THE OCEANOGRAPHIC AND GEOIDAL COMPONENTS OF SEA SURFACE TOPOGRAPHY

by

VICTOR ZLOTNICKI

Agrimensor, Universidad de Buenos Aires, Argentina.
(1974)

SUBMITTED IN PARTIAL FULFILLMENT OF THE
REQUIREMENTS FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

and the

WOODS HOLE OCEANOGRAPHIC INSTITUTION

February, 1983

Signature of author .....
Joint Program in Oceanography,
Massachusetts Institute of Technology-
Woods Hole Oceanographic Institution, February 1983.

Certified by .....
Thesis supervisor.

Certified by .....
Thesis supervisor.

Accepted by .....
Chairman, Joint Committee for Marine Geology and
Geophysics. Massachusetts Institute of Technology-
Woods Hole Oceanographic Institution.

THE OCEANOGRAPHIC AND GEOIDAL COMPONENTS OF SEA SURFACE TOPOGRAPHY

by

Victor Zlotnicki

Submitted to the Massachusetts Institute of Technology/ Woods Hole Oceanographic Institution Joint Program in Oceanography on February 11, 1983, in partial fulfillment of the requirements for the degree of Doctor of Philosophy

ABSTRACT

Altimetric, gravimetric and oceanographic data over the North Atlantic are combined -using techniques of optimum estimation- to infer the surface expression of the time averaged circulation (ζ) and to estimate the marine geoid (γ), both in the wavelength band 100 km-2000 km.

Optimum inverse methods in geophysics are reviewed. They are then used to analyze the estimation of the geoid from gravity data, emphasizing the wavenumber spectrum of resolution functions. It is found that accurate bandpassed versions of the geoid can be recovered from restricted data sets.

The accuracy and distribution of publicly available gravity data are shown to define an estimate $\hat{\gamma}$ whose expected errors, σ_γ, range between 30 and 260 cm, assuming the Wagner and Colombo (1978) spectrum describes the average geoid behaviour. The σ_γ underestimate the actual differences between $\hat{\gamma}$ and an altimetric surface ($\hat{s}$) derived from Seasat, but the spatial variation of σ_γ follows closely the differences $\hat{s}-\hat{\gamma}$. The discrepancy is attributable to a partial failure of the spectral model at short wavelengths.

The differences $\hat{s}-\hat{\gamma}$ are dominated by geoid error that masks much of the signal ζ. The main North Atlantic gyre emerges clearly only after the σ_γ and the simplest model for ζ -as a spatially uncorrelated process with $(30\ \text{cm})^2$ variance- are taken into account. To obtain a corrected geoid, a hydrographic estimate of ζ is combined with $\hat{s}$ and $\hat{\gamma}$, and their expected errors.

Thesis Supervisors: Carl Wunsch
Cecil and Ida Green Professor of Physical Oceanography

Barry Parsons
Associate Professor of Geophysics

TABLE OF CONTENTS

CHAPTER 1: INTRODUCTION

1.1 OVERVIEW

The shape of the surface of the oceans results from the combined effects of many forces. These include the gravity fields of the earth, sun and moon, the rotation of the earth, and forces driven by the sun's heat such as the drag of the winds, atmospheric pressure variations, and pressure gradients associated with the distribution of temperature and salinity of seawater. These forces, and the ocean motions they induce or modify, are related to other phenomena of interest about which the shape of the ocean surface also conveys information. Let $s(\phi,\lambda,t)$ denote the height of the ocean surface, measured from some agreed-upon reference ellipsoid, at latitude ϕ, longitude λ, time t. The earth's gravitational attraction is, by far, the largest of the forces acting on the oceans and makes s closely resemble an equipotential surface of the earth's gravity field. Gravity is related to the density distribution inside the planet, therefore such apparently different phenomena as the direction of lithospheric plate motions, the existence of seamounts on the the ocean floor, or the pattern of convection in the mantle, all produce departures from a uniform density distribution and contribute to the undulations of s. Such geophysical phenomena can therefore be inferred, or at least constrained, from measurements of s.

The surface expression of ocean motions produces small departures of s from an equipotential surface of gravity. Because the ocean exchanges heat with the atmosphere, and ocean currents both redistribute this heat and tilt the ocean surface, s also contains indirect information on weather patterns and the longer time scale climatic variability. Because the roughness of s (its behaviour at wavelength shorter than a few hundred meters) is a direct consequence of local winds acting on the ocean, measurements of s also contain information about winds.

The only measurements of s prior to 1973 were obtained with tidal gages at scattered coastal points, and a few pressure gauges placed on the ocean bottom. These instruments provided excellent time series, but little or no information on the spatial variability of s. The first measurements of sea surface topography on a global scale were made in 1973 using an altimeter aboard the artificial satellite SKYLAB. The accuracy of altimetric measurements improved steadily over the first three missions (SKYLAB, GEOS 3 and SEASAT) and is currently 10 or 20 cm. Such an accuracy has spurred many efforts directed at recovering the wealth of information about the planet that is contained in s. This thesis will concentrate on the discrimination of time-averaged departures of s from the geoid (the equipotential surface of the earth's gravity field to which sea surface would conform if gravity were the only force acting on the oceans†). Both to justify the choice and to provide general

background information for the topics covered in later chapters, a brief review of the main components of s and the methods by which they can be measured will be given in the following sections. To this end, it is convenient to think of $s(\phi,\lambda,t)$ as a sum two terms, both varying in time: the geoid γ and the surface expression of ocean motions, ζ. In symbols,

$$s(\phi,\lambda,t) = \gamma(\phi,\lambda,t) + \zeta(\phi,\lambda,t) \qquad (1\text{-}1)$$

1.2 THE GEOID γ

This particular equipotential surface resembles an ellipsoid of revolution, with an equatorial radius of 6,378,137 ± 1 meters, and a polar radius shorter by 21,385 m. The geoid differs by less than 25 m rms from this ellipsoid. γ measures the differences between the geoid and a reference ellipsoid, adopted by international convention (see Chovitz, 1981). γ, with a spatial variability of 25 m rms, is the dominant term in equation 1-1 when compared to the approximately 1 m rms spatial variability of ζ.

---

† The geoid is usually defined as the equipotential surface of the Earth's gravity field that best fits mean sea surface (e.g., Bomford, 1980). The increasing accuracy in measurements of s will soon require a more accurate working definition. The surface expression of time-averaged motions of the ocean is usually not zero (see section 1.3); uniform changes in temperature or salinity of seawater can also produce small changes in s that do not reflect changes in the gravity field.

Large geoid changes with time occur mostly over geological time scales: mountain building, the drift of the continents and the rebound of portions of the crust following ice melting, are all processes that redistribute large masses within the 'solid' Earth at rates of a few cm per year. Erosion by winds and rivers are processes faster than continental drift, but involve masses that are small relative to the total mass of the Earth. The assumption that γ has changed negligibly during the past 20 years is made in this thesis. 'Negligibly' has the specific meaning that for all t_1, t_2 within this period, the following condition is true:

$$(1/T) \int_0^T |\gamma(\phi,\lambda,t_1)-\gamma(\phi,\lambda,t_2)|^2 \, dt_2 << 10 \text{ cm}$$

10 cm is the nominal accuracy of the measurements of s with Seasat altimetry; most of the gravity acceleration and satellite perturbation data were collected over the last 20 years. With different accuracy requirements, this assumption underlies much current work in geology and geophysics, but the assumption may be wrong. Morner (1982) claims to have identified evidence in the sedimentary record for time changes between 1 and 3 cm/year in γ, a very plausible value (his work has not been published yet).

For the purpose of discussing the different methods of estimating the geoid, it will be convenient to write it as a sum of terms in three nonoverlapping wavelength bands:

$$\gamma = \gamma_1 + \gamma_2 + \gamma_3 \qquad (1\text{-}2)$$

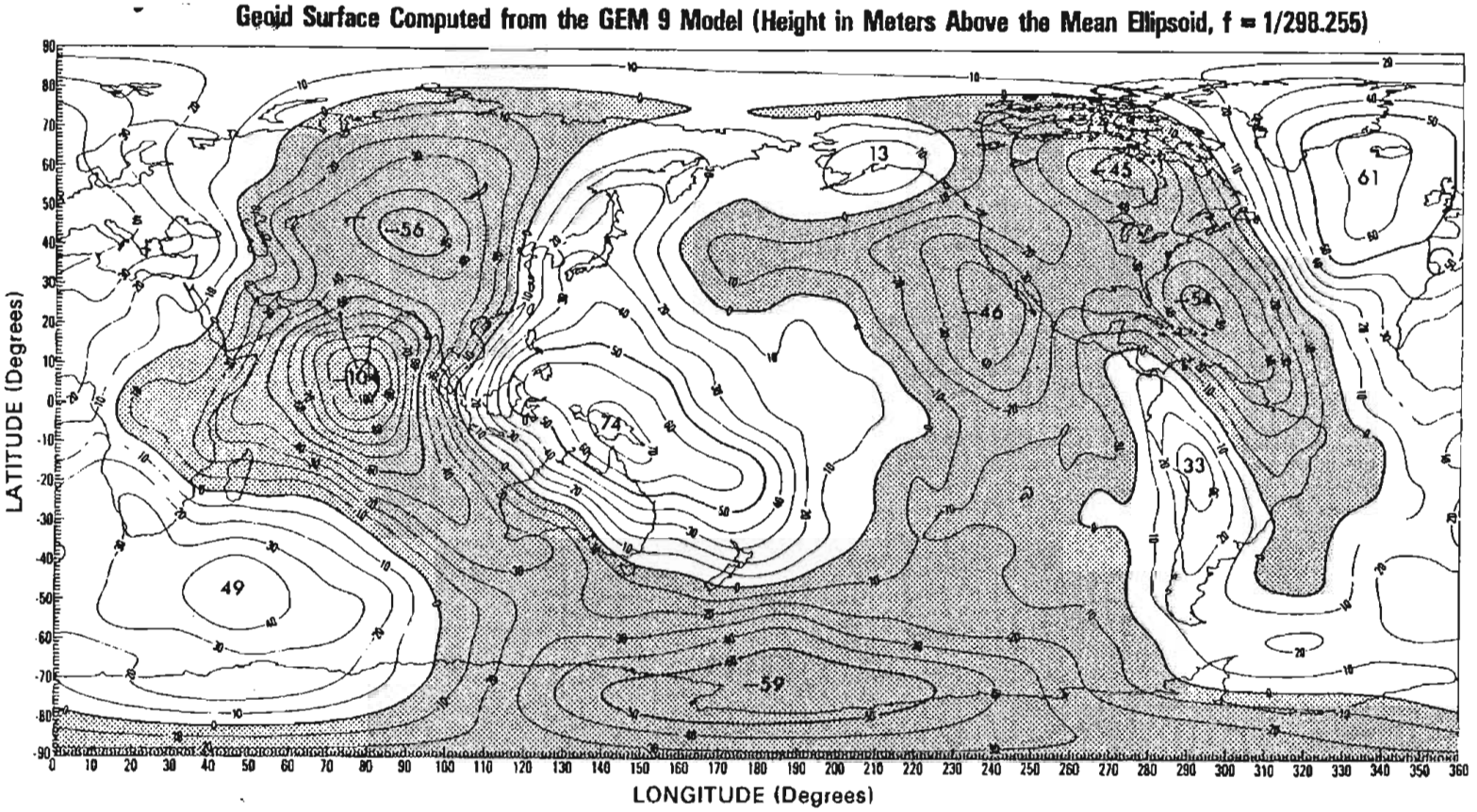

FIGURE 1-1 The long wavelength component of the geoid, estimated from perturbations to satellite orbits. This GEM-9 model (Lerch et al., 1979) is based on 840,000 satellite tracking measurements, 200,000 of which are accurate to 5 cm. Because the satellites' heights ranged between 1000 and 6000 km, and because of the exponential attenuation of high degree energy with height, this model only defines spherical harmonic degrees 2 to 20, or wavelengths longer than 2000 km, approximately. A few isolated higher degree terms to which the orbits were resonant are also defined. The expected uncertainties in this model are:

UP TO DEGREE	4	6	10	15	20
ACCUMULATED ERROR (CM):	16	30	71	128	173

A recent improvement, GEM-L2, has just been published (Lerch et al., 1982).

γ_1 contains energy at wavelengths between 20,000 km and 6700 km, γ_2 between 6700 and 50 km, and γ_3 at wavelengths shorter than 50 km.

The analysis of perturbations to satellite orbits has yielded information on the longer wavelengths of γ (energy at high wavenumbers k is attenuated as exp(-kz) as height z above the earth increases; artificial satellites fly at 700 km height or more). Although estimates of the earth's flattening from observations of the precession of the moon date back to Helmert's textbook of 1884 (quoted by Heiskanen and Moritz, 1967), the bulk of the data comes from observing artificial satellites over the last two decades. The theory of the method is presented in Kaula (1966); Heiskanen and Moritz (1967, hereafter H&M) have a brief, didactical overview chapter; Bomford (1980) surveys both the elements of the tracking techniques and some of the computational details. Formally, solutions such as GEM-9 (Lerch et al., 1979; see figure 1-1 of this chapter) describe the geoid up to spherical harmonic degree and order 20 (an approximate length scale of $2\pi R/20 = 2{,}000$ km; R=6371 km is a mean earth radius). However, the expected relative errors in the coefficients increase rapidly with degree. The choice of a cutoff degree depends only on the signal one wishes to extract from the geoid model. For most geophysical applications, GEM-9 is sufficiently accurate at least to degree 10 (length scale $\approx$ 4,000 km), but this is not so for oceanographic applications.

Tai (1982) analyzed the expected errors in GEM-9, the power at long wavelengths in a time averaged version of ζ, and the power in the difference between GEM-9 and a time average of s. He concluded that noise in GEM-9 masks any oceanographic information at corresponding wavelengths for all degrees larger than 6. Satellite orbit analysis has defined the component γ_1 as defined above with higher accuracy than any other available method.

Surface measurements of gravity accelerations using springs, pendulums and the free fall of objects (and infrequent measurements of the angle between the vertical and the normal to the reference ellipsoid) have steadily accumulated since, at least, the beginning of this century. Bomford (1980) surveys the measurement techniques and their accuracy; Talwani (1971) discussed the special difficulties of measurements at sea. Again, the majority of the available data were collected over the last two decades.

Surface gravity data can, in principle, provide information about γ at all wavelengths (the introduction to chapter 3 reviews techniques for computing geoids from gravity). In practice two problems arise: 1, because entire regions of the earth lack any data, the long wavelengths cannot be defined accurately (this problem is analyzed in chapter 3); 2, existing gravity data are not distributed densely enough to define the shorter wavelengths of the gravity field, nor do gravity measurements filter the short wavelengths (height

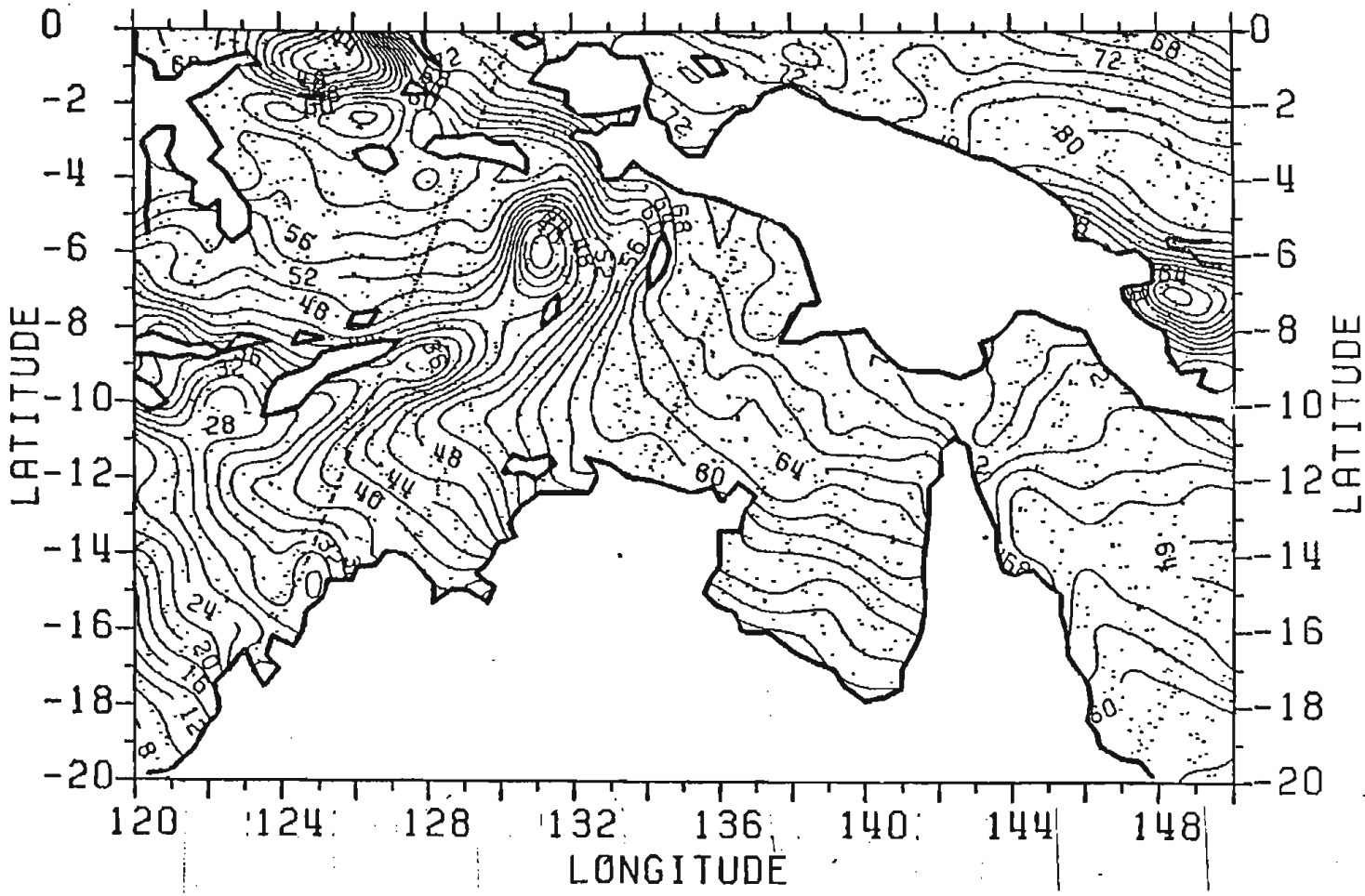

FIGURE 1-2: Seasat altimetric measurements of sea surface topography between Australia and New Guinea, from Rapp (1982a). The altimetric measurements resolve wavelengths longer than 50 km (Brammer and Sailor, 1982), with accuracies between 10 and 30 cm, depending on geographical location and the details of the crossover adjustment. Comparison with figure 1-1 shows much short wavelength energy west of New Guinea, asociated with the southern reaches of the Philippine and Palau trenches on the sea floor.

takes care of this filtering in the case of satellites). The consequence is an aliasing of gravity data by unsampled short wavelength energy acting just as if it were noise (this subject is also analyzed in chapter 3). When publicly available gravity data are used to compute a geoid with wavelengths longer than 100 km in the North Atlantic ocean, its expected errors range between 30 cm and 260 cm, and most of this error is simply due to unsampled short wavelengths (a point discussed in chapter 4).

