Ocean Sci. Discuss., 3, 1543–1568, 2006
www.ocean-sci-discuss.net/3/1543/2006/

Ocean Science
Discussions

Papers published in *Ocean Science Discussions* are under
open-access review for the journal *Ocean Science*

An oceanographer's guide to GOCE and the geoid

C. W. Hughes and R. J. Bingham

Proudman Oceanographic Laboratory, 6 Brownlow St., Liverpool L3 5DA, UK

Received: 5 September 2006 – Accepted: 11 September 2006 – Published: 20 September 2006

Correspondence to: C. W. Hughes (cwh@pol.ac.uk)

Abstract

A review is given of the geodetic concepts necessary for oceanographers to make use of satellite gravity data to define the geoid, and to interpret the resulting product. The geoid is defined, with particular attention to subtleties related to the representation of the permanent tide, and the way in which the geoid is represented in ocean models. The usual spherical harmonic description of the gravitational field is described, together with the concepts required to calculate a geoid from the spherical harmonic coefficients. A brief description is given of the measurement system in the GOCE satellite mission, scheduled for launch shortly, followed by a description of a reference ellipsoid with respect to which the geoid may be calculated. Finally, a recipe is given for calculation of the geoid relative to any chosen ellipsoid, given a set of spherical harmonic coefficients and defining constants.

1 Introduction

Satellite gravity measurements are becoming a very important tool in physical oceanography, with the success of the GRACE mission and the imminent launch of GOCE. Accordingly, it is becoming important for oceanographers to understand satellite gravity. This is not as straightforward as might be thought, since there are a number of subtleties of geodesy associated with the interpretation of gravity data, and the usual product takes the form of a set of spherical harmonic coefficients. Oceanographers are generally not used to working with either of these, so the purpose of this note is to describe the basics of the relevant geodetic issues, with particular reference to GOCE and its measurement system. The aim is to describe the static (time mean) component of the gravity field, without going into the additional detail necessary to understand the time dependent gravity field and its relationship to mass movements in the earth system.

The primary geodetic quantity of interest to oceanographers is the geoid. This is the level surface which would coincide with sea level if the ocean was in a static equilibrium. It is the surface relative to which slopes must be calculated to determine geostrophic currents (with a correction for atmospheric pressure gradients). The geoid can be determined from space by measuring the earth's gravity field via its effect on the motion of satellites and of control masses within those satellites.

This note starts by defining the geoid, and noting some subtleties to its definition. This is followed by a description of the spherical harmonic representation of the geoid and some aspects of that which must be accounted for in interpreting the data. A description of the GOCE measurement system is then given, followed by a précis of the definition of an ellipsoidal reference earth model which can be used to perform geoid computations, and a recipe for computation of geoid heights given a set of spherical harmonic coefficients.

2 Definition of the geoid

The geoid is a "horizontal" or "level" surface, a surface which is everywhere perpendicular to the local direction of gravity. If there were no waves or currents in the ocean, it is where the sea surface would eventually settle in equilibrium. Since dynamics in the ocean make it possible for sea level to depart from the geoid, the actual vertical distance of sea surface height above the geoid is known as the ocean's dynamic topography.

The actual shape of the geoid includes structure at all length scales. To a first approximation it is a sphere with radius about 6371 km. A closer approximation is an ellipsoid, with equatorial radius about 21.4 km longer than the polar radius. Relative to this ellipsoid, the geoid undulates by up to 100 m on the largest scales. On relatively short length scales (a few km to a few hundred km) the geoid is closely related to topography as the gravitational attraction of, for example, a seamount will pull water towards it leading to a bump in the sea surface above it (although gravity is stronger immediately above the seamount, this does not lead to a depression in sea level. Rather, it is

the lateral gravitational force which pulls water from either side of the seamount, leading to a raised level above the seamount). This is the principle behind using sea level measurements from satellite altimetry to help map the sea floor, as used for example by Smith and Sandwell (1997).

The geoid is not, however, simply a gravitational equipotential surface. The earth is rotating, and in the rotating reference frame we feel a centrifugal force which must be added to the gravitational attraction to give what is usually termed "gravity".

To summarise this in mathematical terms, if we write the acceleration due to gravity as the gradient of a potential

$$\boldsymbol{g} = \nabla W, \tag{1}$$

then the geoid is a surface of constant W (note the sign in this equation: the geodetic convention is, counterintuitively, that greater height corresponds to lower potential, unlike electrostatic theory, for example).

There are an infinite number of surfaces of constant W (geopotential surfaces), which results in the question of which one to define as "the" geoid. Although loosely defined as the geopotential closest to observed sea level, it is in practice usually calculated as the geopotential corresponding to the value at the surface of a fictional reference ellipsoidal earth with approximately the same mass, radius, and flattening (i.e. equatorial bulge) as the real earth.

Strictly speaking, for comparison with sea level measurements, we are not interested in differences of sea surface height relative to one particular geopotential, but in differences of geopotential along the sea surface. As long as the sea surface is close to the geoid, these are the same to an accuracy of order 5 mm, but the accuracy decreases if the geoid is further than about 1 m from the sea surface. To see this, consider the potential W_1, W_2 at two points near the earth with heights h_1, h_2 above the geoid, and at points with different strengths of gravity g_1, g_2. Writing $\Delta W = W_2 - W_1$, etc., and approximating $W = gh$, we have

$$\Delta W = (g_1 + \Delta g)(h_1 + \Delta h) - g_1 h_1 = g_1 \Delta h + h_1 \Delta g + \Delta g \Delta h. \tag{2}$$

If we use Δh to estimate the potential difference according to $\Delta W \approx g_0 \Delta h$, where g_0 is some average gravity, then the difference between these two formulae for ΔW gives the error δW resulting from the approximation:

$$\delta W = (g_2 - g_0)h_2 + (g_0 - g_1)h_1. \tag{3}$$

Dividing through by g_0 then gives a value for the size of these errors interpreted as a height. If we consider a geoid close to mean sea level, then the maximum value of h for geostrophic flows is about 1 m, and the maximum fractional change in g, between its average and extreme values, is about 0.25%, so this leads to a maximum possible error of 5 mm. In fact this is likely to be an overestimate, since the calculation assumes both terms in (3) have the same sign, whereas the distribution of h in the real ocean means that they will have opposite signs for the largest values of h_1 and h_2. Nonetheless, an error of this order is unavoidable if a dynamic topography is to be calculated as difference between sea level and a geoid, rather than as the geopotential on the sea level surface.

