0$.

(Note that discussion about whether to adopt a or U_0 as one of the four geometrical parameters is misplaced: W_0 is only a property of the real Earth, and so more fundamental than a , if we define the geoid by the mean sea surface. I believe that this definition is no longer tenable, so a and U_0 are interchangeable and both are equally conventional. The equatorial radius a is much simpler to visualise.)

5 Practical interim height systems and the problem of monuments

How can we make practical use of an *absolute* vertical reference system with a *virtual datum* and no *defining monuments*?

What does absolute mean? The P-S theory of the level ellipsoid makes the reference potential on the ellipsoid an analytical function of the adopted parameters: $U = U_0 \equiv U_0(GM, a, J_2, \mathbf{w})$. Because the equatorial radius of the ellipsoid is given a conventional value rather than being a physical property of the real Earth, defining a vertical

reference system by $W_0 \equiv U_0$ is not 'absolute' in the same physical sense as the absolute origin of the geometrical reference system. There, the origin lies at the Earth's centre of mass. However, in other respects, the analogy with the geometrical reference system is good: the potential on the geoid becomes a function of the geodetic reference system parameters and is therefore 'absolute' in the sense of being error-free and world-wide.

How can we work with a virtual datum? For both a geometrical reference system where the origin is *defined* to be at the centre of mass and a vertical reference system *defined* by $W_0 \equiv U_0$, the datum is inaccessible (and therefore virtual). We use universal phenomena and a collection of physical but not defining sites to get practical quantities.

The concept of a 'virtual datum' becomes more familiar after looking at how geometrical coordinates are determined in practice. After adopting provisional parameters defining a frame of reference, geometrical space geodesy gives the provisional geocentric coordinates of tracking stations as observables. Remembering that geodetic latitude is *defined* by the distance from the Equator over the surface of a chosen ellipsoid, the free-parameter a relates inter-site distances on the Earth to the dynamically determined distances of the satellite orbit. In order to make all observations as consistent as possible with the P-S model Earth, we find 'better' P-S model parameters $\{GM, a, J_2, w\}$ **and better coordinates of the tracking stations**. Now the analogy with a vertical reference system gives useful insight: once the tracking stations are assigned "coordinates in the International Terrestrial Reference Frame" these values become the 'working hypothesis' for 'true' coordinates. For example, we use ITRF values as the *fixed* coordinates of fiducial stations' in a local GPS net. Using the 'working hypothesis' as if it was 'true' is philosophically unsatisfactory but is a necessary consequence of having a reference system with a virtual datum. It has seemed to cause no difficulty for ITRF but only because the errors in the working standards have been negligible. This identifies the key difference between a geometrical reference system like ITRF and our current attempts to create an absolute vertical reference system: the provisionally assigned potential at our working reference sites is not known well enough. I return to this problem.

What about sea level? We can look at the choice of the free parameter a in another way. As before, we

treat the provisional geocentric coordinates of tracking stations as observables but add a second observable at each site: its estimated 'height above sea level'. Combining the two gives a collection of points 'at sea level' whose provisional geocentric radius is known. Real earth properties $\{J_2, GM, w\}$ impose a pre-determined flattening, so we get the equatorial radius a by a geometric fit of a constrained ellipsoid to the zero-height surface. Although the notion that the reference ellipsoid as a representation of mean sea level is several centuries old and is only implicit in the procedure, it is not subsequently important: whatever means is used for getting estimating the value of a , this parameter becomes purely an adopted convention. Any putative relation with the sea surface or the uses of error-prone levelling referred to different datums becomes irrelevant. Nevertheless, it is not coincidental that the GRS80 value of U_0 corresponds to a surface $W = W_0 = U_0$ that only differs by a few decimetres from the current best fit to mean sea surface (Hipkin, 2002a).

The problem of monuments. An integral part of a vertical reference system must be the model $W(x,y,z)$, in which is incorporated the global datum $W_0 = U_0$. Suppose that we knew W perfectly: we could simply insert the latitude, longitude and ellipsoidal height of Amsterdam, Kronstadt, Marseilles or Newlyn and get values for the real Earth's potential at each. These values would then be perfectly consistent with a global datum.

Because the European Vertical Reference System (EVRS) (Ihde & Augath, 2001) has adopted the features of a globally absolute system matching the philosophy I have described, I can illustrate the problem of monuments and practical heighting systems with an example. The attempt to realise EVRS is called the European Vertical Reference Frame 2000 (EVRF2000). The former has a truly global datum defined by $W_0 = (U_0)_{\text{GRS80}}$; the latter adopts a potential value for the benchmark associated with the Normaal Amsterdam Peil while simultaneously treating the mean sea surface, corrected for sea surface topography, as definitive. For the reasons discussed above, these two axioms are inconsistent with each other and with EVRS. We should treat EVRF2000 only as a working hypothesis.

It has become clear that levelling can have very significant systematic errors (for two extreme examples, see Forsberg *et al* (this volume) for Britain; or Kasser (1983) for France). For simplicity, I shall imagine that we are free to devise a height

system without reference to the past so my vision for a Vertical Reference Frame will not involve levelling.