Because s is dominated by the geoid γ, one can interpret measurements of s, from which tides have been removed, as measurements of γ with a 1 m expected error due to ζ, plus any noise associated with the measurements. Analysis of the coherence between overlapping Seasat orbits (Brammer and Sailor, 1983) shows that current altimetry can resolve geoidal length scales larger than 30-80 km, i.e., the components γ_1 and γ_2 of equation 1-2. Figure 1-2 shows a sample of the results of such altimetric measurements, and figure 1-3 shows an altimetry-derived power spectrum of γ. The special problems of altimetry will be reviewed in section 1.4, but a comparison between altimetric and gravimetric data is worth discussing now. A research vessel measuring gravity travels at some 5 m/sec (10 knots), and averages gravity accelerations over 5 minutes (a 1.5 km alongtrack average). A satellite such as Seasat travels at 7 km/sec and takes - roughly - a 10 km average of $s(\phi,\lambda,t)$ every 0.1 sec., but

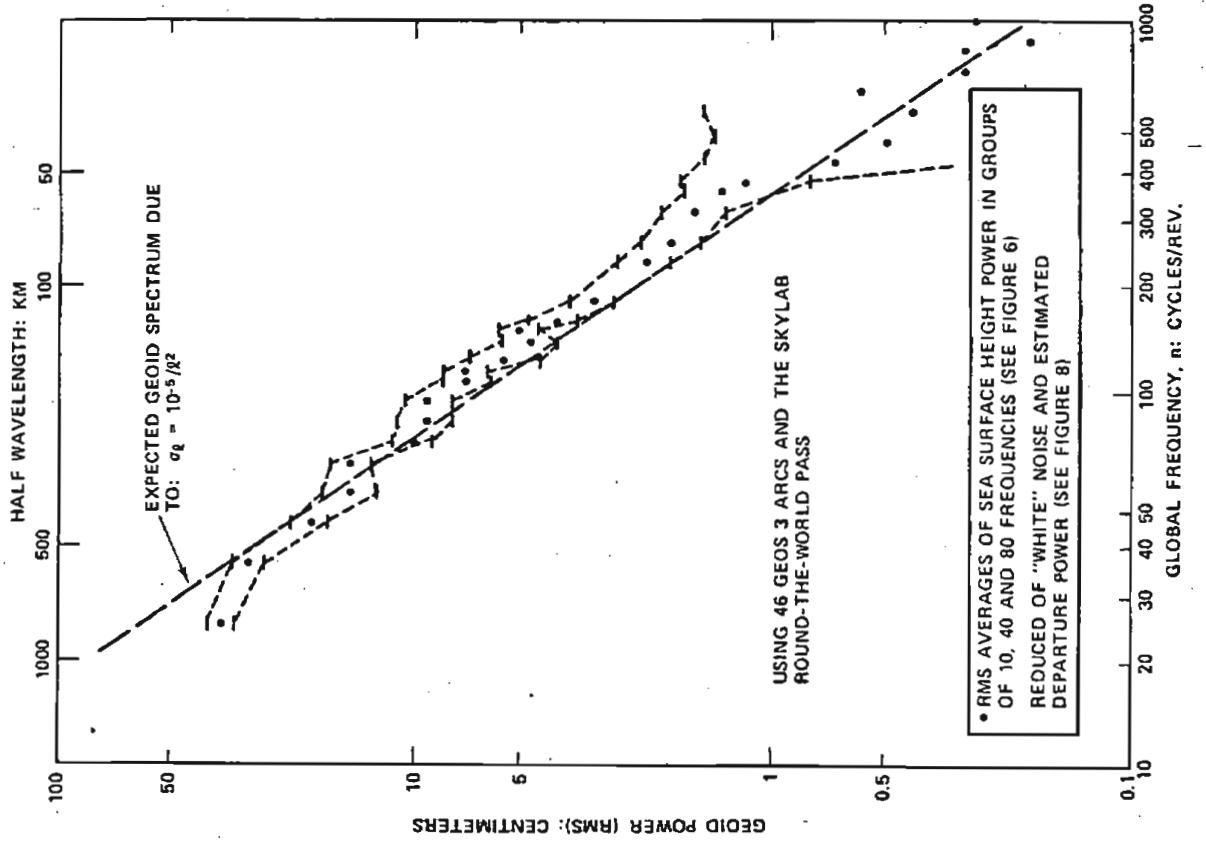

FIGURE 1-3. Geoid power spectrum (Wagner, 1979), from altimetric measurements aboard the Skylab and Geos-3 satellites. Length scales defined: 40 km to 1000 km. The best known previous estimate of the spectrum (Kaula, 1966) is known as "Kaula's rule" and labelled here "expected geoid spectrum". Kaula's rule had been estimated from surface gravity data alone. A more recent estimate, Brammer and Sailor (1983, in press), is based on Seasat altimetry.

it can only resolve gravity features longer than about 50 km. It follows that the alongtrack resolution of altimetry is about 33 times lower than the ship's, but altimetric coverage in one hour is 1400 times greater. For this reason altimeter measurements, subjected to an adequate time average that removes time-varying oceanography, are currently the best available way to describe the gravity field over the oceans.

Satellite-to-satellite tracking and satellite gradiometry are likely to yield much future data on γ. Satellite-to-satellite tracking allows an almost continuous tracking of each satellite's perturbations with very high accuracy; since short wavelengths are highly attenuated, but not eliminated at the satellite's height, they can be recovered given the proper tracking accuracy. Marsh et al. (1981) have published an estimate of γ in the Pacific by this new technique. Satellite gradiometry measures the gradient of gravity accelerations (i.e., components of the tensor of second derivatives of the potential) at satellite heights. This concept has not been implemented yet, but much preliminary analysis has been done, including studies of a proposed gradiometric mission by the French government.

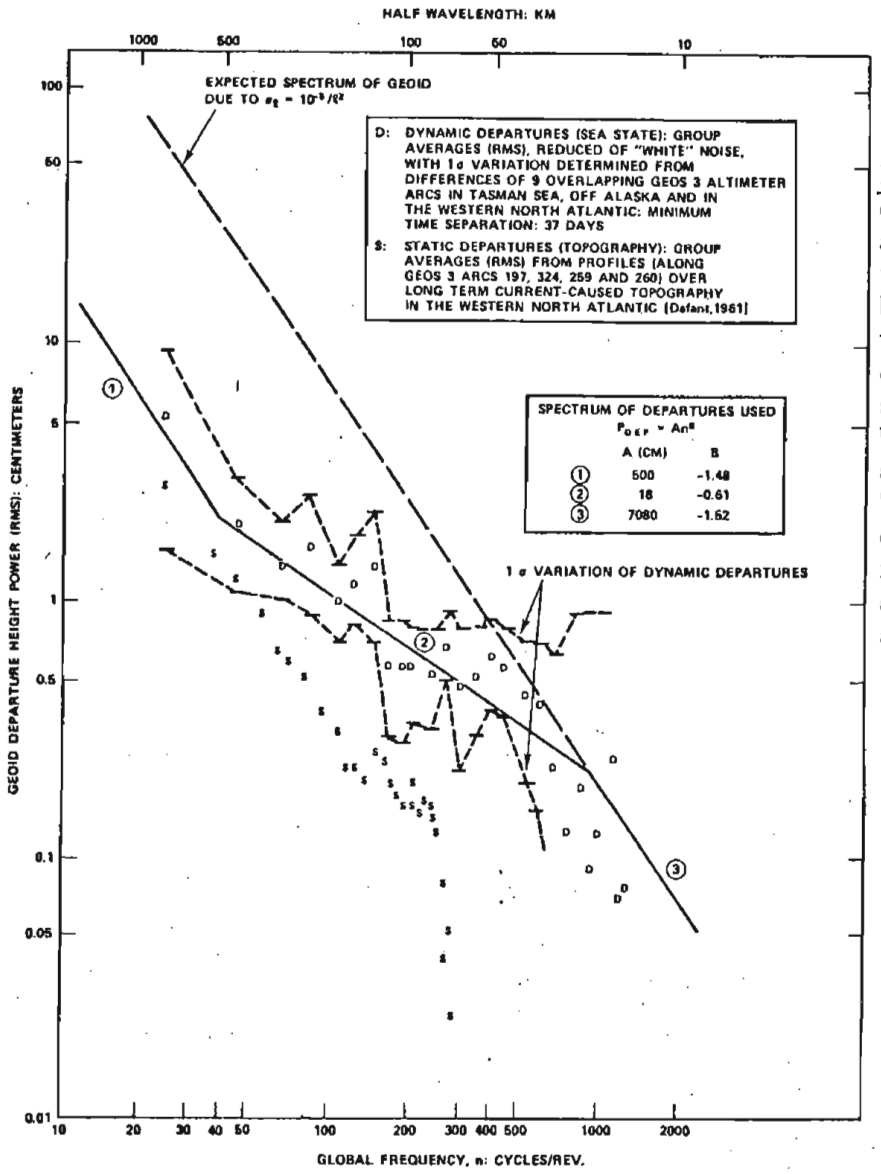

FIGURE 1-4. Comparison of the wavenumber spectra of the geoid, the time-varying and the time-averaged oceanographic components of sea surface topography, at wavelengths shorter than some 1000 km, from Wagner (1979). He estimated the time-varying component (labelled "D") from the height differences in 9 overlapping pairs of Geos-3 arcs, separated by a multiple of 37 days (526 orbits). White noise, inferred from the leveling off of the spectrum at high wavenumbers, has been removed. The time-averaged component (labelled "S") was estimated from a chart computed by Defant (1961) from historical measurements of temperature and salinity of seawater. The geoid spectrum is here approximated by Kaula's rule, for simplicity.

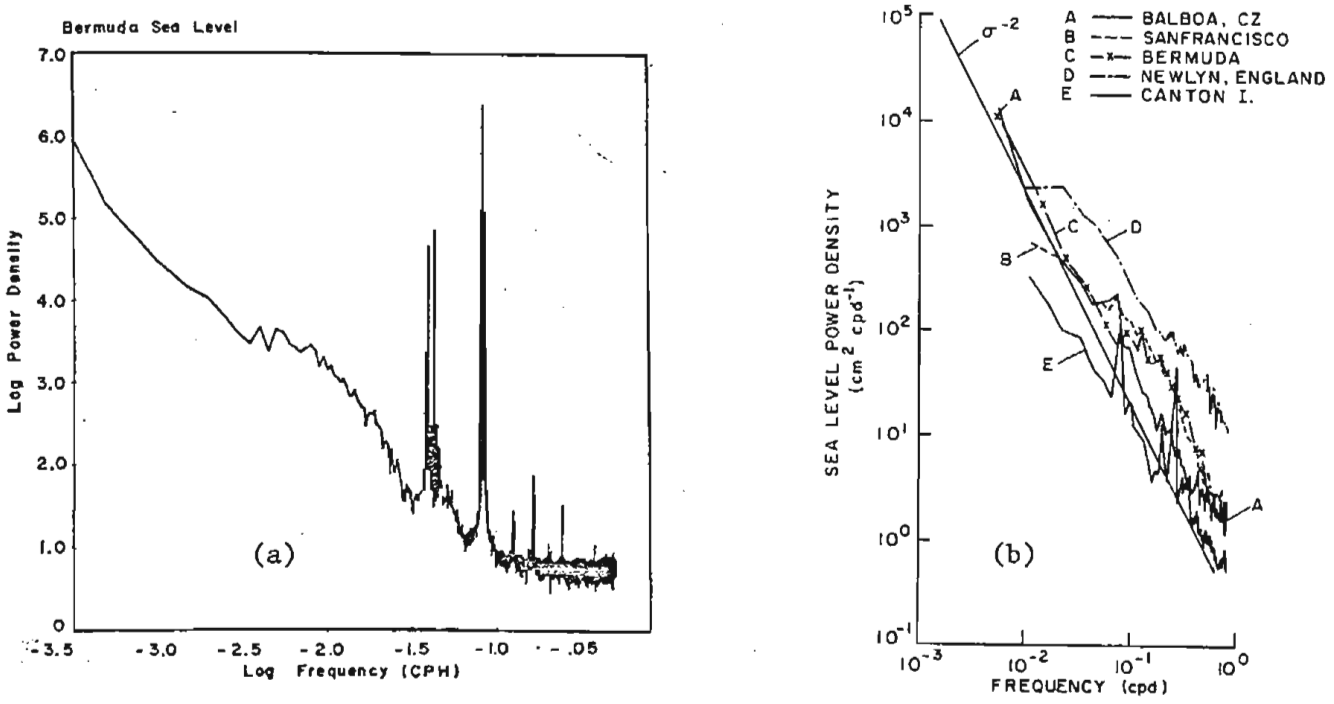

FIGURE 1-5. Frequency spectra of sea-level obtained from tide-gage data at various coastal locations of the Atlantic and Pacific oceans. a) Bermuda. The sharp peaks occur at the tidal frequencies, and for this particular site they carry 70% of the variance of the time series. The total variance is about $(28\ cm)^2$. The record used for this spectrum was sampled every 1 hour for 8 years. Frequencies in cycles per hour.

b) Locations are as indicated above the figure. Frequency is in cycles per day, hence tidal peaks fall to the right of the graph. Notice the change of power by an order of magnitude between different locations.

1.3 THE OCEANOGRAPHIC COMPONENT ζ

Two main differences exist between the oceanographic and geoidal components of s: 1) at any fixed time, the spatial variability of ζ is about an order of magnitude smaller than the spatial variability of γ (a point best illustrated in figure 1-4); 2) at any fixed point on the oceans, the time variability of ζ is much larger than that of γ (but γ is not known accurately enough for a quantitative statement).

Time variability of ζ has been observed at all periods between fractions of a second and several years. In the words of Wunsch (1981): " the ocean is filled with time-varying features, with all space and time scales, whose energy levels vary by an order of magnitude over the ocean basins. To state it slightly differently, the field of variability is locally representable by a continuous frequency-wavenumber spectrum, but the underlying process is not spatially stationary in the statistical sense, and this vitiates much of the utility of the spectral description". In some cases, energy concentration within a more or less narrow frequency band can be associated with an identifiable physical cause, and is given a particular name: wind waves have most of their energy at periods of 1-2 sec (e.g., Kinsman, 1965); tides (e.g., Hendershott, 1981) have their energy concentrated in narrow peaks at periods determined by the relative motion of the earth, moon and sun (but mostly at once or twice per day); mesoscale eddies (e.g., MODE group, 1978) are features with typical widths of 50 to 200 km that

will drift past a gage in 2 or 3 months; the Gulf Stream (e.g., Fofonoff, 1981) is in the same place -within a few hundred km-, and has the same strength -within perhaps 30%- as it had 212 years ago, when Franklin and Folger first mapped it (see Richardson,1980). Figure 1-5 gives examples of measured frequency spectra of sea-surface height at selected locations.

This thesis will concentrate on discriminating time-averaged departures between s and γ (for reasons given in section 1.5); the continuous spectra of figure 1-5 already suggest that the result will be sensitive to the averaging time T. Let

$$\zeta'(\phi,\lambda,t_o) = (1/T) \int_{t_o-T/2}^{t_o+T/2} \zeta(\phi,\lambda,t)\, dt$$

and define γ' and s' with an analogous averaging (but $\gamma=\gamma'$ by our previous assumption). If we choose T of the order of many months, then ζ' will be dominated by the signal oceanographers call 'the general circulation of the oceans'.

The general circulation is a persistent pattern of currents, their return flows, and associated spatial distribution of seawater density (e.g., Defant, 1961, Ch. 18; Warren and Wunsch (ed.), 1981, part 1); the Gulf Stream and Kuroshio currents are perhaps the best known components of the pattern. Except within a few degrees from the equator, this flow is well described by the geostrophic approximation to the equations of motion (e.g. Pedlosky, 1979, Chap. 2), which

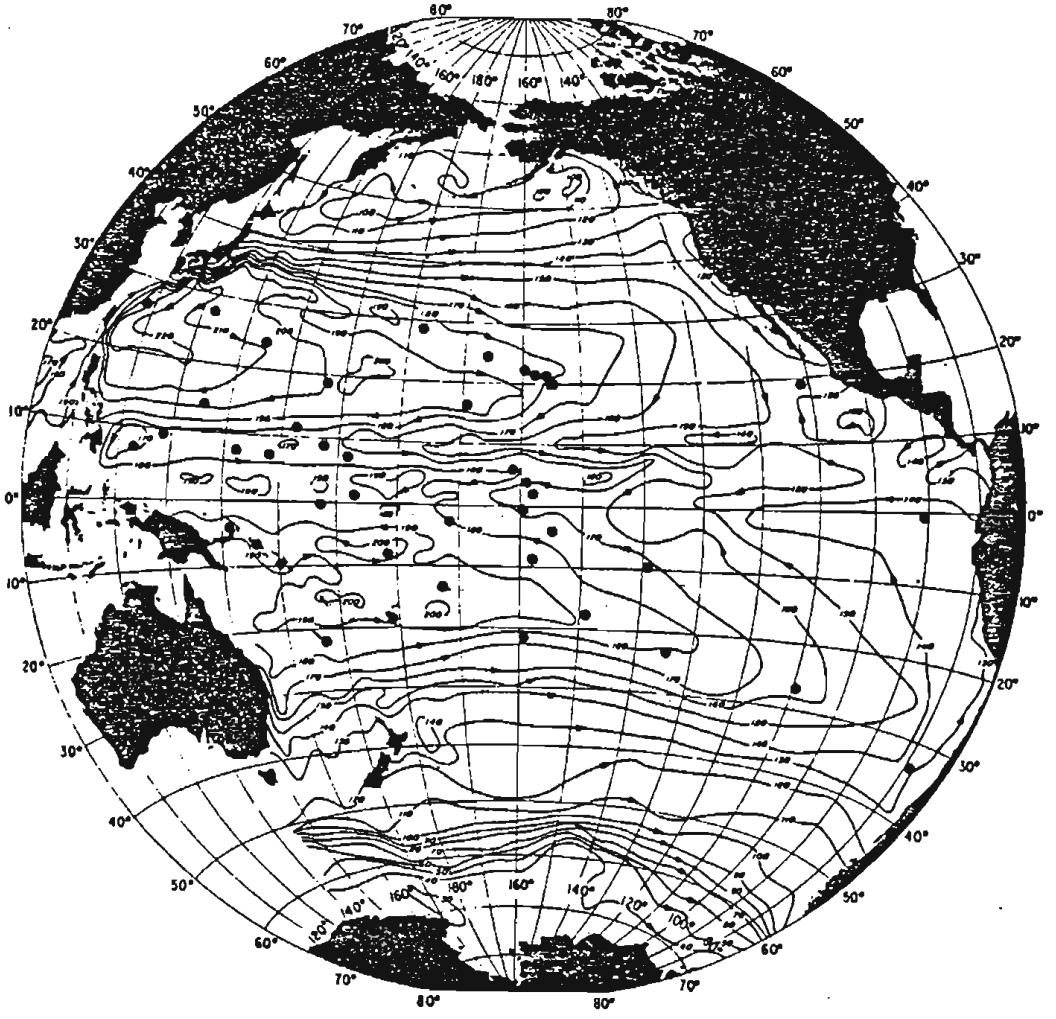

FIGURE 1-6. Estimate of the geostrophic component of sea-surface topography in the Pacific, in dynamic centimeters (a unit of geopotential, numerically equivalent -within 2%- to cm of height), from Wyrtki (1979). This chart is derived from measurements of temperature and salinity of seawater extending over 73 years. Because measurements are seldom repeated at the same station, the chart is time-aliased rather than time-averaged (figure 5-4 shows an equivalent chart for the North Atlantic). In addition, heights are computed as if a deeper surface (the 1000 dbar isobar) were motionless, a reasonable but not exact assumption. If the main assumptions were exactly true (that motion is purely geostrophic, and that no motion occurs at 1000 dbar) and if time aliasing were negligible, then this chart would represent absolute topography relative to the geoid.

essentially states that horizontal pressure gradients in the ocean (caused by variations of temperature and salinity) are balanced exactly by the Coriolis force due to the Earth's rotation; this approximation neglects friction and the transport of momentum and energy and assumes a steady state. The latter condition requires averaging times scales much larger than 1 day/2 $\sin\phi$. The surface expression of this circulation is a slope $d\zeta/dy$, normal to the direction x of the current velocity U, with approximate magnitude $d\zeta/dy \cong U/(7\times10^6 \text{ cm/sec } \sin\phi)$. Typical values of U range between 10 and 50 cm/sec, but western boundary currents, such as the Gulf Stream and Kuroshio can reach 200 cm/sec.