If, however, the geoid is not close to the mean sea level, then the error scales quite differently, as $h\Delta g/g_0$ where h is the displacement of the geoid from sea level. Since $\Delta g/g_0$ can reach 0.5%, this results in an error of up to 5 mm for every metre of displacement of the geoid. For centimetric accuracy, it is therefore necessary to choose for the geoid a geopotential surface which intersects the mean sea surface, preferably close to half way between extreme values of the dynamic topography.

The relationship between geopotential W and the gravitational potential V due to the earth's mass is given by

$$W = V + \Phi, \tag{4}$$

where Φ is the centrifugal potential. The gravitational potential is related to mass by

$$\nabla^2 V = -4\pi G\rho \tag{5}$$

where G is the gravitational constant, and ρ is density of the earth, expressing the fact that mass is the source of gravitational attraction. Outside the earth and its atmosphere, ρ=0, so V obeys Laplace's equation;

$$\nabla^2 V = 0. \tag{6}$$

In this case, V is termed an harmonic function in free space. It is usual to define V such that V tends to zero at infinite distance from the earth. The centrifugal potential is given by

$$\Phi = \frac{\Omega^2 r^2 \cos^2\theta}{2} \tag{7}$$

where Ω is the earth's angular rotation rate, r is radial distance from the earth's centre, and θ is angle subtended at the earth's centre, measured northwards from the equator (this is geocentric latitude, which differs slightly from the geodetic latitude normally used in maps, ocean models, and altimetry products, see next section for more detail). $r\cos\theta$ is the distance from the earth's rotation axis, measured perpendicular to that axis. Φ is zero at the rotation axis, and surfaces of constant Φ are cylinders centred on the axis, with Φ increasing to ∞ as distance from the axis increases.

A second way of decomposing W is

$$W = U + T, \tag{8}$$

where U is the so-called normal gravity potential (sum of gravitational and centrifugal) for an idealised reference earth, and T is the anomalous potential due to non-equilibrium mass distribution in the earth. U is not harmonic, since it includes the centrifugal potential, but T is harmonic outside the earth and atmosphere, obeying

$$\nabla^2 T = 0 \tag{9}$$

in free space, and

$$\nabla^2 T = -4\pi G\rho' \tag{10}$$

elsewhere, where ρ' is the density anomaly compared to the reference earth.

A satellite measures quantities which permit the calculation of V at satellite altitude. Given this boundary condition, and the assumption that the measured V is all due to mass enclosed within the satellite orbit (requiring corrections to be made for the effect of sun and moon, to be discussed in the next subsection), it is possible to solve Eq. (6) to define an artificial V in all space down to some radius beneath the earth's surface (V becomes singular deeper within the earth). In free space, this V will correspond to the true V but on descending beneath the earth's surface they diverge as ρ is no longer zero. This makes little difference down to the surface of the ocean, where the only correction necessary is due to the atmosphere. This correction amounts to a constant lifting of the geoid by about 6 mm over the ocean, plus smaller (<1 mm) adjustments to account for lateral variations in atmospheric mass. Larger adjustments are necessary over land, where the geoid lies beneath the solid earth surface, but we will not be concerned with those corrections here, and will in fact ignore the atmospheric correction as it is dynamically irrelevant (the 6 mm signal being constant over the ocean). This process of taking measurements at satellite altitude and projecting them down to the earth's surface or geoid is known as "downward continuation".

2.1 The permanent tide system

The discussion above relates to the gravitational field of the earth, together with the centrifugal potential due to earth rotation. A complicating factor is that there are also gravitational forces exerted by the sun and moon, and the earth accelerates in response to these forces. This is the phenomenon which produces the tidal forces leading to ocean and earth tides. The usual definition of the geoid averages out the periodic forces, but an issue remains about the permanent tide. This results from the fact that, averaged over a long time, the masses of the sun and moon would appear as bands hovering at great distance over the equator. This results in an addition to the gravitational potential (and in an increase in the earth's equatorial bulge in response to it). There are a number of ways of dealing with this effect.

In the "mean tide" system, the effect of this extra band of mass is included in the definition of the gravity field and geoid. This means that the geoid corresponds to a genuine equipotential surface–the most physically meaningful case for oceanographers and simplest for comparison with satellite altimetry.

In the "zero tide" system, the gravitational attraction to this extra band of mass is removed from the gravity field definition (this correction is precisely known from measurements). This can occur as a side-effect of removing the time-dependent tides, if their average is not explicitly replaced in the calculation. The mean tide should then be added back into any geoid calculated based on a zero tide system. The zero tide system is well-defined, and is the most natural for a representation of the earth's gravity field as a sum of spherical harmonics, as discussed later. It is the system used, for example, for the spherical harmonic representations of the GRACE GGM02 mean geoids.