Suppose instead that 1997 European Gravimetric Geoid (EGG97) (Denker et al, 1997) is *adopted* as the model for W . Combined with GPS coordinates, it becomes the practical way realising of height. Geopotential numbers $c = (W)_{\text{EGG97}} - W_0$ would be described as in the frame of EGG97 and GRS80. We could then determine the offset from the new system of historical levelling referred to NAP by determining $(W_{\text{NAP}})_{\text{EGG97}}$: $Dc_{\text{NAP}} = (W_{\text{NAP}})_{\text{EGG97}} - W_0$. The advantage of this procedure is that it gives the offset a non-zero value, thereby reminding people that the monument identifying the NAP datum is a 'working hypothesis'. The value of the offset will change as our knowledge improves.

To me it seems inevitable that, in the near future, we shall adopt a vertical reference system based on adopting a gravity model and one that incorporates $W = W_0 \equiv U_0$ to define its datum. Some could argue that there is not yet a good enough case to move to such a radical new way of thinking, when, at the moment, the new realisation has doubtful accuracy compared with the present system. It is probably that EGG97 is differentially good to a few centimetres, so is actually not worse than the precision of UELN 95/98 (fig 2, Ihde & Augath, 2001).

However, the available will change rapidly following analysis of the gravity field satellites CHAMP, GRACE and GOCE: apart from continuing improvements in surface gravity data and their adjustment, we anticipate a step change in the quality of our knowledge of W . How would the frame accommodate this improvement?

Suppose that there is a new solution for W , say EGG2005, and this is then adopted for EVRF2008. There will be a direct change in the estimate of heights throughout Europe because W will change at every site where EGG2005 is not equal to EGG97. The new set of heights would be 'in the frame of EGG2005 and GRS80'. In this type of frame, the global datum is *definitive* and we should think of NAP only as a 'working hypothesis'. My prediction is that, post-GOCE, future further improvement of the gravity field model W will correspond to centimetric height changes and the need to change the frame thereafter will be minimal. We shall be in a position rather similar to the 'working hypotheses' of ITRF.

References

- DENKER, H.; BEHREND, D. & TORGE, W. (1997) *EGG97, European Geoid and Quasigeoid Models*, CD-ROM, Inst. für Erdmessung, Hamburg.
- FORSBERG, R.; DTRYKOWSKI, G.; ILLIFE, J.C.; ZIEBERT, M.; CROSS, P.C.; TSCHERNONG, C.C.; CRUDDACE, P.; FINCH, O.; BRAY, C. & STEWART, K. (2002) OSGM02: A new geoid for model for the British Isles, *this volume*.
- HEISKANEN, W. & MORITZ, H. (1967) *Physical Geodesy*, Freeman, San Francisco.
- HEISKANEN, W. & VENING MEINESZ, F. (1958) *The Earth and its gravity field*, McGraw Hill, New York.
- HIPKIN, R. (2002a) Is there a need for a Geodetic Datum 2000: Discussion of a Heiskanen & Moritz proposal, pp 124-127 in *Vistas for Geodesy in the New Millennium* (eds J. ? dàm & K.-P.Schwarz), Springer Verlag, Berlin.
- HIPKIN, R. (2002b) Ellipsoidal geoid computation, *J. Geodesy*, *in press*.
- HIPKIN, R.G. (1988) Bouguer anomalies and the geoid: a reassessment of Stokes' method, *Geoph. J.*, **92**: 53-66.
- HOLT, J.T.; JAMES, I.D. & JONES, J.F. (2001) An s-coordinate density evolving model of the northwest European continental shelf. Part 2: Seasonal currents and tides. *J. Geoph. Res.*, **106C**, 14035-14053.
- IHDE, J. & AUGATH W. (2002) The European Vertical Reference System, its relation to a World Height System and to the ITRS, pp 78-83 in *Vistas for Geodesy in the New Millennium* (eds J. Adam & K.-P.Schwarz), Springer Verlag, Berlin.
- KASSER, M. (1989) Un nivellement de très haute précision: la traversée Marseillais-Dunkirque 1983. *C. R. Acad. Sci. Paris, Ser. II*, **309**, 695-700.
- NAKIBOGLU, S M (1982) Hydrostatic theory of the Earth and its mechanical implications, *Phys Earth Planet Int*, **28**, 302-311.
- SMITH, D.A (1998) There is no such thing as "The" EGM96 geoid: Subtle points on the use of a global geopotential model, *Int Geoid Service Bull.*, **8**: 17-27.
- STOKES, G.G. (1849) On the variation of gravity on the surface of the Earth, *Trans. Cambridge Phil. Soc.*, **8**: 672-695.
- WEBB, D.J.; COWARD, A.C.; DE CUEVAS, B.A. & GWILLIAM, C.S. (1997) A multiprocessor ocean general circulation model using message passing. *J Atmos Ocean Tech.*, **14**, 175-183.