To interpret the results later presented in chapter 5 we need to summarize some features of the way in which oceanographers estimate the general circulation and arrive at pictures such as figure 1-6. From measurements of temperature and salinity of the oceans one obtains density $\rho(\phi,\lambda,z)$, where z is depth, at those (few) points where measurements were taken; the equation of state relating them appears in standard tables (see Neumann and Pierson, 1966, chapter 3). From the vertical profile of density at a 'station' (ϕ_o,λ_o), one can compute pressure $p(\phi_o,\lambda_o,z)$, assuming that pressure is only due to the weight of water above depth z (i.e., neglecting dynamic pressure effects). In differential form,

$$g\,\rho(\phi_o,\lambda_o,z) = -\,\partial p(\phi_o,\lambda_o,z)/\partial z \qquad (1\text{-}4)$$

where g is the acceleration of gravity. In the next step,

one makes the geostrophic assumption -that horizontal pressure gradients are exactly balanced by the Coriolis force-, i.e.,

$$\rho f \vec{U}_H = - \check{k} \times \vec{\nabla} p \qquad (1\text{-}5)$$

Here $f=2\Omega\sin\phi$ is the Coriolis parameter, Ω is the rate of rotation of the earth (once per day), ρ is a mean density, $\vec{U}_H$ is a horizontal velocity vector, and $\check{k}$ is a unit vector down the local vertical direction (the effect of $\check{k} \times \vec{\nabla} p$ is a horizontal velocity at 90° from the horizontal component of the pressure gradient, an idea that takes time getting used to). Equations 1-4 and 1-5 are combined and integrated into the equation finally used for the computations. Taking, for example, the component of $\vec{\nabla} p$ along a local x axis (x and y horizontal) we obtain the y-component U_y of the velocity vector:

$$U_y(x,y,z) = (g/\rho f) \int_{z_o}^{z} (\partial \rho(x,y,z')/\partial x)\, dz' + U_y(x,y,z_o) \qquad (1\text{-}6)$$

The chart of figure 1-6 was obtained using a reference level z_o chosen as the vertical position of the surface of constant pressure=1000 dbar, where $U_y(x,y,z_o)$ was assumed sufficiently small that its neglect in 1-6 (because U was not measured) would not alter the picture dramatically. This 'level of no motion' assumption was a theoretical difficulty that waited until 1977 for a satisfactory solution: when 1-6 is combined with a statemtent of mass conservation, $\vec{\nabla} \cdot (\rho \vec{U}_H) = 0$, into a single equation, the need for assumptions about z_o disappears. Stommel and Schott (1977) combined the

statements into a differential equation; Wunsch (1977) obtained a discretized integral equation. The practical difference between these two versions is another assumption, pointed out by Davis (1978): over what length scales is the combined equation valid when actual data are replaced in it. The relationship between 1-6 and sea-surface topography is this: if z_o is chosen as the plane that (locally) best fits the geoid, and write $z_o=0$ on this plane, then the pressure at z_o is :

$$p(x,y,0) = g\ \zeta(x,y,0)\ \rho(x,y,0)$$

and the surface velocity can be written (neglecting variations in ρ)

$$\vec{U}_H(x,y,0) = (-g/f)\ \hat{k} \times \vec{\nabla}\zeta'(x,y) \qquad (1\text{-}7)$$

If ζ' were measured, the integration 1-6 could be started at $z_o=0$; alternatively, 1-6, 1-7, and mass conservation provide an estimate of $\vec{\nabla}\zeta'$.

It is important to remember that the instantaneous reading of a moored current meter is not $\vec{U}_H$; in fact, to recover $\vec{U}_H$ from current meter data requires averaging the measurements over many years, because the time-varying component of velocity is so energetic that shorter averages do not allow a statistically significant difference between the mean and zero.

No estimate of ζ' obtained from the density field over the oceans, sampled in the traditional oceanographic way, and using the geostrophic equations to compute ζ' can equal a time average of $s-\gamma$, even assuming that both s and γ are

known exactly. The essence of the problem is that the geostrophic equation used to convert ρ into ζ is a very good, but not an exact description of ocean dynamics. Equation 1-5 does not contain $\partial\vec{U}_H/\partial t$, but the position of the Gulf Stream axis has been observed to oscillate (north of Cape Hatteras, by about 300 km over a few months. See figure 5-5, chapter 5). Equation 1-5 does not provide for any component of $\vec{\nabla}p$ parallel to $\vec{U}_H$, but the Gulf Stream is known to flow slightly 'downhill'. These features simply point to a slight difference between the pressure gradient and the Coriolis term in 1-5, a difference that must be balanced by the time evolution of the field, as in the first example, or by nonlinear combinations of $\vec{U}_H$ and its derivatives, as in the second example. The slow changes in ζ' with time are the source of the largest discrepancy between altimetric and oceanographic estimates of the circulation (still assuming that s and γ are known exactly). Because ships are slow and expensive, the relatively few measurements of the density field are widely separated in space and time, precluding a true time average. This aliasing is not really a problem at spatial scales very large compared to the Rossby radius of deformation†, roughly 50 km at mid latitudes (e.g., Charney

---

†The Rossby radius of deformation (e.g., Pedlosky, 1979), is the smallest scale over which geostrophic motions can exist at all. For scales around this radius, a variety of waves will produce time dependence of the geostrophic approximation.

and Flierl, 1981), but the chances of time aliasing increase at wavelengths approaching 50 km.

1.4 SATELLITE ALTIMETRY

The appearance of the first radar altimeter carried aboard a satellite (SKYLAB, in late 1973), to measure its height relative to the Earth's surface, triggered the current convergence of interests between geodesists, geophysicists and oceanographers. SKYLAB's altimeter had an instrumental accuracy - as opposed to overall accuracy - of 5 m (Vonbun et al., 1978). Its successors to date have been only two: the altimeters aboard Geos-3, launched in 1973 with 1 m accuracy, and Seasat-1, in 1978 with 0.1 m accuracy.

The Geos-3 mission has been reviewed by Stanley (1979); this is the first paper in an issue of the Journal of Geophysical Research entirely devoted to Geos-3. The SEASAT mission has been partially reviewed by Lame and Born (1982); details of the four instruments on board (altimeter, scatterometer, scanning multichannel microwave radiometer, and synthetic aperture radar) are given by Barrick and Swift (1980), and in other papers of the same issue of the IEEE Journal of Oceanic Engineering.

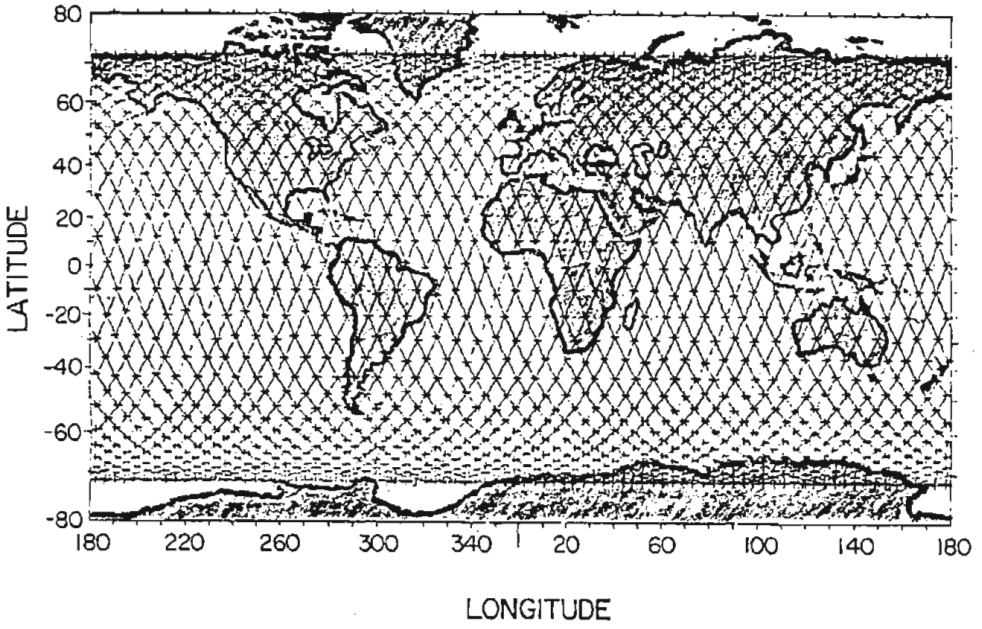

FIGURE 1-7. Seasat ground tracks during august 6-8, 1978, from Schutz et al. (1982). Seasat flew at an altitude of 800 km with an inclination of 108°. Between june 26 and august 25 it completed a revolution in 100.62 minutes, circling the Earth a little more than 14 times a day. The Earth's rotation produced each equatorial crossing of the ground track to be displaced some 25.1° to the west of the previous crossing. In addition, the ascending node of the orbit itself advanced some 2°/day relative to an inertial frame. These values imply that the ground pattern would almost repeat itself every 17 days ('almost' because of an 18 km offset at the Equator). On august 25, a slight maneuver (that increased the period to 100.75 minutes) allowed the ground track pattern to repeat itself every 3 days.

The difference in altimeter measurements on overlapping pairs of arcs measures time-varying oceanography and system errors. The coherent part along overlapping arcs measures the geoid and time-averaged oceanography. The difference in height measurements at the points where ground tracks cross (a "crossover"), after tides are removed, is dominated by the orbit interpolation error, and thus provides the constraints to reduce it (a "crossover adjustment").

The SEASAT altimeter emmited a microwave radar pulse that travelled 800 km through various layers of atmosphere, interacted with the sea-surface and perhaps with clouds, and returned to the altimeter, where both its travel time and shape were measured. Such raw measurements require many corrections before they can describe sea-surface topography; these include instrument delays, effects of the geometry of the satellite (the antenna does not coincide with the center of mass), variations in the speed of light along the path of the radar pulse, and wave-wave interactions between the pulse and sea surface. The reader should consult the recent review by Tapley et al. (1982; also Hancock et al.(1980) and TOPEX Science Working Group (1981)) for the current status of these corrections. After these corrections are applied, however, there still remains an error of some 2 meters rms, with most of its power at a frequency of once per revolution. This error occurs because the satellite is not tracked continuously, hence its position between two consecutive fixes (sometimes many revolutions apart) must be interpolated using the equations of motion and models for the Earth's gravity field, atmospheric drag and solar radiation pressure. This interpolation leaves a residual error in the radius vector, an error usually modelled as a bias and a trend over distances much smaller than its 40,000 km typical scale.

The altimeter should measure the same height where two ground tracks meet (a 'crossover', see figure 1-7), except for time-varying oceanographic features whose amplitude is

usually smaller than the crossover error (after tides are removed). The crossover discrepancies have been successfully used to correct the radial component of the position interpolation error. Rapp (1982; also Rowlands, 1981), whose adjustment of Seasat altimetry is used in this thesis, found post-adjustment discrepancies between 23 and 34 cm rms, depending on the location. Much of this remaining energy is due to time-varying oceanography, as shown by Cheney and Marsh (1981) for Geos-3; Marsh et al. (1982) found that the post-adjustment residual for Seasat data in a small area in the quiet eastern North Pacific was 8 cm if coastal zones were excluded, and 12 cm when the coastal region was included, probably because of incomplete tidal modelling.

It is fair to conclude that, after all corrections are applied, the component of $s(\phi,\lambda,t)$ with wavelength longer than some 50 km can be measured with accuracies between 10 and 30 cm using altimetry. At this writing, however, the spatial structure of this error is not known; we still need to know whether any basin-scale errors, that can mask long wavelength oceanographic information, are left in the adjusted surfaces.

1.5 THIS THESIS

The purpose of this thesis is to present appropriate methods for combining gravity, oceanographic and altimetric information in order to estimate ζ' and γ', and to perform such a computation in the North Atlantic ocean. There are two main motivations for this choice: 1) the time averaged circulation, whose surface expression is ζ', is responsible for much of the heat transport in the oceans, and thus for moderating climate. Altimetry has provided the first opportunity to obtain a global, quantitative picture of the circulation, with good areal coverage -if only ζ' could be recovered. 2) for the reasons explained in section 1.2, the best estimate of the geoid that can be obtained at present is one based on altimetry. This estimate can then be used both to analyze time-dependent oceanography (which requires removing the geoid from individual tracks), or to study solid earth processes such as mantle convection (which requires comparing the geoid to bathymetry). For these uses, the surface expression of the general circulation is a noise that one would like to remove.

Because estimates of both ζ' and γ' have large errors when shipboard data alone are used, Wunsch and Gaposchkin (1980) proposed combining the altimetric, gravity and hydrographic data to produce optimum corrections to the three surfaces involved. The relation of γ to observable gravity accelerations is linear to an excellent degree of approximation, and so is the relation between ζ' and the observable

distribution of density in the oceans; it follows that it is possible to combine all this information using the mathematics of optimum linear estimation. The enormous amount of data that would have to be considered simultaneously, however, precludes a joint inversion from basic data today. It is perfectly feasible, however, to compute initial estimates of the geostrophic component from hydrographic data alone, and of the geoid from gravity and satellite data only, and then combine the initial surfaces and their expected errors in an optimum manner, to compute corrections to the initial estimates. This approach -perturbation of initial estimates- is followed in this thesis.

Chapter 2 reviews the theory of optimum linear estimation that underlies all computations of linear models from discrete and noisy data. The necessity for such a review arises because 'least squares collocation' is used in geodesy, oceanographers follow meteorologists in the use of 'objective mapping', and most geophysicists are convinced that the 'Backus and Gilbert theory' has no equal. The fact that these names, and others, refer to essentially equivalent methods in spite of their different origins, has been recognized only over the last few years, but the literature is still scattered.

Chapter 3 analyzes the optimum estimation of geoids from gravity data, a necessary step towards the computation of an

initial geoidal model and its error structure. The analysis of chapter 3 was prompted by a few problems of available geoids: 1) geoids estimated from orbit analysis are sufficiently accurate only at wavelengths longer than 6700 km (Tai, 1982). In order to define shorter wavelengths in ζ'[†], surface gravity data must be used. 2) very good regional gravimetric geoids have been published (Marsh and Chang, 1977; Chapman and Talwani, 1979), but we know they are good only because they agree with altimetric measurements of s. This argument obviously breaks down when one tries to recover discrepancies between s and γ, but published gravimetric geoids lack error estimates. Chapter 3 starts with a brief review of current methods to construct geoids from gravity. The equations that estimate the expected error of a geoid computed as any (not necessarily optimum) linear combination of gravity data follow. The core of the chapter is an analysis of the maximum amount of information that can be extracted from limited data distributions. The main finding, in addition to the methodology, is that a fairly restricted data set can provide a band-passed version of γ, a most useful characteristic.

---

† Different wavelength bands in ζ' are interrelated both kinematically and dynamically. The Gulf Stream, for example, has a characteristic width of order 100 km. If the total transport in this wavelength band is known, then we also know the transport in the return flow needed to conserve mass, a flow whose scale ranges between these 100 km and the width of the basin. Hence knowledge of an appropriate wavelength band can, in principle, constrain the whole flow.

Chapter 4 is the first one to face real data. All publicly available† gravity data over the North Atlantic are subjected to a crossover analysis to assess their accuracy. We are unable to use optimum methods to estimate the geoid due to computer limitations, and a suboptimum method is chosen. The geoid so constructed describes the wavelength band between 100 and 2000 km with accuracies ranging from 30 cm in the western part of the ocean, to 250 cm in the eastern part. This geoid is then compared to a 3-month average of the Seasat altimetric data (adjusted by Rapp, 1982), and the discrepancies are found to agree with the expected errors, except near Florida, where the geoid has much unsampled short wavelength power associated with the Bahama islands, and the expected errors underestimate the actual discrepancies.

Chapter 5 combines the estimates of s and γ using optimum methods, but disregarding the spatial correlation between errors, again due to computer limitations. The difference $s-\gamma$ does not directly measure ζ because of the variation in errors by an order of magnitude; the effect is somewhat like changing units in the middle of the map. A simple assumption -that ζ is spatially uncorrelated, with $(30\ cm)^2$ variance- produces a believable estimate of ζ,

---

† The data used by Marsh and Chang (1978), and by Brammer (1979) belong to a classified Navy data set.

believable by comparison with the estimate of Wunsch (1981), derived exclusively from hydrographic data. When the hydrographic estimate is included in the computation we lose all independent checks on the accuracy of the result, but obtain a corrected geoid, based on altimetry, hydrography and gravity.

Chapter 6 discusses the various simplifications introduced throughout and their likely effect on the results.

In all following chapters, the unprimed s, ζ and γ will refer to time-averaged quantities.

CHAPTER 2

REVIEW OF LINEAR INVERSE METHODS IN GEOPHYSICS

2.1. INTRODUCTION

Like many other geophysical studies, this thesis deals with a finite number of noisy data (gravity anomalies, satellite heights, temperature and salinity of seawater), and these data are linearly related to either discrete parameters or continuous functions (the geoid, the geostrophic component of sea-surface topography) that we wish to estimate. The need for a brief review arises because names such as least squares collocation, Backus and Gilbert theory, Lanczos' generalized inverse, Moore Penrose inverse, ridge regression, optimal estimation, objective mapping, universal krigging, harmonic splines, and Wiener filtering, all seem to refer to competing methods of dealing with the same basic problem. In some cases the differences arise because these results were derived from different initial assumptions, such as finite versus infinite number of parameters, or discrete versus continuous independent variables, or even deterministic versus random unknowns. In other cases, an older result was rediscovered in a different discipline. Somewhat surprisingly, the optimum estimators derived in each of these cases are fundamentally similar. Broadly speaking, the similarities arise because a) the same error norm is always minimized (least squared error); b) different optimization criteria are equivalent under this

norm; c) the functions dealt with can be approximated to any desired accuracy by a finite number of parameters.

It is the purpose of this review to summarize these results (applied in all later chapters), and emphasize their similarities, differences and practical consequences.

2.2 AN EXAMPLE

To fix ideas, a one-dimensional equivalent of the problem analyzed in chapter 3 -the estimation of geoidal heights from gravity data- will be used as an example throughout this chapter.

Assume we are only interested in the local structure of a strongly lineated feature (for example, the Ninetyeast ridge in the Indian Ocean). Let y be a horizontal axis running along the ridge, x a horizontal axis normal to the ridge, and z the upward vertical direction. Assume gravity accelerations are measured at N_d points x_i, with $L=x_{N_d}-x_1$. Also assume that a long wavelength reference field (such as GEM-9) approximately defines wavelengths longer than L. Let $\tilde{g}_i$ be the difference between measured gravity and the reference gravity at x_i. Let $g(x_i)$ be the exact component of gravity with wavelength shorter than L. Finally, let h(x) be the exact geoid component with wavelengths shorter than L. Then, it is approximately true that

$$g_k = c\,|k|\,h_k \qquad (2\text{-}1)$$

where c is a constant, k is wavenumber along x, g_k is the

k^{th} Fourier coefficient in the expansion of the short wavelength gravity accelerations g(x), and h_k is the Fourier coefficient of h(x), defined by

$$h_k = (1/L) \int_{-L/2}^{L/2} h(x) \exp(-jkx)\, dx \qquad j=\sqrt{-1} \qquad (2\text{-}2)$$

$$h(x) = \sum_{\nu=1}^{\infty} h_k \exp(jkx) \qquad k=2\pi\nu/L \qquad (2\text{-}3)$$

(Chapman (1979) reviews the result 2-1; to see the weakness of a one-dimensional geometry in geoid studies, see McAdoo (1981)).

It follows that we can write g(x) as either

$$g(x) = c \sum_{k} |k|\, h_k \exp(jkx) \qquad (2\text{-}4)$$

or (at least formally, because the series does not converge)

$$g(x) = (1/L) \int_{-L/2}^{L/2} A(x-x')\, h(x')\, dx' \qquad (2\text{-}5)$$

where

$$A(x-x') = [c/(x-x')^2]$$

It is important to notice that, if the g(x) are finite then the h_k must decay fast enough to make 2-4 converge: $|h_k| << k^{-2}$ as $k \to \infty$. In other words, h(x) is a smooth function, a property that makes interpolating between data points easier, and a property that will be central to the later development.

2.3 THE FORWARD PROBLEM

NOTATION

m_p: the unknown function m (model) at point p. If the unknown are discrete parameters, the p^{th} parameter.

m_ν: generalized Fourier coefficients of m_p in some suitable orthonormal basis.

$\psi_{\nu p}$: ν^{th} basis function in the expansion of m, at p.

$\hat{m}_p$: an estimate of m_p.

$\bar{m}$: other functions or vectors in the same Hilbert space as m. In the stochastic approach, other realizations of the same stochastic process.

$\int_p$: mean value of the integrand over the volume of definition of m_p (i.e., the integral divided by the finite volume.

$\sum_p$: when the m_p are discrete parameters, sum over all p.

$\tilde{d}_i$: i^{th} data value.

d_i: linear functional of m_p, of which $\tilde{d}_i$ is a measurement

e_i: $\tilde{d}_i - d_i$, error in the i^{th} equation.

$\sum_i$: sum over all available data.

$\langle\ \rangle$: expected value of the enclosed random variable.

$M_{pp'} = \langle m_p\, m_{p'} \rangle$ (covariance of m_p, assumed random).

M_ν : power spectrum of m_p ($M_\nu = |m_\nu|^2$).

W_ν : an upper bound (not necessarily the smallest) on M_ν

[ ]: a matrix.

$\underline{d}$: vector with elements d_i.