The "tide-free" or "non-tidal" system is a theoretical construct in which the gravity field is calculated by not only removing the mass of the sun and earth from the system, but also allowing the earth's bulge to relax in response to that absence, and adding in the effect of the resulting redistribution of earth mass to the gravity field. This is purely theoretical as it is not known how much the earth would relax in response to such a perturbation, and an assumption has to be made about the size of the (unmeasurable) "zero frequency Love number" in order to calculate this effect. To convert from tide-free to mean tide, it is therefore necessary not only to add back in the effect of the sun and moon mass, but also to know what Love number was assumed in the system. In practice, a form of "tide free" system is often used since, in correcting for the effect of tides, a correction is usually also made for the extra effect due to the tides induced in the solid earth by motions of the sun and moon. This is a simple correction to make, again using a Love number, and (again, unless the mean tide is explicitly replaced) has the effect of producing measurements in the "tide-free" system. However, this is a version of the "tide-free" system which uses a Love number (usually 0.3) appropriate to tidal frequencies instead of the true (unknowable) Love number appropriate to the permanent tide (expected to be closer to a value k=0.94 calculated for a fluid earth).

The geoid in the mean tide system is higher at the equator and lower at the poles than in the zero tide system, the difference being 29.6 cm×$(1/3-\sin^2\theta)$. The difference between mean tide and tide-free geoids is larger by a factor $(1+k)$ where k is the Love number used (usually 0.3).

A further complication occurs in consideration of land movement, for example in GPS coordinate fixing of tide gauges. Absolute positions relative to a reference ellipsoid are the same in both mean tide and zero tide systems. In the tide-free system, however, the equatorial bulge is artificially reduced. Land positions in the tide-free system are thus higher at the equator and lower at the pole than in the other systems, the difference being 29.6 cm×$h(1/3-\sin^2\theta)$, where h is another Love number. The conventional value is about h=0.61, again really appropriate only to relatively high frequencies (the value for a fluid earth is about 1.94). More detail about permanent tides can be found in Ekman (1989) and Rapp (1989).

2.2 The geoid in ocean models

In an ocean model it is usual to use what are thought of as spherical coordinates: latitude, longitude, and vertical. Irrespective of what vertical coordinate system the model uses, there will be a z coordinate implicit in the model which represents distance in the vertical. It is important to recognise that surfaces of constant z are not really determined by distance from the earth's centre. They really represent surfaces of constant geopotential W. The dynamics of the models assume that gravity acts along the z direction, and therefore perpendicular to a surface of constant z. More accurate implementation of the actual geometry of the geoid in an ocean model would not involve adding gravitational forces along the horizontal directions, but involves re-interpreting the geometry of the grid to account for the fact that a given change in z corresponds to different lengths at different positions on the earth. In practice, such a correction makes differences only at the 0.5% level (the effect of the 21 km bulge, smaller again for the smaller-scale effects), and is far from the main source of error in ocean models.

Equally, the latitude in ocean models should be interpreted as geodetic latitude (also sometimes called geographic latitude). That is the latitude used in all maps, and in altimeter products. It is defined as the angle between the normal to the reference ellipsoid and the equatorial plane, which differs from the geocentric latitude because of the departure of the ellipsoid from a sphere. The conversion from geocentric latitude θ to geodetic latitude θ' is given by

$$\tan\theta' = \frac{\tan\theta}{(1-f)^2} \tag{11}$$

where f is the ellipsoidal flattening (defined as $(a-b)/a$ where a is the semimajor axis or equatorial radius of the ellipsoid and b is the semiminor axis or polar radius). The flattening used for GOCE processing is the value from the Geodetic Reference System 1980 (Moritz, 1980a) and is 0.00335281068118, or 1/298.257222101, although other values are used in other circumstances – see Sect. 5 for some examples. The difference between the two latitudes reaches a maximum of about 0.192° at latitude 45° (geodetic latitude is greater than geocentric for a point in the northern hemisphere), corresponding to an offset distance of about 21 km. If misinterpreted, this offset can have dramatic consequences, as the height of the ellipsoid relative to a sphere can change by more than 70 m over this distance. Note also that numerical problems may result if the conversion formula is used at the poles, where $\theta'=\theta$, since $\tan\theta \to \infty$.

3 Spherical harmonics

The usual way to represent the gravity field is in terms of spherical harmonic coefficients. This is because spherical harmonics are solutions to Laplace's equation which are separable in spherical coordinates, which makes them particularly useful for calculations involving downward continuation (although other basis functions are sometimes used, most notably ellipsoidal harmonics). In terms of spherical harmonics, and using spherical coordinates θ (geocentric latitude), λ (longitude), and r (distance from earth's centre), the gravitational potential V is defined by

$$V(r,\theta,\lambda) = \frac{GM}{R}\sum_{l=0}^{\infty}\left(\frac{R}{r}\right)^{l+1}\sum_{m=0}^{l} P_{l,m}(\sin\theta)[C_{l,m}\cos m\lambda + S_{lm}\sin m\lambda], \tag{12}$$

or

$$V(r,\theta,\lambda) = \frac{GM}{R}\sum_{l=0}^{\infty}\left(\frac{R}{r}\right)^{l+1}\sum_{m=0}^{l} K_{l,m}Y_{l,m}(\theta,\lambda), \tag{13}$$

with $(P_{l,m}\cos(m\lambda), P_{l,m}\sin(m\lambda))$ and $Y_{l,m}$ the real and complex valued spherical harmonics of degree l and order m respectively, and $C_{l,m}$, $S_{l,m}$, $K_{l,m}$ numerical coefficients (complex, in the case of $K_{l,m}$). The other terms are GM where G is the gravitational constant and M the mass of the earth + atmosphere (the product is known to much better accuracy than either individually), and R, which is a reference radius, usually taken close to the earth's mean radius or semi-major axis. For a full specification of the gravity field, it is necessary to know the spherical harmonic coefficients, and the values of GM and R with respect to which they were computed.