^{T}: hermitian transpose of a matrix or vector. Also, the adjoint of a linear operator.

Good reviews of the basic formulation are given by Parker (1977), and by Aki and Richards (1980, ch. 12). Only a brief summary follows.

By physical considerations one has arrived at a linear or linearized relation ("the forward problem") between a set of observable quantities $\underline{d} = \{d_1,\ldots,d_{N_d}\}^T$, and a model, described either by a set of discrete parameters $\underline{m} = \{m_1,\ldots,m_{N_p}\}^T$, or by a function m(v) of a continuous variable†, or by an n-tuple or such functions. For generality we assume all quantities are complex; * denotes complex conjugate. Let us first assume discrete parameters m_p. The most general linear relation between the d_i and the m_p is:

$$d_i = \sum_{p=1}^{N_p} A^*_{ip}\, m_p = \sum_{p} A^*_{ip}\, m_p \qquad i=1,\ldots,N_d \qquad (2\text{-}6\text{-}a)$$

When $N_p \to \infty$, equation (2-6-a) is still valid, provided the corresponding sums converge. When a function m(v) is the unknown -such as in our example- the forward problem has the form

$$d_i = (1/V) \int_V A^*_{ip}\ m(v_p)\, dv_p = \int_p A^*_{ip}\, m_p \qquad (2\text{-}6\text{-}b)$$

Here p is the name of a point in the volume V, a point that can vary continuously through V. The second form

---

† Throughout this chapter, many differences between discrete parameters and continuously varying functions defined on a volume V will arise. The latter will be referred to as a 'function m', for brevity.

of 2-6-b will be used throughout the rest of the chapter. No confusion need arise, because the region of integration and the element dv remain fixed throughout a problem. For the remainder of the chapter, p, p',... will denote either model positions or parameters; i, i',... will indicate data values.

In our example, V is the interval $0<x<L$, not because the data were measured in that interval, but because h -as defined- contains no wavelengths longer than L, hence it is sufficient to define it over a length L.

Equation (2-6-b) is valid provided the corresponding Lebesgue integrals exist. There is another restriction that h must satisfy in order to apply the Hilbert space formulation of the next pages: a series expansion of the form

$$m_p = \sum_{\nu=0}^{\infty} \mathcal{m}_\nu^* \Psi_{\nu p} \qquad (2\text{-}7)$$

must be able to compute m_p 'almost everywhere' in V (i.e., except on sets of measure zero, for example an isolated point where m is discontinuous; at the point the series will give a mean value, but m is undefined). The basis functions $\Psi_{\nu p}$ are defined in the same volume V, and they are assumed to be orthonormal (if a basis exists, it can always be orthonormalized). Let $\delta_{\nu,\mu}=1$ if $\nu=\mu$, 0 otherwise (Kronecker delta); then

$$(1/V) \int_V \Psi_{\nu p} \, \Psi^*_{\mu p} \, dv_p = \delta_{\nu,\mu}$$

and the coefficients m_ν are given by the inverse transform

$$\mathcal{m}_\nu = \int_p m_p^* \, \Psi_{\nu p}$$

Also, A_{ip} (for fixed i) is a function defined in the same volume V. We assume it can be expanded in a series like 2-7

$$A_{ip} = \sum_{\nu=0}^{\infty} A_{i\nu}\ \Psi_{\nu p} \qquad 2\text{-}8$$

If 2-8 is absolutely convergent, then sums and integrals can be exchanged and the forward problem can be written

$$d_i = \sum_{\nu} A_{i\nu}\ m^*_{\nu} \qquad 2\text{-}6\text{-}c$$

This expansion allows us to identify the function m(v) with a countable sequence of coefficients m_ν; it also allows us, in principle, to replace the integral equation (2-6-b) with an algebraic system of equations, (2-6-c), with infinitely many unkowns (see Riesz and Nagy (1955), chapter 5).

In our example, the d_i are the (exact) short wavelength gravity values $g(x_i)$. If we identify m_p with h then the kernel A_{ip} is singular (equation 2-5). The singularity reflects a problem in the formulation, not a physical possibility for either g or h, both of which are square integrable (i.e., have finite energy). The relationship

$$g_k = c\ |k|\ h_k \qquad 2\text{-}1$$

is well posed because $\sum_k |k\ h_k|^2 < \infty$, so we will define the forward problem with 2-1 (see sections 2.6.1 and 2.6.2).

For each equation i one has a measured value $\tilde{d}_i$ that differs from the d_i by an error e_i:

$$\tilde{d}_i = d_i + e_i \qquad 2\text{-}9$$

In the example of section 2.2 there are 3 distinct

components to the error e:

a) measurement noise

b) errors in the reference field (removed from the data so that a plane approximation could be used).

c) errors in the one dimensional approximation. Modelling errors are always the hardest to describe, so one must usually assume that they are negligible.

Of course, the values of the e_i are not known, but in what follows it is assumed that their variances are known.

With these definitions, our aim is to compute the exact values of the m_p or, if this is not possible, some best approximation, $\hat{m}_p$. It is also essential that we be able to assess the accuracy of the approximation.

Our example requires the continuous formulation 2-6-b, because there is an integral constraint on the unknown h. Even if we are only interested in the value of h at one point, each constraint equation applies to the whole of h. It is also possible to use the discrete formulation for this problem, inasmuch as h can be approximated arbitrarily closely by a finite number of terms in the series 2-7; we can simply replace h by a vector that differs from h by an acceptable amount, and then estimate this vector.

A discretization of 2-6-b must be done with certain restrictions noted below. First, the integral must be

approximated in the following manner:

$$\sum_p \left(A_{ip}\ (\Delta x_p/L)^{1/2}\right)\ \left(h(x_p)\ (\Delta x_p/L)^{1/2}\right) = d_i \qquad (2\text{-}10\text{-}a)$$

$$\sum_p A'_{ip}\ h'_p = d_i \qquad (2\text{-}10\text{-}b)$$

The reason for splitting the Δx evenly between the kernel and the unknown vector is that the 2-norm of the unknown is minimized in all methods discussed here, hence the 2-norm of the vector h' must be approximately the same as the 2-norm of the function h(x)

$$\sum_p |h'_p|^2 = \sum_p |h_p|^2\ \frac{\Delta x}{L} \cong \int_p |h_p|^2$$

The second restriction is that the chosen Δx must be small enough to approximate the forward problem correctly, i.e., so that if the "true" but unknown h were replaced into 2-10, the g_i that one would compute from 2-10 would differ negligibly from the g_i that one would compute from 2-5; 'negligibly' should be interpreted as 'the error due to the approximation 2-10 must be much smaller than the error in the measurements $\tilde{g}_i$'.

2.4 RESOLUTION, NOISE, AND EXPECTED ERROR

All the approximate solutions to (2-6) or (2-7) summarized here are linear (due to the choice of norm). Any estimate $\hat{m}_p$ obtained as a linear combination of the $\tilde{d}_i$ has the form

$$\hat{m}_p = \sum_{i=1}^{N_d} B^*_{pi}\ \tilde{d}_i = \sum_i B^*_{pi}\ \tilde{d}_i \qquad (2\text{-}11)$$

$\sum_i$, $\sum_{i'}$ will denote sums over all available data throughout the remainder of the chapter.

To avoid cumbersome repetitions, all equations that follow are written for a function m, hence involve $\int_p$. They apply as written to discrete parameters, if $\int$ is replaced by $\sum$. In a few instances the difference between forms is critical and will be pointed out in the text.

For any linear solution, equations 2-6, 9 and 11 imply

$$\hat{m}_p = \int_{p'} \hat{I}^*_{pp'}\, m_{p'} + \sum_i B^*_{pi}\, e_i \qquad (2\text{-}12)$$

$$\hat{m}_p - m_p = \int_{p'} (\hat{I}_{pp'} - I_{pp'})^*\, m_p + \sum_i B^*_{pi}\, e_i \qquad (2\text{-}13)$$

In both the discrete and continuous cases, the resolution operator $\hat{I}$ is defined as the result of applying the inverse B to the forward operator A:

$$\hat{I}_{pp'} = \sum_i B^*_{pi}\, A_{ip} \qquad (2\text{-}14)$$

and the identity operator $I_{pp'}$ is a Kronecker delta for discrete m, and a Dirac delta for a function m.

Equation (2-12) is the fundamental relationship between what we wanted to find, m_p, and what we computed, $\hat{m}_p$; 2-12 is valid no matter how the coefficients B_{pi} were computed. The resolution operator $\hat{I}$ describes a systematic distortion in $\hat{m}_p$, produced because B is not the exact (left) inverse of A. This can be due both to the lack of enough data, and to the criterion by which B was chosen. For fixed p, $\hat{I}_{pp'}$ is a filter (discrete or continuous) through which we must view

m_p. This term makes $\hat{m}_p$ a weighted average of the $m_{p'}$, possibly scaled (i.e., the sum or integral of weights may not be equal to 1). Perfect resolution requires that $\hat{I}_{pp'}$ be the corresponding identity: a Kronecker or Dirac delta centered at p. The first term in 2-13 will be called 'resolution error' (it is also called 'omision error').

The second term in (2-12) is the error due to noise in the data and to modelling error (and influenced by the choice of B); this term will be called 'noise' (it is sometimes called "comission error") and must be treated statistically. Suppose we understand $\underline{e}$ well enough to compute its expected value and covariance matrix [E]

$$\langle e_i \rangle = 0 \; ; \quad [E]_{ii'} = \langle e_i e_{i'} \rangle \qquad 2\text{-}15$$

With (2-15), the expected value of $\sum_i B^*_{pi} e_i$ is also zero, and its covariance function E' is

$$E'_{pp'} = \sum_i \sum_{i'} B_{ip} \, E_{ii'} \, B^*_{ip} \qquad 2\text{-}16$$

If the e_i were stationary and had a gaussian distribution, equation (2-16) would describe the noise in $\hat{m}$ completely, hence only 2-16 would need to be minimized in order to bring noise in $\hat{m}$ down to acceptable levels. This is the usual assumption, based on the property of sums of random numbers with any probability density function (pdf) to approach a gaussian pdf, and it is the one followed in the rest of this summary. But it is not always a good assumption. Claerbout and Muir (1973) give good examples of the effect of blunders in the data, and of the robust properties of minimizing a 1-norm, rather than a sum of

squared errors as (2-16) is.

Suppose 2-13 is multiplied by its complex conjugate; we then take expected values over the noise process. This yields a formal equation for the expected error in $\hat{m}_p$

$$\langle|\hat{m}_p-m_p|^2\rangle = M_{pp} + \sum_i \sum_{i'} \int_{p'} \int_{p''} B_{pi}\, A_{ip'}\, M_{p'p''}\, A^*_{i'p''}\, B^*_{pi'} - \tag{2-17}$$

$$- 2 \sum_i \int_{p'} B_{pi}\, A_{ip'}\, M_{p'p} + \sum_i \sum_{i'} B_{ip}\, E_{ii'}\, B^*_{i'p}$$

to write 2-17, we called

$$M_{pp'} = m_p\, m_{p'} \tag{2-18}$$

Without further assumptions, 2-17 is useless, because the covariance M as defined in 2-18 is unknown.

The statistical methods (Liebelt, 1967; Moritz, 1978) start with the assumption that the m_p themselves are random numbers. Hence, one also takes expected values over the ensemble of m in 2-17 (an equation that retains its form if $\langle m_p\, e_i\rangle=0$). This yields the expected value $\langle m_p\, m_{p'}\rangle$ which can, presumably, be estimated from data that we do not wish to include in the inversion; this subject will be expanded later.

In this section we have summarized the functional description of the resolution error (equation 2-13) and the statistical description (equation 2-17). In practice both lines of reasoning have many connections, as will become obvious later, and equations (2-13) and (2-17) simply describe two aspects of the same problem.

2.5 THE STATISTICAL METHODS.

The statistical approach described in this section has a long history of success when applied to a variety of geophysical problems: predictive decomposition of seismic signals (Robinson, 1954), interpolation of gravity anomalies (H&M (1967), chapter 7), design of oceanographic experiments (Bretherton et al., 1976), and others. The various names: discrete Wiener filtering, least squares collocation, optimal estimation, objective mapping, universal kriging, etc. refer to the same basic result (the Gauss-Markov theorem, equation 2-22). Each one has added peculiarities owing to the specific problem to which the technique was applied; for example, collocation uses a spherical geometry, where invariance under rotations must be required; the original Wiener filters were designed for continuous (rather than discrete) data, and requiring causality.

Two related arguments are frequently raised against the use of probabilistic methods: 1) that the attribution of statistical properties to the physical model is very artificial, for example, when density or seismic velocity vary with depth; 2) what is the physical meaning of an ensemble of earths that differ randomly from ours?.

In response to the first concern, Papoulis (1965) responds: " the student accepts readily this separation between the conceptual ... model and the physical world

for the so-called deterministic phenomena, but in probabilistic descriptions he confuses the two". A statistical model may accurately describe the density vs. depth profile, without implying that density is a random property in a philosophical sense. With appropriate constraints, the infinitely many realizations of a random process, one of which is the density profile we are interested in, are entirely equivalent to the infinitely many elements of a Hilbert space, one of whose elements is the density profile. In other words, they are all functions potentially capable of satisfying our data. These points will be expanded in this and following sections.

***

2.5.1 THE GAUSS-MARKOV RESULT

Both for discrete parameters m_p and for a function m, the Gauss-Markov theorem (Liebelt, 1967, chapter 5) applies, because in this chapter we assume the number of data is finite. When a function is the data, the formulation of Wiener and Kolmogorov (Liebelt, 1967, chapter 7) must be used. The book by Liebelt (1967) is an excellent introduction to this subject. A rigorous treatment of probability measures for infinite-dimensional spaces can be found in Wong (1971) and in Gihman and Skorohod (1974, chapters 5 and 8). Moritz (1980) and Luenberger (1967) have good summaries, with the added advantage that these books assume deterministic unknowns m, but Luenberger only applies the statistical method to a finite number of parameters.

Assume the measured values $\tilde{d}_i$ are random numbers with <u>any</u> probability density function (pdf), zero expected value, and with known covariance matrix $[\tilde{D}]$, whose elements are

$$\tilde{D}_{ii} = \langle \tilde{d}^*_i \, \tilde{d}_{i'} \rangle \qquad 2\text{-}19$$

assume <u>each</u> unknown m_p is also a random number, with zero expected value; assume each m_p is correlated to each $\tilde{d}_i$ with a known crosscovariance function C_{ip}

$$C_{ip} = \langle \tilde{d}^*_i \, m_p \rangle \qquad 2\text{-}20$$

We now seek the estimate $\hat{m}_p$ of each m_p that satisfies two properties: 1) $\hat{m}_p$ is a linear combination of the data; 2) $\hat{m}_p$ has smaller expected squared error than any other linear combination of the data:

$$\hat{m}_p = \sum_i B^*_{pi} \, d_i; \quad \langle |\hat{m}_p - m_p|^2 \rangle \quad \text{minimum} \qquad 2\text{-}21$$

The Gauss-Markov theorem states that $\hat{m}_p$ can be computed with

$$B^*_{pi} = \sum_{i'} C^*_{pi'} \, [D]^{-1}_{ii'} \qquad 2\text{-}22$$

where $[\ ]^{-1}$ is the inverse of the matrix in brackets.

Remarks:

** No assumptions were made about the pdf of either the $\tilde{d}_i$ or the m_p. This gives great generality to the result 2-22, but it should also be remembered that for a very general pdf the sum of squared errors may be the wrong quantity to minimize. Also, for an arbitrary pdf a nonlinear combination of the data may yield a smaller squared error than 2-22.

** If both the $\tilde{d}_i$ and the m_p have a gaussian pdf, then 2-22 has other desirable properties:

•• 2-22 yields a smaller mean squared error than any other estimate of m_p, linear or not (Van Trees, 1968, chapter 6).

•• 2-22 maximizes the likelihood, i.e., the probability of observing the actually observed d_i.

•• 2.22 maximizes the entropy or information content of the data (B. Cornuelle, 1982, pers. comm.)

** It is not necessary that the m_p and the $\tilde{d}_i$ be linearly related for the G-M result to hold, only that the covariances C_{ip} and $\tilde{D}_{ii}$ exist and be known, and that the matrix $[\tilde{D}]$ be invertible (in section 2-7 this constraint will be relaxed).

** When the m_p and the $\tilde{d}_i$ are linearly related, then the required covariances can be easily written in terms of $M_{pp'}$, the covariance of the m_p. Let

$$M_{pp'} = \langle m_p^* \, m_{p'} \rangle \qquad 2\text{-}23$$

When m_p and its conjugate are expanded in a series like 2-7, both series are multiplied together, expected values are taken, and the assumption is made that terms with different index ν are uncorrelated

$$\langle (m_\nu \, \psi_{\nu p})^* \, (m_{\nu'} \, \psi_{\nu p'}) \rangle = 0 \qquad 2\text{-}24$$

then 2-23 gives the spectral representation of M_{pp}

$$M_{pp'} = \sum_\nu M_\nu \, \psi_{\nu p} \, \psi^*_{\nu p'} \; ; \quad M_\nu = \langle |m_\nu|^2 \rangle \qquad 2\text{-}25$$

** Assuming linearity of the forward problem, and using the result 2-25

$$C_{ip} \equiv [AM]_{ip} = \sum_{\nu} A_{i\nu} M_{\nu} \psi_{\nu p} = \langle d^*_i m_p \rangle \qquad 2\text{-}26$$

$$D_{ii'} \equiv [AMA^T]_{ii'} = \sum_{\nu} A_{i\nu} M_{\nu} A^*_{i'\nu} = \langle d^*_i d_{i'} \rangle \qquad 2\text{-}27$$

(the notation in square brackets is a shorthand; for discrete m_p it indicates the required matrix operations; for a function m it indicates linear operators defined by 2-26, 2-27 (A^T is the adjoint of A).

The further assumption that the m_p are uncorrelated with data errors e_i, $\langle m^*_p e_i \rangle = 0$, yields

$$[\tilde{D}]_{ii'} = [D]_{ii'} + [E]_{ii'} \qquad 2\text{-}28$$

With 2-26 through 2-28, and assuming $N_p > N_d$, the Gauss-Markov result can be written

$$B^*_{pi} = \sum_{i'} [AM]^*_{i'p} ([AMA^T + E]^{-1})_{ii'} \qquad 2\text{-}29$$

For $N_p < N_d$, the classical least squares overdetermined case, the result is somewhat different from 2-29. See Luenberger (1969), chapter 4.

** The Gauss-Markov result in any of its forms yields a biased estimate ($\langle \hat{m}_p \rangle \neq \langle m_p \rangle$), unless one of the following is true:

$$\cdot \ \langle \hat{m}_p \rangle = 0 \qquad 2\text{-}30$$

$$\cdot \ \hat{I}_{pp'} = I_{pp'} \qquad 2\text{-}31$$

Condition 2-30 can be satisfied, for example, when the linear problem is the result of linearizing a nonlinear

problem about a reasonable initial value. Another case (that will appear in chapter 3) occurs when the < > operation is defined as an average over the volume of definition of m (this property, ergodicity, is discussed below), and the mean value of m is zero. Condition 2-31 can only be satisfied if the m_p are discrete parameters, and $N_p=K\leq N_d$ where K is the rank (e.g., Strang, 1980, chapter 2) of the A matrix in 2-6; in other words, when there are, at least, as many independent equations as there are parameters.

** Stationarity and ergodicity. Let $m_{p,k}$ indicate the k^{th} realization of a random process at point p, for an unknown function m. The expected value can be defined either as an average over the ensemble of realizations, or in terms of the pdf $\rho(m)$

$$\langle m_p \rangle = \lim_{N\to\infty} (1/N) \sum_{k=1}^{N} m_{p,k} = \int_{-\infty}^{\infty} m\ \rho(m)\ dm \qquad 2\text{-}32$$

To use the Gauss-Markov result for deterministic functions m, we must add two conditions to the probabilistic model

1) stationarity:

$$\langle m_p \rangle = \langle m_{p'} \rangle;\ \langle |m_p|^2 \rangle = \langle |m_{p'}|^2 \rangle;\ \ldots \qquad 2\text{-}33$$

for any p, p' where m_p is defined. When this condition is only true for the m_p themselves and their squares, but not for higher powers, the process is called 'weakly stationary' (see Wong (1971), chapter 2).

2) ergodicity

$$\langle m \rangle = (1/V) \int_V m_{p,k}\ dv(p) \qquad 2\text{-}34$$

$$\langle |m|^2 \rangle = \ldots$$

in words: the expected values can be computed as integrals over the volume of definition of m, on <u>any</u> realization k. This property requires stationarity. It is 2-34 that connects the Hilbert space approach described in the following sections and the probabilistic approach of this section; it states that the average properties of the (conceptual, not physical) ensemble $m_{p,k}$ must be the average properties of the only function we are interested in, m_p. To see this point, notice that 2-34 implies that the condition 2-24 on which later equations are based, is trivially satisfied.

** Assume m is ergodic, in the notation of 2-6

$$\langle m \rangle = \int_p m_p \qquad 2\text{-}35$$

Assume the forward problem is linear (equation 2-6). Then the minimum variance estimate of m_p among all <u>unbiased</u> linear combinations of the $\tilde{d}_i$ can be computed with

$$B_{pi'} = \sum_i ([AM]_{ip} + Q_{ip}) ([AMA^T + E]^{-1})_{ii'} \qquad 2\text{-}36$$

where $$Q_{ip} = a_i \Lambda_p, \quad a_i = \int_p A_{ip} \qquad 2\text{-}37$$

and Λ_p is a Lagrange multiplier that must be adjusted to satisfy the 'unimodular' condition

$$\int_{p'} I_{pp'} = \int_{p'} (\sum_i B_{pi} A_{ip'}) = \sum_i B_{pi} a_i = 1 \qquad 2\text{-}38$$

(the result is $\Lambda p = (1 - \underline{c}_p^T [\tilde{D}]^{-1} \underline{a}) / (\underline{a}^T [\tilde{D}]^{-1} \underline{a})$, where $\underline{c}_p$ is a vector with the C_{ip}, $[\tilde{D}]$ is the matrix of the $\tilde{D}_{ii'}$, and $\underline{a}$ is a vector with the a_i. C and $\tilde{D}$ were defined in 2-26 through 2-28).

** Note that $\langle|\hat{m}_p - m_p|^2\rangle$ is larger when the unbiased result 2-36 is used than when the original Gauss-Markov version 2-29 is used. A choice between these two boils down, in practice, to a matter of judgement, because $M_{pp'}$ is never known, but estimated from other data:

a) if we have a very good estimate of $M_{pp'}$ then obviously the G-M result 2-29 is preferable, because its rms discrepancy from the desired m_p is smaller. The fact that $\hat{m}_p$ cannot equal m_p --on the average-- is a small price to pay for higher overall accuracy.

b) if the estimate of $M_{pp'}$ is poor, then neither of the two estimates we are discussing has high overall accuracy; in this case I would choose the unbiased result, because at least one condition is satisfied (condition 2-38, which implies unbiasedness under the ergodic assumption, is satisfied even if a very poor estimate of $M_{pp'}$ is used in the computations). This point will be made again in section 2.8, where the Backus-Gilbert conditions are discussed.

2.5.2 EXAMPLE

A cursory overview of the application of these results to the example of section 2.2 completes this section

I.-DATA COVARIANCE. From the data d_i and from any other data of the same type that we have but do not want to use in the inversion (for example, because it would produce

a $[\tilde{D}]$ matrix too large to be inverted with the available computer), one computes the power spectrum $\tilde{D}_\nu$ of the data (the methods for one dimensional data are standard; see Bendat and Piersol (1971), section 9.6.2). The inverse Fourier transform of the $\tilde{D}_\nu$ is a piecewise constant estimate of the covariance function $\tilde{D}_{ii'}$ of the data. From this estimate we build the covariance matrix of the data vector $\underset{\sim}{\tilde{d}}$.

II.-ERROR COVARIANCE. We need the error covariance matrix of the data. Knowledge of the measuring process should provide the component due to measurement noise. To this term must be added the error due to the incomplete formulation of the forward problem. In our example, modelling error includes the residual power at long wavelengths owing to errors in the reference field, and the error of the 1-D approximation. When using GEM9, a reasonable estimate of this error accompanies the reference field. For lack of information, we must assume the error in the one-dimensional equation is negligible. Let us assume the error is stationary, i.e., the variance is the same for each $\tilde{d}_i$ and the correlation only depends upon the distance between the $\tilde{d}_i$. Then we can also compute the power spectrum E_ν of the errors.

III.-CROSSCOVARIANCE. To estimate C_{ip} the following steps can be used.

1) the coefficients of the forward kernel

(equation 2-4) are $A_{i\nu} = c\ |k_\nu|\ \exp(jk_\nu x_i)$, i.e.,

$$A_{i\nu} = A_\nu\ \Psi_{\nu i};\quad A_\nu = c\ |k_\nu|$$

hence, the covariance D can be written as

$$D_{ii'} = \sum_\nu \underbrace{A_\nu\ A^*_\nu\ M_\nu}_{D_\nu}\ \Psi_{\nu i}\ \Psi^*_{\nu i} \qquad 2\text{-}40$$

and the crosscovariance C can be written as

$$C_{ip} = \sum_\nu \underbrace{A_\nu\ M_\nu}_{C_\nu}\ \Psi_{i\nu}\ \Psi^*_{p\nu} \qquad 2\text{-}41$$

It follows that the coefficients of the crosscovariance can be computed from the already computed spectra of the data and of its noise

$$C_\nu = (\tilde{D}_\nu - E_\nu)\ A_\nu\ /\ |A_\nu|^2 \qquad 2\text{-}42$$

2) The inverse Fourier transform of the coefficients 2-42 gives a piecewise continuous estimate of the covariance function C_{ip}, from which the required values can be interpolated.

3) The $[\tilde{D}]$ matrix is formed with the values of $\tilde{D}_{ii'}$ previously computed, sampled at the appropriate data positions. $[\tilde{D}]$ is mathematicaly positive definite, because it is the sum of the error matrix [E], which is positive definite, and the errorless D, which is nonnegative definite (this point is expanded in section 2.6.2). To invert $\tilde{D}$, a Cholesky decomposition (e.g., Dongarra et al., 1979) is an accurate and efficient method, if indeed $\tilde{D}$ is computationally positive definite. The meaning and solution of computational singularities will be discussed in section 2.6.2.

4) The expected error of $\hat{m}_p$ can be computed using 2-17 with the quantities just defined. The resolution function has little meaning for this problem, because we can argue that it is a bandpass filter whose boundaries are the length L and the data spacing Δx.

2.6 HILBERT SPACE METHODS

In the statistical methods the unknown $\hat{m}$ is thought of as one realization of a random process, which is the set of infinitely many realizations having in common means, variances, etc. The pdf of the random process effectively summarizes the common features of all the realizations. In the Hilbert space methods, m is thought of as one element of a Hilbert space $\mathcal{M}$ that contains other functions $\bar{m}$; all the $\bar{m}$ are defined in the same volume, can be expanded in a series 2-7, and are at least as smooth as the desired m (a concept made precise below). As such, all the $\bar{m}$ are reasonable candidates to be on the right side of the forward equation 2-6, at least before the data are taken into account. The data values, $\bar{d}_i$, are also considered as elements of a finite dimensional Hilbert space $\mathcal{D}$ (a classical vector space).

The author's favorite introduction to the subject are chapters 3 to 5 of Lanczos (1961), but the modern use of reproducing kernel spaces must be sought elsewhere. Davis (1975, chapters 8,9,13 and section 12.6) is more concise,

discusses more properties of Hilbert spaces, but fails to emphasize the fundamental reciprocities between under and overdetermined problems. Luenberger (1968) is far more complete and concise than the preceding two; as such, it is an excellent reference, but also a little harder to follow as a primer. Moritz (1980) covers applications to a spherical geometry, compares statistical, Hilbert space and discrete methods, and does the above in a leisurely and didactical style. The article by Freeden (1981) summarizes many recent developments (reproducing kernel spaces, harmonic splines).

The key operation in both the $\mathcal{D}$ and $\mathcal{M}_6$ spaces is the norm: in $\mathcal{D}$ the norm of $(\tilde{d}-d)$ measures how close a vector d is to the data vector. In $\mathcal{M}_6$, the norm of an estimate $\hat{m}$ of m is used to choose the estimate with smallest norm -physically, the estimate with the minimum possible structure required by the data; e.g., Wunsch (1978). The inner product is the other essential property of Hilbert spaces, but this summary will circumvent the concept with algebra.

The norm of an element $\underline{d} \varepsilon \mathcal{D}$ will be computed with

$$\|\underline{d}\|_{E^{-1}} = \left(\sum_i \sum_{i'} d^*_i \, [E]^{-1}_{ii'} \, d_{i'} \right)^{1/2} = \left(\underline{d}^T \cdot [E]^{-1} \cdot \underline{d}\right)^{1/2} \qquad 2\text{-}43$$

Notice that 2-43 can also be written:

$$\|\underline{d}\|_{E^{-1}} = \left(([E]^{-1/2} \cdot \underline{d})^T \cdot ([E]^{-1/2} \cdot \underline{d})\right)^{1/2} = \left(\underline{d}'^T \cdot \underline{d}'\right)^{1/2}$$

The geometrical effect of $E^{-1/2}$ is to rotate the vector $\underline{d}$ into $\underline{d}'$, whose components refer to an orthonormal set of

axis. In this case, the components of $\underline{d}'$ have uncorrelated errors with unit variance, while those of $\underline{d}$ do not. In addition, if the errors of $\underline{d}$ are gaussian, those of $\underline{d}'$ also are.

The norm of $\bar{m} \varepsilon \mathcal{M}_0$ will be computed as follows

a) for discrete parameters m_p

$$\|m\|_{W^{-1}} = \left(\sum_{pp'} m^*_p \, [W]^{-1}_{pp'} \, m_{p'}\right)^{1/2} = \left(\underline{m}^T \, [W]^{-1} \, \underline{m}\right)^{1/2}$$

b) for a function m

$$\|m\|_{W^{-1}} = \left(\sum_{\nu=0}^{\infty} m^*_\nu \, m_\nu \, / \, W_\nu \right)^{1/2} \qquad 2\text{-}45$$

In 2-44, $W^{-1/2}$ usually provides a normalization (e.g., if the m_p have different units). In 2-45, the W_ν are mostly used to provide a physically required upper bound on the spectrum of m, and thus define the smoothness of the functions in the space $\mathcal{M}_0$ (only functions with finite norm are acceptable).

2.6.1 CONSTRAINED MINIMUM NORM

With these definitions, one chooses as optimum solution an $\hat{m}_p = \sum_i B^*_{pi} \, d_i$ such that

$\|\hat{m}_p\|_{W^{-1}}$ is minimum among all $\bar{m}$ that satisfy the data within a bound S:

$$\|A\hat{m} - d\|_{E^{-1}} = S < \overline{S} \qquad 2\text{-}46$$

(where $(Am)_i = \int_p A^*_{ip} \, m_p$ for a function m). See Moritz (1980) or Shure et al. (1982). Using a Lagrange multiplier μ^{-1} we can write

$$\text{minimize } \{ \|\hat{m}\|_{W^{-1}} + (1/\mu) \|A\hat{m}-d\|_{E^{-1}} \} \qquad 2\text{-}47$$

choosing μ to satisfy 2-46.

In order to write the solution to 2-47 for a function m, let

$$[AW]_{ip} = \sum_{\nu} A_{i\nu} W_{\nu} \Psi_{\nu p} \qquad 2\text{-}48$$

$$[AWA^T]_{ii'} = \sum_{\nu} A^*_{i\nu} W_{\nu} A_{i'\nu} \qquad 2\text{-}49$$

(when the m_p are discrete parameters, AW, AWA^T indicate the required matrix operations). With this notation, the result of minimizing 2-47 is:

$$B^{\blacktriangleright}_{pi} = \sum_{i'} [AW]^*_{i'p} ([AWA^T + \mu E]^{-1})_{ii'} \qquad 2\text{-}50$$

** 2-50 is similar in form to the Gauss Markov result 2-29, in spite of the fact that 2-29 optimizes <u>each</u> $\hat{m}_p$, while 2-47 is a global constraint on all $\hat{m}_p$.

** The essential difference is that the G-M result presumes we <u>know</u> the expected covariance $M_{pp'}$ (or its equivalent for a function, the expected spectrum M_ν); there is no parameter μ to be adjusted. In 2-50, W is arbitrary, within a class of reasonable constraints. Any (small) upper bound on the spectrum M_ν is a reasonable choice for W_ν. It is a subtletv of the Hilbert space formulation that the spectrum M_ν is not a valid choice of weights, because in these "units" the norm of m is ∞. Because W does not have to be closely related to m, the Lagrange multiplier μ must be chosen on the basis of the only known quantities, the $\tilde{d}_i$.

** The choice of μ is known as 'ridge regression' (Marquardt (1970); Lawson and Hanson (1974), chapter 25). The following summary is based on Shure et al. (1982). If the errors in the $\tilde{d}_i$ are gaussian, and if [E] is their covariance, then the expected value of $S(\mu)$ in 2-46 is N_d, the number of data. Calling

$$[AWA^T+\mu E]^{-1} = [Z]; \quad [Z]\underline{d} = \underline{h}$$

then S and $\partial S/\partial\mu$ can be written

$$S = \mu^2 \, \underline{h}^T \, [E] \, \underline{h} \qquad 2\text{-}51$$

$$\partial S/\partial\mu = 2\mu \, \underline{h}^T \, [AWA^T][Z][E] \, \underline{h} \qquad 2\text{-}52$$

Equations 2-51 and 2-52 can then be used in a Newton iteration to find a μ that makes $S=N_d$. The iteration is guaranteed to converge because all matrices in 2-52 are positive definite when $\mu>0$. $[AWA^T]$ may be positive semidefinite (see section 2.7), but $[AWA^T+\mu I]$ is mathematically positive definite. The range $0<\mu<\infty$ corresponds to $0<S<\underline{d}^T[E]^{-1}\underline{d}$; in that range, $\partial S/\partial\mu>0$, guaranteeing convergence of the Newton iteration.

***

An overview of the computational steps needed to solve the example of equation 2-5 using equation 2-50 concludes this section. The data are the $\tilde{g}_i$, and their error covariance E_{ij}, is presumed known (see section 2.5). The key choice to make is W. Because the coefficients h_k must decay at least like

$$|h_k| << c_k \, k^{-2} \text{ as } k \to \infty \qquad 2\text{-}53$$

for the g_i to be finite (the c_k are numbers of order 1), one could set $W_k = k^{-4.001}$. If the data are marine gravity, it is likely that the sources of significant anomalies are at least z_o=2 km below the surface. This allows us to impose a stronger constraint for short wavelengths:

$$|h_k| = c_k \exp(-|k|z_o) \text{ as } k \to \infty \qquad 2\text{-}54$$

(variations in water density, etc. become data noise). Hence

$$W_k = \exp(-|k|z_o) \qquad 2\text{-}54$$

is another valid choice. No units are necessary: μ is adjusted to yield the correct units. With the choice 2-54,

$$(AW)_{ip} = \sum_k |k| \exp(-|k|z_o \exp(jk(x_p - x_i)) \qquad 2\text{-}55$$

$$(AWA^T)_{ii'} = \sum_k |k|^2 \exp(-|k|z_o \exp(jk(x_i - x_{i'})) \qquad 2\text{-}56$$

Equations 2-55 and 56, the $\tilde{d}_i$ and their x_i, and $E_{ii'}$ are all the information required to solve 2-50; μ is chosen with 2-51, 52 (any singularities in $[Z]^{-1}$ are discussed in section 2-6-2).

Shure et al. (1982) have excellent examples of physically required upper bounds for the spectrum of the geomagnetic potential.

Because the W_ν are not necessarily close to the spectrum of m, the expected error equation 2-17 cannot be used. Our two choices are: a) find and describe the resolution function 2-14, or b) find an upper bound to $\|\hat{m}-m\|_{W^{-1}}$. Upper bounds, however, are almost always unrealistically large unrealistically large.

2.6.2 THE SINGULAR VALUE DECOMPOSITION

The singular value decomposition (SVD) underlies all Hilbert space methods; the SVD also allows us to do away with the often-seen requirement that all equations in 2-6 be linearly independent when $N_d < N_p$, so that the Gram matrix $[AWA^T]$ can be inverted. Even $[AWA^T+E]$, although always mathematically positive definite, may be computationally singular.

For finite N_p, the matrix version of 2-6-a,

$$\underline{d} = [A]\ \underline{m} \qquad 2\text{-}57$$

will be replaced by

$$[E]^{-1/2}\ \underline{d} = ([E]^{-1/2}\ [A]\ [W]^{1/2})\ ([W]^{-1/2}\ \underline{m}) \qquad 2\text{-}58\text{-}a$$

$$\underline{d}' = [A']\ \underline{m}' \qquad 2\text{-}58\text{-}b$$

In 2-58, [E] and [W] are positive definite matrices, hence $[W]^{-1}$, $[W]^{-1/2}$ are well defined. The meaning of [E] and [W] is the same as in section 2.6. For a function m, and in order to avoid describing the properties of reproducing kernel Hilbert spaces (see Freeden, 1981) the integral equation

$$d_i = \int_p A^*_{ip}\ m_p \qquad 2\text{-}59$$

will be replaced by

$$d_i = \sum_\nu A'^*_{i\nu}\ m'_\nu \qquad 2\text{-}58\text{-}c$$

with

$$d'_i = \sum_{i'} [E^{-1/2}]_{ii'}\ d_{i'} = E^{-1/2}\ d \qquad 2\text{-}60$$

$$m'_p = \sum_{\nu=0}^{\infty} \left(m_\nu\ /\ \sqrt{W_\nu}\right)\ \Psi_{\nu p} \qquad 2\text{-}61$$

$$A'_{ip} = \sum_{\nu=0}^{\infty} \sum_{i'} [E^{-1/2}]_{ii'}\ A_{i'\nu}\ \sqrt{W_\nu}\ \Psi_{\nu p} \qquad 2\text{-}62$$

$A_{i\nu}$ was defined in 2-8; m_ν in 2-7, $[E^{-1/2}]_{ii'}$ is an element of the $E^{-1/2}$ matrix. Equations 2-58 relate elements of a space $\mathcal{D}'$ ($d' \in \mathcal{D}'$) to those of a space $\mathcal{M}'$. A norm in $\mathcal{D}'$ will be computed as

$$\|d'\| = \sum_i |d'_i|^2$$

For function m, the norm in $\mathcal{M}'$ will be computed as

$$\|m'\| = \sum_\nu |m'_\nu|^2$$

it is clear that such norms correspond to weighted norms in the unprimed spaces.

The primes in d', A', m' will be dropped in the rest of this section, but the discussion will refer to the rotated equations 2-58, 2-59; m will be assumed to be a function; the changes needed for discrete m_p are trivial.

The function A_{ip} (for fixed i) as defined in 2-62 must have a finite norm in $\mathcal{M}'$, i.e.:

$$\sum_\nu |A_{i\nu}|^2 < \infty$$

In the example of section 2.2, the unrotated kernel of the integral equation is not even integrable; after a W_ν is introduced that requires the measured g_i to be finite everywhere, the new, "rotated" A_{ip} is square integrable.

The following results are based on Lanczos (1961, chapter 3 for finite N_p; chapters 4 and 5 for infinite N_p and N_d). See also Parker (1977) for infinite N_p and finite N_d.

Any element $d \in \mathcal{D}'$ can be expressed as a linear combination of N_d basis vectors $u_k = \{U_{1k}\ U_{2k} \cdots U_{N_d k}\}^T$; the u_k are the eigenvectors of the symmetric Gram matrix AA^T:

$$[AA^T]_{ii'} = \int_p A_{ip}\, A^*_{i'p} \qquad 2\text{-}63$$

$$\sum_{i'} [AA^T]_{ii'}\, U_{i'k} = \lambda^2_k\, U_{ik} \qquad 2\text{-}64$$

$$d_i = \sum_{k=1}^{K} \alpha_k\, U_{ik} + \sum_{K+1}^{N_d} \alpha_k\, U_{ik} \qquad 2\text{-}65$$

Equation 2-64 yields $K \leq N_d$ nonzero eigenvalues λ_k, and $N_d - K$ zero eigenvalues with their corresponding eigenvectors (let $\lambda_1 > \lambda_2 > \ldots > \lambda_K$; $\lambda_{K+1} = \ldots = \lambda_{N_d} = 0$). The u_k associated with nonzero eigenvalues (i.e., u_k, $k \leq K$) span a subspace $\mathcal{D}'_A$ of $\mathcal{D}'$; $\mathcal{D}'_A$ will be called 'activated subspace'. The u_k, $K < k \leq N_d$ span the 'null subspace' $\mathcal{D}'_o$. K is the rank of the matrix $[AA^T]$; K indicates the number of independent linear combinations one can form with the N_d equations 2-58. Equation 2-65 can also be written

$$\underline{d} = \underline{d}_A + \underline{d}_o \; ; \quad \underline{d}_A{}^T \cdot \underline{d}_o = 0 \qquad 2\text{-}65\text{-}b$$

with $\underline{d}_A \varepsilon \mathcal{D}_A'$, $\underline{d}_o \varepsilon \mathcal{D}'_o$. In other words: $\mathcal{D}'_A$ and $\mathcal{D}'_o$ are orthogonal complements.