The spherical harmonic representation is analogous to a Fourier representation of a field on a plane. The Fourier coefficients describe the amplitude of each wavelength on the plane. If the field obeys Laplace's equation, then it can be calculated above that plane from the same coefficients multiplied by $e^{-i\kappa z}$ where $\kappa=\sqrt{k^2+l^2}$ is the total horizontal wavenumber and z the vertical distance above the original plane (this assumes the field decays to zero as $z\to\infty$, otherwise there can also be exponentially growing solutions). In spherical harmonics, we can think of a field defined on a spherical surface R. If that field obeys Laplace's equation then the value at r can be calculated by multiplying each coefficient by $(R/r)^{l+1}$, showing how the field decays as r increases (again, there is another, growing solution possible if the field is not required to decay at infinity. For our purposes, the growing solution applies to masses outside the satellite orbit, while the decaying solution applies to the part of the potential resulting from the

earth's mass). The degree l is therefore analogous to the total horizontal wavenumber κ, whereas the order m is like k, being a zonal wavenumber. The main difference from the plane case is the way in which the spherical harmonics depend on latitude and longitude. On a plane, the functions of x and y are both sine waves. On a sphere, the function of λ is a sine wave, but the function of θ is a more complicated function of $\sin\theta$ (the associated Legendre functions). Furthermore, each pair (l,m) defines a different associated Legendre function.

For $m=0$ the harmonics have no dependence on longitude, and are therefore functions of latitude only. These harmonics are known as "zonals". For $m=l$, the associated Legendre function is positive everywhere (although its amplitude becomes concentrated close to the equator for high degree l), resulting in harmonics with nodes only along meridians, known as sectorial harmonics. Other harmonics have both zonal and meridional nodes, and are called tesseral harmonics. See Fig. 1 for examples of degree 3 harmonics.

The spherical harmonics are usually used in "fully normalised" form, which is defined so that the square of a spherical harmonic function, integrated over a unit sphere, integrates to 4π. The functions are orthogonal, meaning that the product of two different harmonics integrates to zero over the unit sphere.

This representation has the advantage of reducing an apparently three-dimensional problem (the potential is a field in three dimensions) to two dimensions (zonal and meridional). For example, if the potential is known on some spherical surface $r=R_0$, it can easily be calculated on another spherical surface $r=R_1$, by multiplying all the coefficients $C_{l,m}$ and $S_{l,m}$ (or $K_{l,m}$) by $(R_0/R_1)^{l+1}$.

In principle, the calculation of geoid height from these coefficients cannot be performed in a single step, as it involves calculating the potential at an unknown position. In practice it can be simplified by a linearisation about a known position, the reference ellipsoid, since the total potential W is known to be close to a constant there. This is done by using the anomalous potential $T=V+\Phi-U$ where U is a reference potential consistent with the chosen reference ellipsoid and earth rotation rate (U also includes

Φ, so the centrifugal potentials cancel in the calculation of T). The linearized geoid height N above the ellipsoid is then given by the Bruns formula

$$N(\theta, \lambda) = \frac{T(\theta, \lambda)}{\gamma(\theta)} \tag{14}$$

where γ is the gravity taken from the reference earth. For geoid heights of up to 100 m this approximation is good to sub-centimetre accuracy (having found the position of the geoid to this accuracy, it is then possible to use this more accurate position to evaluate the potential much closer to the true geoid, after which a further application of the Bruns formula results in accuracy well below millimetric).

Spherical harmonic representation also has the advantage of neatly identifying the effect of length scale. The degree l is an inverse measure of horizontal length scale of geoid anomalies associated with a particular spherical harmonic. At each degree l there are $2l+1$ coefficients corresponding to different orders m, but all have in a sense the same characteristic length scale. That "in a sense" comes from counting the number of circular nodes in each spherical harmonic. The nodes lie along either circles of latitude, or great circles through the poles (meridians), and the total number of such nodes in a harmonic of degree l is simply l (it must be remembered that, on many map projections, a great circle through the poles would appear as two vertical lines, giving the impression of two nodal lines where in fact there is only one).

Although the individual harmonics appear to treat the poles in a special way, the sum of all harmonics at a particular degree does not. For example, a spherical harmonic of degree l calculated from a rotated coordinate system in which the poles lie at 45° latitude would look unlike any of the conventional spherical harmonics, but could be calculated as a weighted sum of only the conventional harmonics of degree l, another reason for associating "degree" with "inverse length scale".

The length scale associated with harmonics of a particular degree $l=L$ is usually quoted as the half wavelength D, given in km by

$$D=20\,000/L. \tag{15}$$

Given the different geometries of different harmonics, this is rather hard to relate to an actual wavelength of any particular spherical harmonic, and is really a qualitative guide to the associated length scale. Another way of thinking of this is in terms of the number of independent pieces of information. The weighted sum of spherical harmonics up to degree $l=L$ involves $\sum_{l=0}^{L}(2l+1)=(L+1)^2$ coefficients. The area of the earth's surface is $4\pi R^2$, so the same amount of information would be provided by dividing the earth up into areas of size $4\pi R^2/(L+1)^2$ and assigning a number to each such area. This is the area of a square of side $2R\sqrt{\pi}/(L+1)=22\,585/(L+1)$ km, so a sum of all spherical harmonics up to degree $l=L$ provides the same amount of information as a grid at resolution $22\,585/(L+1)$ km. In fact, this is also the estimate of "half wavelength associated with L" that one arrives at by pursuing the analogy between l and κ for a Fourier transform on a plane square domain.