A similar basis can be found in $\mathcal{M}$. Let (A^TA) denote the symmetric function

$$(A^TA)_{pp'} = \sum_i A_{ip}\, A^*_{ip'} \qquad 2\text{-}67$$

then any element $m \varepsilon \mathcal{M}'$ can be written in terms of the eigenfunctions v_k of the self-adjoint operator

$$\int_{p'} (A^TA)_{pp'}\, V_{kp} = \lambda^2_k\, V_{kp} \qquad 2\text{-}67$$

Equation 67 has an infinity of solutions v_k, but only K of these are associated with nonzero λ_k; the nonzero eigenvalues of 2-67 are equal to those of 2-64. The unknown m can be written as the sum of two orthogonal parts

$$m_p = \sum_{k=1}^{K} \beta_k V_{kp} + \sum_{K+1}^{\infty} \beta_k V_{kp} \qquad 2\text{-}68\text{-}a$$

$$m = m_A + m_o \qquad 2\text{-}68\text{-}b$$

(We assumed in section 2.3 that m could be expanded in an infinite series 2-7; this assures discrete eigenvalues in 2-67).

***

The key result of this section is the singular value decomposition of A:

$$A_{ip} = \sum_{k=1}^{K} U_{ik} \lambda_k V^*_{pk} \qquad 2\text{-}69$$

i.e., A_{ip} can be computed exactly using only the activated eigenvectors and eigenfunctions. Equation 2-69 has many consequences:

** The matrix or integral operator A, does not act upon the whole of $\mathcal{H}'$, but only upon components $\varepsilon\mathcal{H}'_A$, and does not yield as a result vectors 'anywhere' in $\mathcal{D}'$, but only those lying in $\mathcal{D}_A$.

** 2-69 implies that $\underline{d}$ does not have any component lying in $\mathcal{D}_o'$. The data, however, are in the vector $\tilde{\underline{d}} = \underline{d}+\underline{e} = \underline{d}+\underline{e}_A+\underline{e}_o$, where $\underline{e}_o \in \mathcal{D}_o'$. If the error vector does indeed have a nonzero component $\underline{e}_o$, then the system of equations 2-58 is incompatible (e.g., two different measurements of the same quantity). The error component $\underline{e}_A$ cannot be distinguished from 'legitimate' data, hence $\underline{e}_A$ will always map into erroneous components of $\hat{m}$.

** 2-69 implies that any component m_o ($m = m_A + m_o$, $m_o \in \mathcal{M}_o$) present in m has no expression in $\underline{d}$ (or $\tilde{\underline{d}}$); in other words,

$$\int_p A_{ip} (m_o)_p = 0$$

Conversely, with the given data $\underline{d}+\underline{e}$, nothing can be said about those parts of m lying in $\mathcal{M}_o'$ (e.g., $g(x_i)$ sampled every Δx gives no information about components of h whose wavelengths are shorter than $2\Delta x$).

The SVD inverse computes $\hat{m}$ with

$$B^*_{pi} = \sum_{k=1}^{K} V^*_{pk} (1/\lambda_k) U^*_{ik} \qquad 2\text{-}70$$

hence 2-70 only acts between the activated subspaces. Equation 2-70 can also be written as

$$B^*_{pi} = \lim_{\mu \to 0} \left(\sum_{k=1}^{(N_p, N_d)} V_{pk} (\lambda_k/(\lambda_k{}^2+\mu)) U^*_{ik} \right) \qquad 2\text{-}71$$

the upper limit of the sum is the smaller of N_p or N_d, but the choice is irrelevant because the $\lim_{\mu \to 0}$ operation eliminates all terms whose λ_k equals zero.

The relation between this and previous results can be seen by rewriting 2-71. When $N_p > N_d$, and $\mu \neq 0$, and we revert to our primed notation, then 2-71 can be written in the form

$$B'_{pi} = \lim_{\mu \to 0} \{ \sum_{i'} A'_{i'p} ([A'A'^T + \mu I]^{-1})_{ii'} \} \qquad 2\text{-}72$$

see Luenberger (1969, ch. 6) for some properties of 2-72. When $N_p < N_d$ (the overdetermined case) 2-71 is equivalent to

$$B'_{pi} = \lim_{\mu \to 0} \{ \sum_{p'} ([A'^T A' + \mu I]^{-1})_{pp'} A'_{ip'} \}$$

When 2-72 is written in terms of the 'unrotated' variables,

we obtain the now-familiar form

$$B'_{pi} = \lim_{\mu \to 0} \{ \sum_{i'} [AW]^{*}_{i'p} \, ([AWA^{T}+\mu E]^{-1})_{ii'} \} \qquad 2\text{-}73$$

Equation 2-73 is a mathematical definition, not an efficient computational algorithm. In practice, two alternatives have been used:

1) the 'tapered or damped' inverse: is 2-73 with a finite μ (i.e., the limit operation is eliminated). The names of this inverse derive from the effect of μ on the factor $\lambda_k/(\lambda_k^2+\mu)$ of equation 2-71. This is, of course, the same equation 2-50 previously discussed. The choice of μ was discussed in equations 2-51 and 2-52.

2) the 'truncated' inverse: follows the suggestion of Lanczos (1961, chapter 3) that any eigenvalue small enough to cause trouble when inverted, should be considered a zero eigenvalue, hence the whole term associated with this eigenvalue should be removed from the inverse. For $N_d << N_p$ this algorithm is implemented as follows:

a) all original variables are 'rotated' into primed variables.

b) the Gram matrix $[A'A'^{T}]$ is set up. Its eigenvalues and eigenvector are computed (e.g., routine EIGRS of the commercial package IMSL).

c) 2-70 is equivalent to

$$B^{*}_{pi} = \sum_{i'=1} A'^{*}_{i'p} \left(\sum_{k=1}^{K} U^{*}_{ik} \, (1/\lambda_k^2) \, U_{i'k} \right) \qquad 2\text{-}74$$

All terms in 2-74 with $\lambda_k^2 < \Lambda^2$ (whose choice is discussed below) are eliminated from the sum. Lawson and

Hanson (1974) point out that for a finite P which is not too large, it is more efficient to avoid building the Gram matrix. It is sufficient to build the A' matrix, and compute its singular value decomposition directly.

The limiting Λ is chosen with an argument similar to that used to choose μ. The data misfit can be written

$$S(K) = \sum_i \left| \sum_{k=1}^{K} U_{ik} \sum_i U_{i'k}\, d_{i'} - d_{i'} \right|^2 \qquad 2\text{-}75$$

Assuming the λ_k are ordered ($\lambda_1 > \lambda_2 > \ldots$), the rank K is chosen as the smallest K' for which S(K') achieves the value N_d, the number of data and the expected normalized misfit. Keeping more λ_k will only fit data noise better.

*****

2.6.3 THE BACKUS-GILBERT CONSTRAINT.

The resolution equation 2-12 will be needed again:

$$\hat{m}_p = \int_{p'} \hat{I}_{pp'}\, m_{p'} + \sum_i B^*_{pi}\, e_i \qquad 2\text{-}12$$

$$\hat{I}_{pp'} = \sum_i B^*_{pi}\, A_{ip} \qquad 2\text{-}14$$

The physical problem Backus and Gilbert (1968, 1970) dealt with was the distribution of densities and seismic velocities as a function of depth in the earth. They chose not to impose any constraint on the smoothness of these functions (other than requiring that they be square integrable). Let us say m is density as a function of depth.

Backus and Gilbert argued that an estimate $\hat{m}_p$ gave useful information about density at depth p if: a) $\hat{m}_p$ was a local average of m, and, b) the effect of noise e was as small as possible. They showed that it is not possible to optimize resolution and noise independently of each other. The Backus-Gilbert constraint, for <u>each</u> $\hat{m}_p$, can be written

$$\text{minimize} \left\{ \alpha \int_{p'} |\hat{I}_{pp'} - I_{pp'}|^2 + \beta \sum_i \sum_{i'} B^*_{pi} E_{ii'} B_{pi'} \right. \qquad 2\text{-}76$$

$$\text{subject to} \int_{p'} \hat{I}_{pp'} = 1 \qquad 2\text{-}77$$

$$\text{with} \quad \alpha + \beta = 1 \; ; \; \alpha, \beta \geq 0. \qquad 2\text{-}78$$

Setting $\alpha = 1$ yields the maximum resolution that the given data can provide for $\hat{m}_p$; setting $\beta = 1$ yields the minimum noise that $\hat{m}_p$ can have. Infinitely many choices of α, β yield an equal number of solutions, all equally 'optimum' in the sense that for a given acceptable noise level, 2-76 gives maximum resolution, and for a given acceptable resolution, 2-76 gives minimum noise level. Unless the constraint 2-77 is imposed, setting $\beta = 1$ yields $\hat{m}_p = 0$ (an estimate insensitive to data noise ...); indeed, for any value of β other than zero, 2-77 is required to yield an unscaled local average.

The integral in 2-76 is the so-called first Dirichlet criterion of deltaness. "Many such δ-ness criteria are available, so we are free to choose one which facilitates numerical computation" (Backus and Gilbert, 1968). This criterion requires only one matrix inversion for all $\hat{m}_p$.

Let $\mu = \alpha/\beta$, $\beta \neq 0$. The solution to 2-76 with 2-78 can then be written

$$B_{pi} = \sum_{i'} (A+Q)_{i'p} \, ([AA^T+\mu E]^{-1})_{ii'} \qquad 2\text{-}79$$

where

$$Q_{ip} = a_i \, \Lambda_p, \quad a_i = \int_p A_{ip}$$

and Λ_p is a Lagrange multiplier that must be adjusted to satisfy the 'unimodular' condition 2-77.

** In the original Backus-Gilbert papers no use is made of weighting functions in M. When this concept is introduced, the deltaness criterion must be changed in form because the weighting function $W_{pp'}$, for fixed p, is itself a reproducing kernel for all functions belonging to (see the article by Freeden, 1982). The same result can be obtained by 'rotating' m as was done in the previous section.

** The particular deltaness criterion used makes the form of equation 2-79 very similar to all previous results, in particular to the unbiased inverse (equation 2-36), obtained through statistical arguments.

** The fundamental difference between the Backus-Gilbert line of reasoning and the previous results, particularly those reached at by statistical methods, is this: there is no unique, overall best inverse. The statistical methods have one and only one answer, because the covariance M is presumed known. The minimum norm method of section 2.6.1 also yields only one answer, but many reasonable

weights W can be chosen. Backus and Gilbert invite the user to explore the range of possible solutions.

** If a unique answer is needed, the argument by which μ (i.e., α/β) was chosen in section 2.6.1 is still valid: the solution should not fit the data any better than the expected errors of the data predict, otherwise B is, most likely, reproducing noise.

***

2.7 CONCLUSIONS

All the methods discussed in this chapter are least squares methods, and the optimum inverses they yield have essentially the same form.

The fundamental difference between methods lies in how much we can assume known. In addition to the data d_i, the following are required, in order of increasing information

1) $E_{ii'}$: a covariance matrix describing both measuring noise of the $\tilde{d}_i$, and modelling error of the forward problem. This latter component is usually unknown.

2) W_ν : an upper bound on the spectrum of the unknown function m.

3) M_ν : the spectrum of the unknown function m.

If both 1 and 3 are known with reasonable accuracy, the Gauss Markov result 2-22 is applicable. Furthermore, in this case, valid expected errors (rms errors over the volume of definition of m) can be computed. This case occurs when very large amounts of data are available (e.g., satellite altimetry) because it is not possible to use all the data in one inversion; although a stepwise inversion can include more equations, the total number is still well below the millions of data values that a satellite can collect in its lifetime. The 'statistical' approach uses all the data to compute the spectra, and only some to provide the correct phases of the different basis functions in a small area.

If only 1 and 2 are known (e.g., Shure et al., 1982), then the 'damped' inverse is a better choice. Ridge regression provides a value of μ that scales the solution to fit the data only to the extent that their errors permit. The unbiased damped inverse is a better choice when the mean value of m is not zero, and W is far from M. If no constraint on m is known, the Backus-Gilbert philosophy applies, and all one can ask is: what is the range of possible solutions?. This approach is still applicable when more information is available.

The difference between an infinite dimensional formulation and a finite one has two different aspects.

1) inasmuch as all functions dealt with can be approximated arbitrarily closely by a finite vector (either Fourier coefficients, or discrete samples), the forward problem can always be replaced by a finite dimensional problem that is arbitrarily close to the original. Many times the forward kernel A_{ip} is only known at discrete points p (e.g., Wunsch and Grant, 1982), so little is gained by an infinite dimensional formulation whose kernel is unknown.

2) if the discretization chosen is too coarse or extreme (e.g., describing the ocean floor as three layers only) then the modelling error is large. Since the modelling error is always very difficult to describe, because it requires knowledge about m that one does not have, this error component should be kept small. If the modelling

error is large, but our description of it (i.e., the matrix E) is incorrect, then we are solving the wrong problem. In this sense, an infinite dimensional formulation (when the unknown is a function) is a safe way to keep the modelling error negligibly small.

The elegance of the theory of Hilbert spaces and the simplicity of the Gauss-Markov result should not obscure the fact that the only norm being minimized is the sum of (weighted) squares. A humorous "example" of the inapplicability of the L_2 norm to fit the data will be borrowed from Claerbout and Muir (1973): '... when a traveller reaches a fork in the road, the L_1 norm tells him to take one way or the other, but the L_2 norm instructs him to head off into the bushes'. The main requirement to fit the data in a least squares sense is the removal of outliers from the data set prior to inversion, because the L_2 norm does not yield 'robust' estimates.

The L_2 norm is not the only way to choose among different estimates of m either. Density inside the Earth is a positive quantity, but none of the methods discusses here addressed inequality constraints. The singular value decomposition allows one to add null space eigenfunctions v_k (which the data do not constrain) until an inequality constraint is satisfied. But linear programming -for finite N_p and N_d is a systematic way to incorporate inequality constraints, and minimize the 1-norm of m (see Sabatier (1977)).

Furthermore, statistical methods are well suited to include inequality constraints, inasmuch as the pdf of m can be set to zero in the forbidden regions. Tarantola and Valette (1982 a,b) have proposed -for finite N_p- to use the pdf of the data and the a-priori pdf of m to compute the pdf of m given the data. Such an approach is computationally very cumbersome, but it has the ability to solve nonlinear problems without linearization and gives an excellent description of the set of acceptable models.

CHAPTER 3

OPTIMUM GEOID ESTIMATION

3.1 INTRODUCTION

This chapter analyzes the construction of geoidal estimates from surface gravity data (it is a slightly revised version of Zlotnicki et al (1982)). At present, the analysis of satellite orbits has been able to yield the spherical harmonic coefficients of the earth's gravity field associated with wavelengths longer than 2000 km (such as GEM 9 and GEM10, Lerch et al., 1979) with more accuracy than surface gravity measurements can provide. Because of the height at which satellites fly and the exponential decrease of high frequency coefficients with height, coefficients in the wavelength range 2000-4000 km have large uncertainties and those with wavelengths shorter than 2000 km are practically unknown. In order to define these short wavelengths, surface gravity data must be used.

This chapter does not analyze the estimation of satellite-derived coefficients because they are always accompanied by realistic formal error estimates (see figure 1-1). The best gravimetric geoids, on the other hand, lack any error estimate, and their quality has only be assessed by their similarity to sea-surface topography, an argument

that breaks down when one is trying to recover oceanographic discrepancies between the two.

Geoidal heights and gravity accelerations are both expressions of the earth's gravity field; as such, one of them can be computed if the other is known fully. To compute a geoidal height at just one point, however, knowledge of gravity accelerations over the entire surface of the Earth is required. Unfortunately, gravity measurements are always discrete and noisy, and entire regions of the earth lack any data. Using the methods discussed in chapter 2, we can obtain optimum estimates of geoidal heights from any discrete distribution of data. What is more important in order to extract small signals, realistic estimates of the systematic error owing to the incompleteness of the data set (the "omission" or "resolution" error) and the uncertainty owing to noise in the measurements can be given. The equations for the omission error are valid for any other computational scheme that involves a linear combination of gravity data, such as the non-optimum method actually used to compute a North Atlantic geoid in chapter 4.

The goal of inferring the geoid from gravity is not new, and four methods have been used to solve it approximately: Stokes integral, Molodenskii's series, modified Stokes integrals, and least squares collocation.

Stokes integral [Stokes, 1849; Heiskanen and Moritz, 1967 (hereinafter H&M)] is the exact solution to the problem of computing the disturbing potential (proportional to geoidal height, within excellent accuracy), assuming exact

gravity data are available everywhere on a sphere and no masses exist outside this sphere. A discrete version of Stokes integral has been the most commonly used procedure, usually with no analysis of the omission error (but Sjoberg [1979] discussed the error for a particular distribution of data). Marsh and Chang [1978] computed a detailed local geoid with a variation of this technique. A long wavelength error (usually a strong bias and a trend) are the result of integrating only over a fraction of the Earth.

Molodenskii et al. [1967] break up the integration over the sphere into two terms: an integral over a small spherical cap around the point of interest and a series in 'truncation' functions for the remaining effect. This approach shows that short wavelength information is contributed by data at neighboring points and long wavelength information by distant data; it also allows one to evaluate the omission errors due to lack of data outside the spherical cap. Jekeli [1979] evaluated omission errors in this manner, assuming perfect data (i.e., continuous and noiseless) around the point of interest.

Molodenskii's expansion can be used to find modifications of Stokes kernel that reduce the error due to lack of remote data. Optimum algorithms for continuous point data inside a spherical cap were discussed by Jekeli [1981], based upon a minimization also due to Molodenskii et al.

[1967]. See also Rapp [1980]. These methods do not address discrete (and irregularly spaced) data and require the point of interest to be in the center of a circle of data.

Least squares collocation [Moritz, 1978, 1980] has been reviewed in Chapter 2. It is closely related to the Backus-Gilbert method, but they differ in the error analysis: collocation depends upon knowledge of the covariance function of the model, whereas the Backus-Gilbert method emphasizes the resolution function, which retains its significance even when nothing is known about the expected behaviour of the geoid. Both these methods are more accurate and data adaptive (for real data sets) than those described above but they require more expensive computations. Although this chapter emphasizes optimum methods, the geoid computation of Chapter 4 was performed with a suboptimum modified Stokes kernel.

All these methods include other assumptions whose violation is a source of systematic error.

1.-Usually a spherical geometry is used; when an integral such as Stokes' is used to compute geoidal heights, the gravity data are assumed to lie on a sphere (the optimum methods do not require this assumption). The 21 km difference between equatorial and polar radii introduces a long wavelength error of order f=1/298. Rapp (1981) derived and computed the corrections needed when a partial Stokes integration

is performed over a cap. For a 10^{o} cap (which is used in chapter 4) and the unmodified Stokes kernel, the maximum and rms corrections needed are (-26 cm, ±6 cm); using a modified Stokes kernel the corrections were slightly lower. For larger caps the error increases because the integration relies more on gravity data for its long wavelength information: for an integral over the whole Earth the corrections have (-59 cm, ±18 cm) values. Although these numbers are relatively low, they appear as an E-W slope that can be easily mistaken for the longer scales of the general circulation.

2.-The other major assumption is the absence of masses outside the geoid (needed for Laplace's equation to remain valid). Continents, islands, and the whole atmosphere are violations of this assumption. According to H&M (section 8-2) this error can be corrected by a 3-step process:

-remove from measured gravity the attraction of the exterior masses M_E, either by moving them out to infinity (mathematically, of course) or by pushing them inside the Earth.

NOTE: in this chapter, geoidal heights will be denoted by N, as is traditional in the geodetic literature. In all other chapters, they are denoted γ.

$$g'_i = g_i - \Delta g_i(M_E)$$

-compute from g' the geoid N' of an Earth lacking M_E (called the 'cogeoid'), for example with Stokes integral.