This is not a fair comparison to an ocean model, however, as an ocean model cannot be said to have useful independent information at each grid point. Ocean models often suffer from "chequerboard" errors at the grid scale, and always use artificial diffusivity to damp out errors at the shortest scales. It is probably safe to say that any feature with fewer than 3 grid points per half wavelength is unreliable in an ocean model. Taking this rough guide, the ocean model resolution equivalent to a degree L is approximately $20\,000/3L$ km, giving an equivalent model resolution of 33 km for degree $L=200$. Model studies indicate that the mean dynamic topography contains substantial variability (amplitudes over 10 cm in the Southern Ocean and subpolar latitudes) at the short wavelengths corresponding to degree 80 and higher (half wavelength less than 250 km).

3.1 Complications with spherical harmonics

The fact that the geoid, a globally defined field, is most naturally given a spherical harmonic representation, while the mean sea surface with which it is to be compared is defined in a point-wise fashion only for the ocean, presents a number of difficulties for oceanographers. To compute the difference between these two fields clearly requires that one of them be transformed into the domain of the other, while ultimately the difference between them – the mean dynamic topography – will be expressed geographically.

For any field that includes steep gradients or discontinuities, the transformation between spectral and spatial representations will lead to the well known problem of Gibbs' fringes, familiar from Fourier analysis. The Fourier series approximation of a field with a discontinuity includes large overshoots as the sine waves adjust from the vertical step to the near-horizontal neighbouring values. For a finite number of sine waves (or a finite degree spherical harmonic approximation) this results in ringing, or large oscillations at the smallest wavelength used, that decay slowly away from the discontinuity (in fact, for a true discontinuity, the overshoot remains even with infinite Fourier series, but becomes more tightly confined near to the discontinuity).

Given that it is most usefully expressed geographically, it might be thought that the most straightforward way to calculate a dynamic topography would be by reconstituting the geoid as a function of space, and then subtracting it from the measured mean sea surface. However, because the geoid can only be determined to some finite degree, a number of problems arise with this approach. Firstly, the computed geoid has an "omission error" insofar as it does not contain any information at scales shorter than those of the highest degree included. The resulting dynamic topography would therefore have highly unrealistic short length scale features resulting from this missing geoid information which would remain in the mean sea surface. Secondly, the spectrally-truncated geoid would contain Gibbs' effects due to steep gradients in the Earth's gravitational field, which occur in the vicinity of mountain ranges, subduction trenches, and seamounts. Although a small component of the geoid itself, these effects would nonetheless significantly contaminate the derived dynamic topography (these effects are large: about 20 cm root-mean-square for degrees above 250, although perhaps half that over much of the ocean).

For these reasons a more appropriate approach to calculating the mean dynamic topography is to first derive a spectral model of the mean sea surface and then re-

express it in the spatial domain but truncated at the same degree as the geoid with which it is to be compared. This method too is not without problems.

If a mean sea surface is to be represented as a sum of spherical harmonics, an implicit value over land is required. Because of the Gibbs' phenomenon, the value chosen over land will also have an effect in ocean regions, and for this reason one might think it best to choose a smooth function over land. However, Gibbs' fringes in the geoid will also appear over the ocean as a result of the truncated representation of the geoid over the same land areas. An improved method is therefore to fill land areas with geoid heights. This means that the contaminating Gibbs' fringes in the spectrally-truncated geoid and those in the spectrally-truncated sea surface height will be very similar, and will to a large extent be removed when taking the difference to calculate dynamic topography.

A second disadvantage of spherical harmonics is that they lead to a lack of transparency about the spatial distribution of errors in the geoid. For example, GOCE will not be in a precisely polar orbit, and will therefore leave patches near the poles where the geoid is poorly determined. This results in a large error in the estimated coefficient for any spherical harmonic (especially the zonals). However, combinations of harmonics which have small projection into these polar regions are well-defined, and other combinations which project strongly onto the poles are extremely poorly defined. In order to extract the spatial information (that the geoid is well-defined everywhere except near the poles) it is necessary to look at not just the errors in individual coefficients, but at the covariances of errors among combinations of coefficients. This highlights the importance of treating errors carefully.

A complication concerns the handling of the permanent tide in spherical harmonics. The simplest thing to do here is to use the zero-tide system, in which the direct gravitational effect of sun and moon is subtracted out. That is because the mass of sun and moon lie outside the satellite orbit altitude, so the spherical harmonics (in practice just the $C_{2,0}$ term) representing the effect of this mass should be the alternative ones which decay downwards. The correct way to represent this in a mean-tide system would be to have two $C_{2,0}$ terms, one for the upward-decaying effect of the earth's mass, and one for the downward-decaying effect of the sun and moon. In practice, both mean-tide and tide-free systems are sometimes artificially generated by altering the $C_{2,0}$ coefficient for upward-decaying harmonics in such a way as to give the correct effect at the geoid. This works for calculating the geoid, but is wrong for any other geopotential surfaces.

Finally, something more should be said about omission error. The error covariance provided with a set of spherical harmonic coefficients is a measure of the errors in those coefficients only, and is known as "commission error". In addition, the true geoid contains spatial scales at smaller length scales than those represented by any finite set of spherical harmonics. Errors due to this missing information are omission errors. As noted above, these can be large, and it is important to be clear about what is being compared with what, when discussing errors. A point measurement of sea level should only be compared with a point estimate of the geoid, incurring the full omission error in addition to the commission error. Almost the full omission error is incurred by a satellite altimeter measurement, which is an average over a circular area of diameter typically about 7 km.

The effect of omission error can be reduced by comparing spatial averages of sea level and geoid. Although a simple average over a defined area will have smaller omission error than a point measurement, there will still be significant error due to the interaction between small wavelength features and the sharp cut-off at the area edge. This can be reduced further by comparing weighted averages of geoid and sea level, where the weighting is by some smooth function which reduces the effect of short wavelengths. The extent to which this reduces omission error will need to be determined for different weighting functions, but can be substantial if the typical length scale of the weighting function is longer than the longest wavelength contributing to omission error.

Unfortunately, the mean sea surface has not been measured at uniformly high resolution. There are poorly-sampled gaps between satellite altimeter tracks of the repeat

missions, and the so-called "geodetic" missions of ERS-1 and Geosat, although producing a densely-sampled grid in space, did not sample at enough times to produce a genuine time-mean, so the accuracy of the mean sea surface from altimetry varies strongly from place to place. In addition, sea ice and the non-polar nature of satellite orbits leads to poorer sampling at high latitudes, and limitations of the measurement system near land, coupled with the large amplitude, high-frequency sea level variations often observed in shallow water, mean that coastal mean sea level is particularly poorly determined. This is a particular problem for comparison of tide gauge data with a mean dynamic topography derived from satellite gravity and altimetry. Omission error in coastal regions might only be reduced by recourse to local (airborne, or terrestrial and marine) gravity data at high resolution.

4 The GOCE measurement system

The GOCE satellite measures the earth's gravity field in two ways, by satellite-satellite tracking (SST) plus accelerometer, and by gradiometry. The former is the more familiar technique (the same as that used by CHAMP). The acceleration of the satellite is due to a combination of gravitational forces and body forces (such as atmospheric drag and thruster forces). Using the onboard accelerometers to determine the acceleration due to body forces, the GPS tracking of the satellite then constrains the estimation of gravitational accelerations, permitting the earth's gravitational field to be determined. This technique is particularly suited to longer wavelength parts of the gravity field.

The second method used by GOCE is gradiometry, and it is this method which permits the recovery of short wavelength features in the gravity field. Gradiometry uses a pair of accelerometers to measure the difference in gravitational acceleration between two nearby points (separated by 0.5 m for GOCE). There are three such pairs in GOCE, arranged along mutually orthogonal axes, resulting in a full measurement of the three-dimensional gradient of gravity (9 numbers, each representing the gradient of one component of gravity along one particular direction). In terms of potential, this can be represented as a 3×3 symmetric tensor with terms T_{ij} where $T_{1,2}=\partial^2 V/\partial x \partial y$, etc.

In addition to gravity gradients, the accelerometers are affected by the rotation of the satellite. This arises from the centrifugal force, the effect of which can also be represented as a symmetric tensor in apparent gravity gradients, and from rate of change of rotation, the effect of which can be represented as an antisymmetric tensor. Since all components of the tensor are measured, the antisymmetric component can be extracted and integrated with respect to time to produce a measure of the rotation rate, from which the centrifugal term can be calculated and therefore removed from the measurement. In order to avoid long-term drift in this estimate of rotation rate, and to supply the integration constant, star tracker data are also incorporated into the integration. Each accelerometer has two sensitive axes and one less sensitive axis. These are arranged so as to provide the most accurate values for the diagonal terms T_{ii} of the tensor, and for the off-diagonal term corresponding to the largest rotation rate (that due to the orbital rotation). The other off-diagonal terms are less well determined (although accurate enough for calculation of rotation rate), so the primary output of the gradiometer measurement is the three diagonal components of the gravity tensor, after correction for rotational effects.

A good check on the accuracy of removal of the rotational effects results from the fact that (ignoring the gravitational effect of the accelerometer itself), V obeys Laplace's equation $\nabla^2 V=0$. This means that the sum of the three diagonal terms (the trace of the tensor) should be zero. In contrast, the apparent gravity gradient due to a rotation with angular speed ϖ would lead to a trace of $2\varpi^2$.

There is a further redundancy in the measurement in that, in principle, any one of these diagonal components, if measured with sufficient density over a sphere enclosing the earth, is sufficient to determine the entire gravity field outside the earth. In practice, each component is sensitive to errors in a different way, and an optimal combination must be found.

Being a differential measurement of the gravity field, the gravity gradients are rela-

tively more sensitive to short wavelength features than other forms of measurement. This means that the useful accuracy of the derived geoid can be pushed to smaller scales than previously. The nominal GOCE accuracy is 2 cm to degree and order 200 (half wavelength 100 km). This requires a low orbit, expected to be around 250 km altitude. The satellite will be maintained in this orbit by a drag-compensating ion thruster system which acts to minimise the total measured acceleration. This has the dual effect of maintaining the altitude of the satellite, while increasing the sensitivity of the gradiometer.

The orbit will be sun-synchronous, with an inclination of 96.5°, meaning that there will be polar gaps within about 6.5 degrees of the poles. Gravity in these regions must be taken from previous satellite, airborne, and/or terrestrial gravity measurements to permit the calculation of a global solution. The science part of the mission will consist of two, six-month periods of measurement.

The two measurement methods provide complementary information, with SST providing more accurate long wavelength information and the gradiometry constraining the shorter wavelengths. The two contribute equally at half wavelengths near 500 km. More detailed information can be found in the GOCE mission selection report (ESA, 1999).

5 A reference ellipsoidal earth

In order to use the Bruns formula Eq. (14), it is necessary to have a good description of the gravity field associated with a reference earth with ellipsoidal geopotentials. One such reference is GRS80 (Moritz, 1980a), which will be briefly described here.