-add to N' the geoidal effect of the external masses, $\Delta N(M_E)$, usually called 'indirect effect'

$$N_p = N'_p + \Delta N_p(M_E)$$

The main practical problem encountered in computing these corrections is the lack of knowledge about M_E. One can easily measure the topography of a mountain, but it is much more difficult to obtain its density. H&M (sections 3-6 and 3-7) argue that the best among the existing corrections is the 'condensation' correction, which compresses M_E onto a surface layer on the geoid, because both $\Delta g(M_E)$ and $\Delta N(M_E)$ are small, hence the error introduced by inaccurate knowledge of M_E is a fraction of a small number. For the atmosphere $\Delta N(M_E)$ varies with the seasons between -0.1 and -1.3 cm, and $\Delta g(M_E)$ is about 0.87 mgal (Christodoulidis, 1979). For mountains or islands, ΔN is of order 1 m per 3 km of average elevation of M_E. Over the oceans this effect is important only near coastlines.

3.2. FORMULATION OF THE SPHERICAL PROBLEM

This section derives and discusses various forms of the basic integral equation needed for the inversion. A spherical geometry is used.

The total potential $V(\theta,\lambda,r)$ of the earth's gravity field is split into a reference component U and a disturbing potential T. U is the attraction of a field of which the reference ellipsoid is an equipotential surface (H&M).

$$V(r,\theta,\lambda) = U(r,\theta,\lambda) + T(r,\theta,\lambda)$$

Here r is radial distance as a fraction of the radius of the limiting sphere, θ and λ are colatitude and longitude, respectively.

The required solution of Laplace's equation for T, in terms of spherical harmonics, valid on and outside the sphere, and assuming no masses exist outside this sphere, is (H&M)

$$T(r,\theta,\lambda) = \sum_{n=0}^{\infty} \sum_{m=0}^{n} (a_{nm} \cos(m\lambda) + b_{nm} \sin(m\lambda))P_{nm}(\cos\theta)/r^{n+1}$$

where P_{nm} are fully normalized Legendre functions and a_{nm} and b_{nm} are real coefficients. For notational convenience we use complex coefficients T_{nm} and functions Φ_{nmp} (Jackson [1975] but with the normalizations of H&M; the asterisk denotes complex conjugate)

$$T_{nm}(r) = (-1)^m T^*_{n,-m}(r) = (1/\sqrt{2})(a_{mm} - ib_{nm}) \;/\; r^{n+1} \qquad (3\text{-}1)$$

$$\Phi_{nmp} = (-1)^m \, \Phi^*_{n,-mp} = (1/\sqrt{2}) P_{nm}(\cos \theta_p) \exp(jm\lambda_p)$$

$$j=\sqrt{-1}$$

so that the expansion of T reads

$$T(r,\theta_p,\lambda_p) = \sum_{n=0}^{\infty} \sum_{m=-n}^{n} \Phi_{nmp} \, T_{nm}(r) \qquad (3\text{-}2)$$

The coefficients T_{nm} satisfy the inverse transformation

$$T_{nm}(r) = \int_p T(r,\theta_p,\lambda_p) \, \Phi^*_{nmp} \qquad (3\text{-}3)$$

where $\int_p (\cdot) = \int\int(\cdot)d\sigma_p$ recovers the mean value of the integrand on the sphere of radius r (the element of area $d\sigma_p$ around point p is measured as a fraction of the area of the sphere). The geoidal heights, N, are computed from Brun's approximate formula (H&M)

$$N_p = T(1,\theta_p,\lambda_p)/\gamma \qquad (3\text{-}4)$$

where γ is the mean value of gravity acceleration on $r = 1$. Equation 3-4 is a spherical approximation. The ellipsoidal relationship can be found in Moritz (1980, section 39). Equation (1) becomes the spherical harmonic expansion of N when r=1, after dividing by γ:

$$N_{nm} = T_{nm}(1)/\gamma \qquad (3\text{-}5)$$

Studies of gravity acceleration customarily use gravity 'anomalies' (see below), but this discussion will be restricted to gravity 'disturbances'; use of the former requires only minor changes. The gravity disturbance, g, is defined as

$$g(r,\theta_i,\lambda_i) = -(\partial T/\partial r)|_{r,\theta_i,\lambda_i} = \sum_n \sum_m g_{nm}(r)\Phi_{nmi} \quad (3\text{-}6)$$

The coefficients in the expansion of gravity and the geoid are related by

$$g_{nm}(r) = (\gamma/R)\,[(n+1)\,/\,r^{n+2}]\,N_{nm} \quad (3\text{-}7)$$

(R=6371 km is a mean earth radius). If we wish to assume that gravity data are given on the surface of the sphere r=1, then 3-7 becomes

$$g_{nm}(1) = (\gamma/R)\,(n+1)\,N_{nm} \quad (3\text{-}8)$$

(both gravity anomalies and disturbances are defined as $\partial/\partial r(V_P - U_{P'})$; for disturbances P'=P; for anomalies P' and P are on the same vertical but at different heights, with P' on the reference ellipsoid. When gravity anomalies are on the left hand of (7) and (8), (n+1) on the right must be replaced by (n-1), and both equations are valid for $n > 2$. See H&M).

To use gravity measured on r=1 we would have to back transform equation (8). Clearly, the kernel whose coefficients are (n+1) is singular (it is unbounded at the origin, H&M, equation (1-96)). We know, however, that g has finite energy, hence g_{nm} must fall off as $|g_{nm}(1)|^2 << n^{-2}$ as $n\to\infty$, and N_{nm} must decay even faster. The difficulty, therefore, lies only in the very high degrees, which the kernel attempts to amplify n+1 times, whereas neither g nor N have significant energy at those high degrees. For this reason all terms with $n > \bar{n}$, a very large but finite cutoff will be deleted from the formulation.

We define a function m (the model) that contains all spherical harmonic terms of N, in the range $-n_1 \leq n \leq \bar{n}$. The inversion procedure of section 3 will attempt to estimate m rather than N. Let

$$d_i = \sum_{n=n_1}^{\bar{n}} \sum_{m=-n}^{n} g_{nm}(1)\,\Phi_{nmi}$$

$$A_{ip} = \sum_{n=n_1}^{\bar{n}} \sum_{m=-n}^{n} A_n\, \Phi^*_{nmi}\, \Phi_{nmp} \qquad A_n = \frac{\gamma}{R}(n+1)$$

$$m_p = \sum_{n=n_1}^{\bar{n}} \sum_{m=-n}^{n} N_{nm}\, \Phi_{nmp}$$

Assume first that $n_1=0$. Because g and N are bounded and square integrable, d and m can approximate them as closely as desired by taking $\bar{n}$ sufficiently large. It is enough to choose $\bar{n}$ so that the remainder in the series for d is much smaller than measurement noise. Measurements of g can then be interpreted as measurements of d. We also remove terms with $n<n_1$ because analysis of satellite orbit perturbations can provide these components of g and N, and their removal allows us to use fewer data in the inversion.

Although $\bar{n}$ eliminates the mathematical singularity, A_{ip} can still be larger than the computer can represent, or produce instabilities in the inversion (large changes in m due to minor perturbations in the data).

This 'computational singularity' can be avoided either

by filtering d or by weighting m; at least one of these alternatives must be used. We multiply (8) by the coefficients F_n of a filter and apply a weighting W_n, both functions of angular distance only (hence with no order m dependence), and both $F_n \to 0$, $W_n \to 0$ as $n \to \infty$.

$$F_n G_{nm}(1) = \left(\frac{\gamma}{R} F_n (n+1) W_n^{1/2}\right) (W_n^{-1/2} N_{nm}) \qquad (3\text{-}9)$$

Applying the inverse transform to (9) yields

$$(Fd)_i = \int_p (FAW^{1/2})_{ip} (W^{-1/2}m)_p \qquad (3\text{-}10)$$

where we have called

$$(Fd)_i = \sum_{n=n_1}^{\bar{n}} \sum_{m=-n}^{n} F_n G_{nm}(1) \Phi_{nmi} \qquad (3\text{-}11)$$

$$(FAW^{1/2})_{ip} = \sum_{n=n_1}^{\bar{n}} F_n A_n W_n^{1/2} \frac{1}{\xi_n} P_n(\cos \psi_{ip}) \qquad (3\text{-}12)$$

$$(W^{-1/2}m)_p = \sum_{n=n_1}^{\bar{n}} \sum_{m=-n}^{n} W_n^{-1/2} N_{nm} \Phi_{nmp} \qquad (3\text{-}13)$$

where ψ_{ip} = angular distance between points i and p, $\xi_n = 1/\sqrt{2n+1}$. See also Appendix 1.

We first discuss filtering only (setting $W_n=1$ in (9), all W disappear in (10)-(13)). An example at hand is equation (7), where $1/r^{(n+2)}$ acts as a filter by attenuating high degree components of g as distance to the earth increases; this filter appears because of the physical constraints on g. Another type of filter is the degree average [e.g.,

Rapp, 1978] but its coefficients do not decay with n fast enough to make (12) computationally efficient when $W_n = 1$. For our examples we will use a spherical equivalent of the Gaussian filter, described in Appendix 3 and characterized by its half width ψ_o. When a filter that makes 3-12 converge is used, the artificial cutoff at n becomes unnecesary. Filtering data is a standard procedure for reducing both noise and the number of values that must be handled. The additional advantage here is the reduction of high degree content in (FA). The main disadvantage is that a sampling error is introduced, because we assume $(Fd)_i = \int_{i'} F_{ii'} d_{i'}$ with $\&_{i'} F_{ii'} = 1$ but we must use

$$(\tilde{Fd})_i = \sum_{i'} F_{ii'} \tilde{d}_{i'} \Big/ \sum_{i'} F_{ii'}$$

where $\sum_{i'}$ is a sum over the available data. The tilde on $\tilde{d}$ indicates that it contains measurement noise, but that over $(\tilde{Fd})$ reflects both the effect of noise and incomplete sampling. (The integral over i' need not cover the sphere if $F_{ii'}$ is negligibly small for $\psi_{ii'} > \bar{\psi}$).

We now focus on W alone (setting F_n=1 in (9) eliminates F from (10)-(13)). Because $W_n \to 0$ as $n \to \infty$, high degrees are eliminated from $(AW^{1/2})$, thus removing the singularity. Its effect will be discussed further in later sections. W may be any reasonable upper bound on the power spectrum of N; when used, it becomes unnecesary to cut off the sums at $\bar{n}$.

But I feel that the high degree behaviour of N has not yet been adequately modelled, so that cutting off at some high degree is the simplest possible description, and the one whose error (the rms power above $\bar{n}$) is easiest to describe.

The computations of section 3 require such functions as $(FAWA^TF^T)$. We list here their Legendre series for reference. See also Appendix 1.

$$
\begin{aligned}
A_n &= \frac{\gamma}{R}\,(n+1) \\
\xi_n &= 1/\sqrt{2n+1} \\
\sum &= \sum_{n=n_1}^{\bar{n}} \\
W_{pp'} &= \sum W_n \,\frac{1}{\xi_n}\, P_n(\cos\psi_{pp'}) \\
(FAW)_{ip} &= \sum F_n\, A_n\, \xi_n \,\frac{1}{\xi_n}\, P_n(\cos\psi_{ip}) \\
(FAWA^TF^T)_{ii'} &= \sum F_n{}^2\, A_n{}^2\, W_n \,\frac{1}{\xi_n}\, P_n(\cos\psi_{ii'})
\end{aligned}
\qquad (3\text{-}14)
$$

In these equations it is easy to take into account gravity accelerations measured at different distances from the Earth's center, thus eliminating the need for free air or ellipsoidal corrections: one simply replaces the kernel coefficients by $A_n=(\gamma/R)\,(n+1)\,r^{-(n+1)}$ (and use the ellipsoidal approximation to Brun's formula).

The same reasoning described in this section can be used to pose integral equations for gravity accelerations, given geoid data from satellite altimeters. We need only exchange m and d in (11) and 13), and replace $(n+1)\gamma/R$ by its reciprocal in (12) and (14). Truncation is still needed

because the kernel whose coefficients are 1/(n+1) (similar to Stokes function, H&M, equation (2-169)) is also singular because of its high degree content. Filtering and weighting operate in the same manner.

3.3. RESOLUTION AND NOISE

Equation (10) can be written, after an obvious change of variables,

$$d_i = \int_p A_{ip}\, m_p \qquad (3\text{-}15)$$

where A_{ip} is a square integrable function, as are d and m. The whole machinery of chapter 2 can now be applied.

Integrals are taken over the surface of the sphere, and they recover the mean value of the integrand.

We assume N_D measurements $\tilde{d}_i$ are available,

$$\tilde{d}_i = d_i + \varepsilon_{d_i} \qquad i = 1,2,\ldots, N_D \qquad N_D << N_P \qquad (3\text{-}16)$$

where ε_{d_i} are measurement errors, and N_P is the dimension of the space in which m lies.

We compute $\hat{m}_p$ with

$$\hat{m}_p = \sum_i B_{pi}\, \tilde{d}_i \qquad (3\text{-}17)$$

(again indices i, i', ... indicate data positions and p, p', ... for model positions). Notice that B_{pi} can be Stokes kernel sampled at the data positions i, any of its modifications, or the optimum inverses. The fundamental relation between, $\hat{m}$ and m (equation 2-12) follows

$$\hat{m}_{p_o} = \int_p \hat{I}_{p_o p}\, m_p + \sum_i B_{p_o i}\, \varepsilon_{d_i} \qquad (3\text{-}18)$$

where

$$\hat{I}_{p_o p} = \sum_i B_{p_o i}\, A_{ip} \qquad (3\text{-}19)$$

Data noise ε_d is described by its covariance matrix $[E_d]$,

$$[E_d]_{ii'} = \langle \varepsilon_{d_i} \varepsilon_{d_{i'}} \rangle \qquad \langle \varepsilon_{d_i} \rangle = 0$$

and its effect on m is described by the covariance function E"

$$E"_{p_o p} = \sum_i \sum_{i'} B_{p_o i}\, [E_d]_{ii'} B_{pi'}; \quad \sigma"_{p_o} = \sqrt{(E"_{p_o p_o})} \qquad (3\text{-}20)$$

Using the definition of ϕ_{nmp} in (1), and calling M_{nm} the Legendre coefficients of m_p, the resolution error becomes

$$\varepsilon'_{p_o n_1}^{n_2} = \sum_{n=n1}^{n_2} \sum_{m=-n}^{n} m_{nm} (\hat{I}^*_{p_o nm} - \Phi_{nmp_o}) \qquad (3\text{-}21)$$

where

$$\hat{I}_{p_o nm} = \sum_i B_{p_o i}\, A_n\, \Phi^*_{nmi}$$

$$\hat{I}_{p_o p} = \sum_n \sum_m \hat{I}_{p_o nm}\, \Phi_{nmp} \qquad (3\text{-}22)$$

Because we are not interested in reproducing low degrees $n<n_1$ (provided by the analysis of perturbation to satellite orbits) we will consider m a useful estimate of m if a range $n_1 \leq n \leq n_2$ exists over which $\varepsilon'_{p_o n_1}^{n_2}$ is small. If we do not know the M_{nm} we must require

$$\hat{I}^*_{p_o nm} \simeq \Phi_{nmp_o} \qquad \begin{matrix} n_1 \leq n \leq n_2 \\ -n \leq m \leq n \end{matrix} \qquad (3\text{-}23)$$

The right hand side of (23) is the spherical harmonic coefficient of the identity operator (Dirac delta) on the sphere, centered at p_o (Appendix 1).

Direct verification of condition (23) for all n, m of interest would be prohibitively expensive. We will use two alternatives: looking at the degree variances of $\hat{I}$, and estimating the rms value of the error (21) betwen n_1 and n_2.

A necessary condition on the degree variances can be obtained from (23) (see also Appendix 2)

$$|\hat{\delta}_{p_o n}|^2 = \sum_{m=-n}^{n} |\hat{I}_{p_o nm}|^2 \simeq 2n+1 \qquad n_1 \leq n \leq n_2 \qquad (3\text{-}24)$$

(necessarily, 2n+1 are the degree variances of a Dirac delta). This condition is not sufficient for (23); it is also necessary to ascertain whether $\hat{I}_{p_o p}$ is peaked at p_o, indication that many terms in the expansion of $\hat{I}$ are in phase at that point.

3.3.1 RMS RESOLUTION ERROR

The error analysis described before has a fundamental advantage: it requires no knowledge about the model m. For example, if condition (23) is satisfied to a specified accuracy for all n between n_1 and n_2, if the degree variances are negligibly small when n is not between n_1 and n_2, and if these hold for all p_o inside some area, we can

claim that $\hat{m}$ is a band-passed version of m inside the same area, a most useful property.

An estimate of the spectrum of m is needed, however, if we want an rms value of the systematic omission error in the proper units. Given two points inside the data region, p_1 and p_2, we imagine their error computed from (25) (we really do not know all the M_{nm}). We then 'slide' the data distribution to another place in the earth, where p_1 and p_2 have different latitude and longitude, but the same position relative to the data points; again we compute their errors. The average over the earth of the products $\varepsilon_p \varepsilon_p$ is defined as the covariance of the omission error between p_1 and p_2. It is difficult to compute it unless we also average over all azimuths, a more meaningful operation if m is isotropic. With this assumption

$$\langle |M_{nm}|^2 \rangle = \frac{|M_n|^2}{2n+1} \quad \text{for all orders } m.$$

Therefore, multiplying (25) by its complex conjugate, taking 'expected values' and summing between n_1 and n_2, one obtains the rms error equation 2-17 [e.g., Moritz, 1976)]

$$E'_{p_o p}\Big|_{n_1}^{n_2} = \sum_i \sum_{i'} B_{p_o i} (ACA^T)_{ii'} \Big|_{n_1}^{n_2} B_{pi'}$$

$$+ C_{p_o p}\Big|_{n_1}^{n2} - \sum_i B_{p_o i}(AC)_{ip}\Big|_{n_1}^{n2} - \sum_i B_{pi}(AC)_{ip_0}\Big|_{n_1}^{n2} \qquad (3\text{-}25)$$

where $E'\Big|_{n_1}^{n_2}$ is the component of the covariance function of the resolution error with $n_1 \leq n \leq n_2$, $C\Big|_{n_1}^{n_2}$ is the same component of the expected covariance function of m; $(AC)\Big|_{n_1}^{n_2}$ is the same component of the forward kernel function weighted with C; $(ACA)\Big|_{n_1}^{n_2}$ is the same component of the inner product function also weighted with C. Formulas for these are given in section 2; the Legendre series are summed between n_1 and n_2 (in our computations we made tables with these functions, from which the needed values were later interpolated.

The total rms error of m is obtained by summing (25) and (20). The limits n_1 and n_2 in (25) will be used to study the error in a particular frequency band. The total error of a point estimate requires $n_2 \to \infty$.

3.3.2 INVERSE OPERATORS

The three inverse operators that will be used are of the form

$$B_{pi} = \sum_{i'} A_{i'p} \, [AA^T]^-_{ii'} \qquad (3\text{-}26)$$

where $[AA^T]^-$ is a generalized inverse of the inner product or Gram matrix (see section 2.6.2).

The singular value decomposition (SVD) inverse of AA^T is (equation 2-72)

$$[AA^T]^-_{ii'} = \sum_{k=1}^{K} U_{ik} \, \lambda_k^{-2} \, U_{i'k} \qquad (3\text{-}27)$$

The tapered least squares (TLS) inverse (eq. 2-50) is

$$[AA^T]^-_{ii'} = \sum_{k=1}^{D} U_{ik} \left(\frac{1}{\lambda_k^2 + \beta^2/\alpha^2}\right) U_{i'k} \quad (3\text{-}28a)$$

$$[AA^T]^- = \left[AA^T + \frac{\beta^2}{\alpha^2}\right]^{-1} \quad (3\text{-}28b)$$

The constraint $\int \hat{I} = 1$ cannot be applied to this problem because the lack of a zero degree term in m and A yields $\int \hat{I} = 0$. This produces an undesired effect that is better explained in section 4. To avoid it, an 'unbiased TLS inverse' will be defined by

$$[AA^T]^-_{ii'} = \left(1 + \frac{\beta^2/\alpha^2}{\lambda_1^2}\right) \sum_{k=1}^{D} U_{ik} \left(\frac{1}{\lambda_k^2 + \beta^2/\alpha^2}\right) U_{i'k} \quad (3\text{-}29)$$

As written, these inverses are optimum only when data noise is an uncorrelated process with constant variance. When the noise covariance E_D, the weights W and the filter F are made explicit, we write

$$B = [FAW]^T E_D^{-1/2} [E_D^{-1/2} FAWA^T F^T E_D^{-1/2}]^- E_D^{-1/2} \quad (3\text{-}31)$$

Notice that the elements of FAW and $FAWA^TF^T$ are computed from (14).