The reference earth is based on Newton's postulate, subsequently proved by Clairaut, that a rotating fluid planet can reach equilibrium as a spheroid. The resulting external gravity field is completely defined by 4 parameters, without any need to know how density varies with depth in the earth. The 4 parameters chosen for GRS80 are:

Equatorial radius of the earth a=6378, 137 m.

Product of the gravitational constant and mass of (earth plus atmosphere) GM=3.986005×0^{14} $\mathrm{m}^3\,\mathrm{s}^{-2}$.

Dynamical form factor J_2=1.08263×10^{-3}.

Angular rotation speed of the earth Ω=7.292115×10^{-5} rad s^{-1}.

The dynamical form factor can be written as $J_2=(C-A)/Ma^2$ where C is the earth's moment of inertia about its axis of rotation, and A is moment of inertia about an equatorial axis. It is actually defined as $J_2=-\sqrt{5}C_{2,0}$, i.e. the coefficient of the corresponding spherical harmonic in the less convenient conventional (rather than fully normalized) form. Note that, since the only gravitational attractions involved in this idealized model are those due to the earth itself, this is a tide-free earth, and the corresponding ellipsoid and geoid are tide-free. No correction for this is necessary, since it is simply a reference ellipsoid and field. As long as it is within about a metre of the sea surface, it is sufficient for accurate application of the Bruns formula to calculate the true geoid.

From these parameters, chosen exactly as above, it is possible to derive all other dimensions and properties of interest. Of particular interest are:

Polar radius of the earth b=6356, 752.3141 m.

Reciprocal flattening f^{-1}=298.257222101.

Equatorial gravity γ_e=9.7803267715 ms^{-2}.

Polar gravity γ_p=9.8321863685 ms^{-2}.

A formula (Somigliana's formula) for gravity γ on the ellipsoid is:

$$\gamma=\frac{a\gamma_p\sin^2\theta+b\gamma_e\cos^2\theta}{\sqrt{a^2\sin^2\theta+b^2\cos^2\theta}}, \tag{16}$$

which can be re-expressed in terms of geodetic latitude θ' rather than the spherical coordinate geocentric latitude θ as

$$\gamma=\frac{a\gamma_e\cos^2\theta'+b\gamma_p\sin^2\theta'}{\sqrt{a^2\cos^2\theta'+b^2\sin^2\theta'}}. \tag{17}$$

The spherical harmonic coefficients of the corresponding gravitational potential $U-\Phi$ can also be derived. Since the ellipsoid is independent of longitude and symmetrical about the equator, the only non-zero coefficients are those of the form $C_{2n,0}$ following Eq. 2.92 on p.73 of Heiskanen and Moritz (1967), these are given by:

$$C_{2n,0}=(-1)^n\frac{3e^{2n}(1-n+5J_2/e^2)}{(2n+1)(2n+3)\sqrt{(4n+1)}}, \tag{18}$$

where e is the first eccentricity defined as $e=\sqrt{a^2-b^2}/a$. Only a few coefficients are needed as the amplitude decreases rapidly with n.

The same equatorial radius and flattening are also used to define the WGS84 reference ellipsoid. Care must be taken when comparing with altimetry though, as this has been defined relative to a variety of reference ellipsoids over the years. For example, the orbits in Topex/Poseidon products are given relative to an ellipsoid with a=6378136.3 m (70 cm smaller than GRS80) and $1/f$=298.257, making the polar radius about 1.5 cm greater than it would be assuming the GRS80 flattening. GRACE products use the same equatorial radius as Topex/Poseidon, together with GM=3.986004415 × 10^{14} $\mathrm{m}^3\mathrm{s}^{-2}$.

6 A recipe for computing geoid heights

To finish with, we will provide a practical summary of what has been discussed, together with a few more detailed formulae, in the form of a recipe describing how to calculate geoid height at a particular geodetic latitude θ' and longitude λ, relative to a chosen ellipsoid, given a set of spherical harmonic coefficients $C_{n,m}$ and $S_{n,m}$ of a satellite-derived gravitational potential V together with the corresponding values of GM and R (we assume access to a subroutine to calculate the fully-normalised forms for the spherical harmonic functions).

The first operation is to choose a reference ellipsoid, either by choosing equatorial and polar radii a and b, or by choosing one of either a or b, together with the flattening

$f=(a-b)/a$ or its reciprocal. Alternatively, these can be derived from other defined quantities as in the GRS80 ellipsoid above.

Given a reference ellipsoid, it is now possible to convert from geodetic latitude θ' to geocentric latitude θ using Eq. (11), and to calculate the distance from earth's centre r of the chosen point on the ellipsoid from $r^2=a^2\cos^2\theta + b^2\sin^2\theta$.

To calculate geoid heights, we will also need the normal gravity at this point, given by Somigliana's formula Eq. (16). In this formula, equatorial and polar gravity are given by

$$\gamma_e = \frac{GM}{ab} - \Omega^2 a\left(1+\frac{e'q_0'}{6q_0}\right), \tag{19}$$

$$\gamma_p = \frac{GM}{a^2} + \Omega^2 b\left(\frac{e'q_0'}{3q_0}\right), \tag{20}$$

where e' is the second eccentricity defined as $e' = \sqrt{a^2+b^2}/b$, $q_0=0.5(1+3/e'^2)\tan^{-1}e' - 1.5/e'$, and $q_0'=3(1+1/e'^2)(1-(\tan^{-1}e')/e')-1$. A choice must be made here of earth rotation rate Ω to accompany the ellipsoid definition and chosen value of GM.

Next, the coefficients $C^0_{n,0}$ for the reference gravitational field corresponding to the chosen ellipsoid etc. should be calculated using Eq. (18), and subtracted from the given cosine coefficients $C_{n,m}$ to give the coefficients $C'_{n,m}$ corresponding to the anomalous potential T. In order to use Eq. (18) it is necessary to know the value of J_2, which is defined by

$$J_2 = \frac{1}{3}e^2\left(1-\frac{2me'}{15q_0}\right). \tag{21}$$

The formulae given above, and more information, particularly concerning the normal potential and related variables, can be found in Heiskanen and Moritz (1967) and Moritz

(1980b). Given this information at the chosen point, it is then possible to calculate the geoid height at the chosen point from a combination of Bruns' formula Eq. (14) with Eq. (12) as modified to represent the anomalous potential T rather than V:

$$N(\theta,\lambda) = \frac{GM}{R\gamma(\theta,\lambda)}\sum_{l=0}^{L}\left(\frac{R}{r}\right)^{l+1}\sum_{m=0}^{l}P_{l,m}(\sin\theta)[C'_{l,m}\cos m\lambda + S_{lm}\sin m\lambda]. \tag{22}$$

Note that the sum in Eq. (22) is over a finite number L of spherical harmonic degrees, and is equivalent to a sharp truncation of the spherical harmonic expansion. Far better is to weight the coefficients $C'_{l,m}$ and $S_{l,m}$ by a function $w(l)$ which is 1 for small l and 0 for large l, reducing smoothly between the two near to some value of l which corresponds to the smallest length scale at which the geoid data contain useful information. This is equivalent to applying an isotropic smoothing function. The best form for this function, and questions of whether non-isotropic functions would be better, are current subjects of research.

One final correction may be necessary, depending on the permanent tide system used in defining the spherical harmonic coefficients. If a tide-free system has been used, then the permanent tide effect should be added back into the geoid, by adding $(1+k)(1/3-\sin^2\theta)\times 29.6$ cm, to represent the true position of a geopotential surface, usually with $k=0.3$ (setting $k=0$ gives the correction if a zero-tide system has been used).

There are many more subtleties to be explored concerning optimal filtering and combination with measured sea surface heights, not to mention the extra complications of combination with surface gravity data which may be necessary to produce the highest resolution geoid, but the information presented here should be adequate to help the interested oceanographer make his or her first steps in making use of satellite gravity data.

7 Conclusions

We hope that this brief guide to some of the geodetic subtleties involved in the interpretation of satellite gravity data will make it easier for oceanographers to exploit these exciting new data sets, without falling into some of the traps which are obvious to experienced geodesists, but less clear to oceanographers coming to the subject with a different set of background knowledge.

Acknowledgements. Thanks to R. Rummel, C. Tscherning, G. Balmino, T. Baker and P. Woodworth for helpful discussions which aided in the preparation of this document. This is a contribution of the Proudman Oceanographic Laboratory's research programme: "Sea-level, bottom pressure and space geodesy".

References

Ekman, M.: Impacts of geodynamic phenomena on systems for height and gravity, Bulletin Géodésique, 63, 281–296, 1989. 1551

ESA: Gravity field and steady-state ocean circulation mission, report for mission sellection of the four candidate earth explorer missions, (available at http://www.esa.int/livingplanet/goce), ESA report SP-1233 (1), 217 pp, 1999. 1562

Heiskanen, W. and Moritz, H.: Physical Geodesy, W. H. Freeman and Co, San Francisco and London, 364pp, 1967. 1564, 1565

Moritz, H.: Geodetic Reference System 1980, Bulletin Géodésique, 54, 395–405, 1980a. 1552, 1562

Moritz, H.: Advanced Physical Geodesy, Herbert Wichmann Verlag (West Germany) and Abacus Press (Great Britain), 500 pp, 1980b. 1565

Rapp, R. H.: The treatment of permanent tidal effects in the analysis of satellite altimeter data for sea surface topography, Manuscripta Geodaetica, 14, 368–372, 1989. 1551

Smith, W. H. F. and Sandwell, D. T. : Global sea floor mapping from satellite altimetry and ship depth soundings, Science, 277(5334), 1956–1962, 1997. 1546

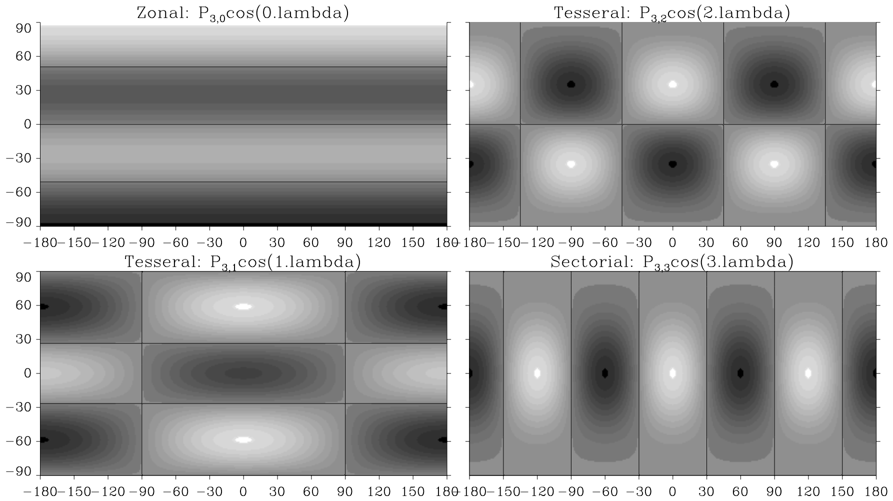

Fig. 1. Four of the seven spherical harmonics of degree 3. The remaining three are produced by shifting the patterns to the east by a quarter of a zonal wavelength. The number of circular nodal lines (horizontal lines plus half the number of vertical lines) is three in each case.