3.4 EXAMPLES

This section illustrates how the different parameters associated with the computational scheme and the data distribution affect the geoidal estimate.

3.4.1 SVD, TLS, AND UNBIASED TLS INVERSES

Assume a filtered data value is given every 1° both in latitude and longitude between 6°N to 6°S, 6°W to 6°E (169 data points). Each datum is a weighted average, the filter is that of equation (3-A15), with $\nu = 10{,}000$ ($\psi_o \simeq 0.81°$).

In this chapter (except in section 'Random noise') random errors in the filtered data values are always assumed to be uncorrelated with constant variance. Except where noted, no model weighting W is imposed. The only filter applied to the data is that of equation 3-A15, appendix 3-3. Most figures in this chapter present the square root of the ratio between the degree variances of the resolution functions (equation 3-A14) and (2n+1), the degree variances of a Dirac delta; they will be called 'degree responses.'

Figure 3-1 shows the degree responses for $(\text{lat, lon})_p$ = (0°,0°), at the center of the data region, computed by the SVD equation (3-26)/(3-27) for three different ranks: 169, 146, 97. Figure 3-2 shows the range of eigenvalues for this example. These figures show that the full rank solution retains as much high degree information as this filtering and distribution of filtered values allow to be defined. This choice of data spacing and filter width produce

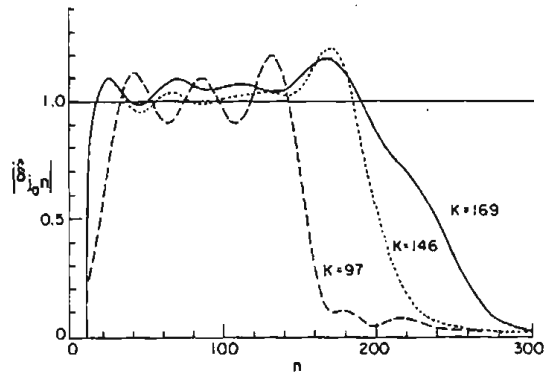

Fig. 1. Degree responses at $(\text{lat, lon})_{i_0} = (0°, 0°)$; data distribution: 6°W to 6°E, 6°S to 6°N, every 1°. Filter: $\nu = 10000$; SVD inverse, ranks 169, 146, and 97. Random noise sensitivities of 194, 46, and 18 cm/mgal, respectively. Data are assumed to lack degrees 0–9.

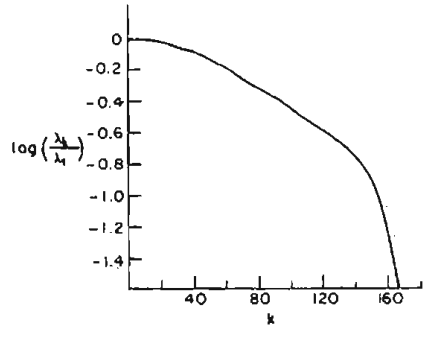

Fig. 2. Range of eigenvalues λ for the problem of Figure 1; $\lambda_1 =$ 13.1 mgal/cm.

redundant combinations of data; for K=169 very small eigenvalues are retained, and 1 mgal random noise in the filtered data produces 194 cm random noise in the geoidal estimate. Dropping smaller eigenvalues (Figure 3-1) eliminates high degree information, yields a sharper high degree cutoff, eliminates a small positive bias of the full rank solution, and brings noise down to 46 cm/mgal. Reduction of the high degree cutoff is the main effect of decreasing the rank up to K = 120. Beyond that, the low degree cutoff is affected too, as the curve for K = 97 shows; also, the ripples at the plateau become larger. Higher degrees are the least well defined and thus are associated with the smallest eigenvalues; the very low degrees are also associated with small eigenvalues because of the limited extent of the data. Notice that the plateau remains around 1 when the rank is decreased. (In Figures 3-1, 3-3, and 3-4 it is assumed that degrees 0-9 have been (exactly) removed from the data. This produces $\delta_n = 0$ for $n \leq 9$. However, these terms had not been removed from the functions that make up the inverse operator (equation 3-26).

The inverse operators for Figure 3-3 were computed by the TLS equation (26)/(28) and three different values of β/α: 0.7, 4, and 8 mgal/cm, respectively, equal to the eigenvalues λ_{169}, λ_{146}, λ_{111}. The $\beta=0$ solution is obviously the same as the full rank solution of Figure 1. Increasing β/α shifts the high and low degree cutoffs toward medium degrees, it decreases noise (63, 20, 11 cm/mgal,

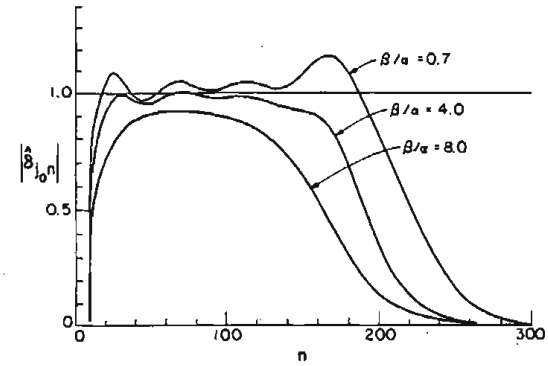

Fig. 3. Degree response. Same data as Figure 1. Degrees 0–9 removed from data and kernel. TLS inverse, β/α: 0.7 ($=\lambda_{169}$), 4 ($=\lambda_{146}$), 8 ($=\lambda_{111}$), all in mgal/cm. Random noise: 63, 20, 11 cm/mgal.

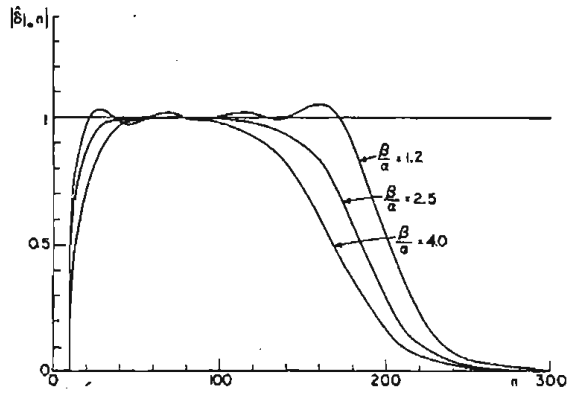

Fig. 4. Degree responses. Same data as Figure 3, but using the unbiased TLS inverse. β/α = 1.2, 2.5, 4.0 mgal/cm. Noise: 54, 35, 24, cm/mgal.

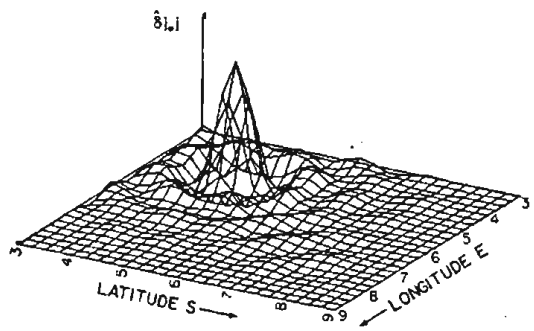

Fig. 5. Resolution function on the sphere. Same data as Figure 1. SVD inverse, K = 152. (lat, long)$_{j_0}$ = (4°, 4°).

respectively); but the most striking effect is the bias it introduces, shown by the decrease in the value at the plateau. The reason for this bias is the inability to force $\hat{I}$ to integrate to 1.

The unbiased TLS inverse, equation (3-26/28), eliminates this problem (Figure 3-4).

For illustration, Figure 3-5 shows the resolution function corresponding to one of these degree responses. Notice that it is difficult to infer anything from it, besides its center, which is at the correct location (showing that the phase condition is satisfied), its width, and the presence of sidelobes. Degree responses are easier to interpret for this problem.

3.4.2 QUALITY VARIATION ACCROSS THE DATA REGION

Quality of the geoidal estimate is degraded away from the center of the data region. To see this explicitly, consider the same filtering and data distribution of the previous example. (In addition, degrees 0-9 were removed from the kernel and its inner product to compute the inverse operator B. When this B is applied on data d', which completely lack the first 30 degrees, instead of acting on data which only lack degrees 0-9, $\sum_i B_{pi} A'_{ip'}$ has zero power at $n < 30$. This is best seen in equation (3-A14)).

Three zones can be distinguished: an inner zone, surrounding the center of the data region; an intermediate

zone, the outer rim of the data region; and, an outer zone, where no data are given.

Quality in the inner region is fairly uniform both in resolution and in sensitivity to noise; at data points, a band-passed version of the real geoid is reproduced. Figure 3-6 compares the degree responses in the inner region, using an unbiased TLS inverse, β/α = 3.4 mgal/cm. The response at (2°,2°) contains less information at low degrees than that at (0°,0°), the center of the data region.

A geoidal height at a point in this inner region, but where no gravity data is given, say at (0.5°,0.5°), is a weighted average of the closest neighbors that do have data, and its high degree content is diminished (Figure 3-7).

In the intermediate zone and toward the edge of the data region, resolution worsens noticeably, but we can still recover some useful information at intermediate degrees (Figure 3-8).

The data contribute no useful information about the outer zone (curve for (8°,8°) in Figure 3-8).

Degree responses do not tell the whole story because they lack phase information (both of which are expensive to compute). We then compute the rms value of the resolution error, equation (3-25). The degree response clearly show that $30 \leq n \leq 160$ is the range accurately resolved by the data. Assume that we know the component with $n<30$ from other

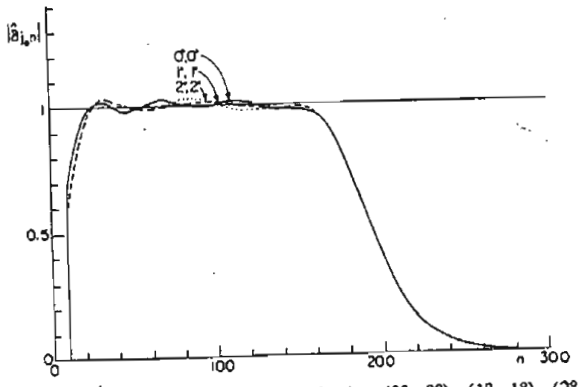

Fig. 6. Degree response. (lat, lon)$_{j_0}$: (0°, 0°), (1°, 1°), (2°, 2°). Unbiased TLS inverse, β/α = 1.7 mgal/cm. Data distribution: 6°N to 6°S, 6°W to 6°E, every 1°. Filter: ν = 10000, degrees 0–9 removed. Noise: 46 cm/mgal.

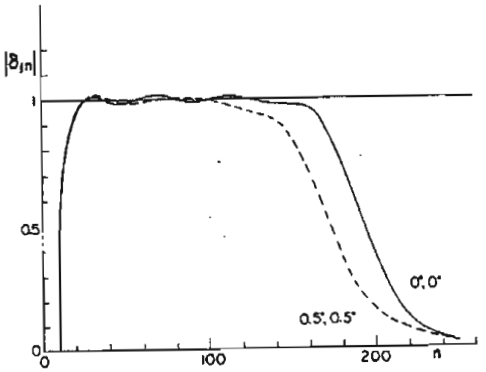

Fig. 7. Degree response. Same data and inverse as Figure 6. (lat, lon)$_{j_0}$: (0°, 0°), (0.5°, 0.5°). Noise: 46, 28 cm/mgal.

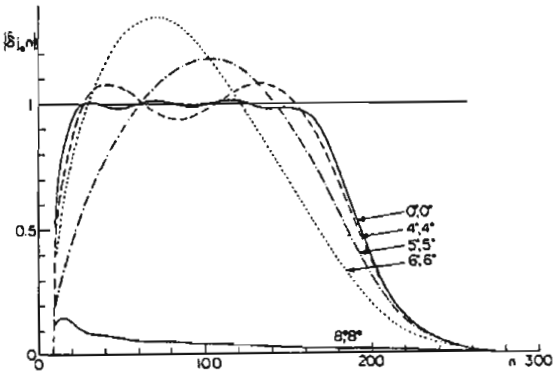

Fig. 8. Degree responses. Same data and inverse as Figure 6. (lat, lon)$_{j_0}$ = (0°, 0°), (4°, 4°), (5°, 5°), (6°, 6°). Noise: 46, 44, 40, 29 cm/mgal. Also shown, curve for (8°, 8°) outside the data region.

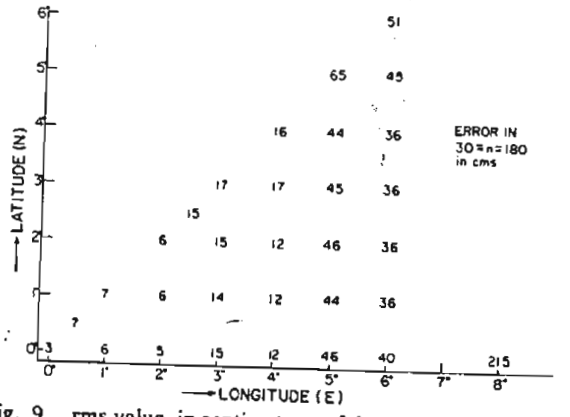

Fig. 9. rms value, in centimeters, of the systematic error in $30 \leq n \leq 150$ for the example of Figures 6–8. Total rms height by Kaula's rule: 213 cm. The figure can be completed by arguments of symmetry.

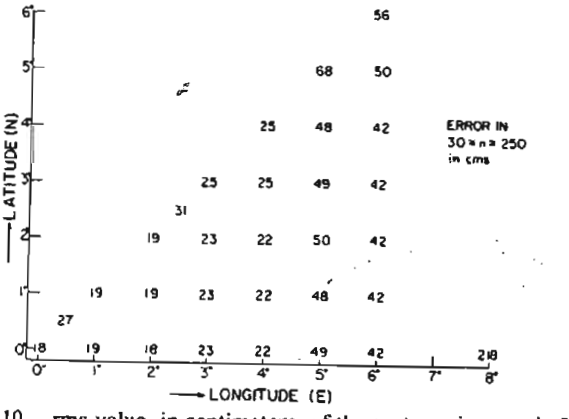

Fig. 10. rms value, in centimeters, of the systematic error in $30 \leq n \leq 250$ for the example of Figures 6–8. Total rms height by Kaula's rule: 216 cm. The figure can be completed by arguments of symmetry.

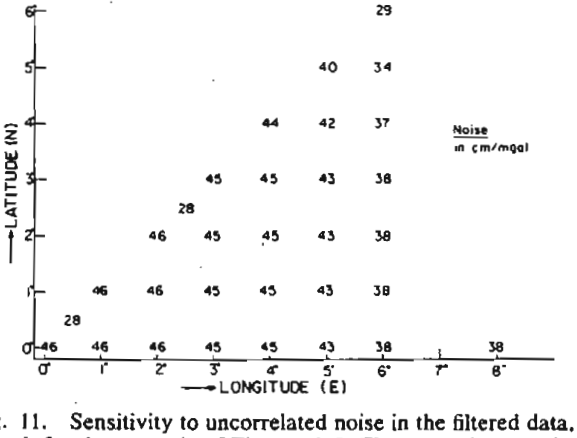

Fig. 11. Sensitivity to uncorrelated noise in the filtered data, in cm/mgal, for the example of Figures 6–8. Figure can be completed by symmetry.

measurements and that it can be (exactly) removed from the data. Figure 9 shows a map of the rms value of the resolution error in the range $30<n<160$, assuming Kaula's rule ($W_n=R^2\ 10^{-10}n^{-4}$; Kaula [1966]) approximates the spectrum of m. Noise is not included in figure 9. Because the high degree 'tail' cannot be removed the way the low degrees can, Figure 3-10 shows the same error in $30<n<260$ (the degree response is negligibly small for $n>260$). Finally, Figure 11 shows the effect of uncorrelated noise in the filtered values. A discussion of the effect of uncorrelated noise in the original point data is postponed until the last section of this chapter.

3.4.3 DATA COVERAGE

Now we shrink the data region to 5°S to 5°N, 5°W to 5°E, again sampled every 1° (121 points), applying the same filter as in the previous example. Figure 3-12 compares the solutions of maximum resolution for the central point with this reduced coverage and with the slightly larger coverage used before. The most significant difference is that the smaller data coverage distorts the low degree information in the geoidal estimate. The high degree cutoff has not changed, and the sensitivity to data noise has decreased only slightly. This same behavior is retained when dropping smaller eigenvalues, but the intermediate degrees are better defined when more data are given.

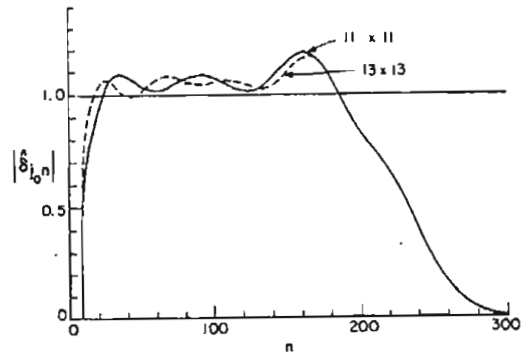

Fig. 12. Degree responses for (lat, long) = (0°, 0°). Data: sampled every 1°, filter $\nu = 10000$, degrees 0–9 removed from data and kernel. Full curve: data, 5°W to 5°E, 5°S to 5°N; SVD inverse, $K = 121$; noise 184 cm/mgal. Dashed curve: data, 6°W to 6°E, 6°S to 6°N; SVD inverse, $K = 169$; noise 194 cm/mgal. Labels refer to the number of data points.

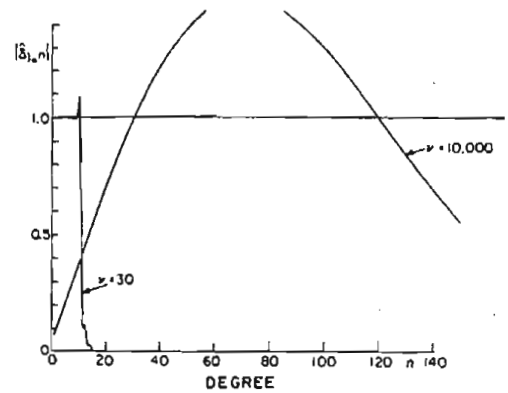

Fig. 13. Degree responses at (0°, 0°). Data: covering the globe, every 15° in latitude, and approximately every 15°/cos(lat) in longitude. Filters: $\nu = 10000$ and $\nu = 30$. SVD inverse. The fundamental differences between the curves is the filtering, but their respective ranks are $K = 191$ and $K = 121$.

For the corresponding problem on a plane, these conclusions are easy to infer from the analysis of time series; here they have been quantified for a spherical geometry.

Both Molodenskii's approach and these examples show that the information about the component of a geoidal height with $n>n_1$ is mostly given by the data around the point. Roughly, the minimum width of the data region must be about $2\pi R/n_1$ and is governed by the reference field that is available.

3.4.4 DATA SPACING: OVERSAMPLING

What is the effect of sampling the filtered data more closely than required by their filter? The sampling distance, $\Delta\psi$, should be smaller than one half the shortest wavelength passed by the filter (Nyquist frequency); however, the filter of equation 3-A15 does not have a sharp high degree cutoff, but the half-width ψ_o (equation 3-A17) can be used as a reference. For the following examples, we fix $\Delta\psi=1°$, and the extent of the data region as 5°N to 5°S, 5°W to 5°E.

The examples in the previous sections used a filter with $\psi_o=0.81°$ ($\nu=10,000$), with satisfactory results. The computations corresponding to $\psi_o=1.0°$ ($\nu=6500$) yield similar degree responses as with $\psi_o=0.81°$, but noise sensitivity is 5 times larger for equivalent degree responses. Furthermore,

if a filter with ψ_o=1.15° is used, not only the sensitivity to noise increases, but the computation itself may become unstable (in the full rank solution) because the ratio of the largest to the smallest eigenvalues of AA^T is 10^{-7}. Oversampling produces small eigenvalues, because it generates redundancies among the data.

Filtering the data with a very narrow filter (ψ_o=0.68°) yields the same full rank degree response as when ψ_o=0.81° is used, but lower noise sensitivity (35 cm/mgal). For smaller ψ_o however, the undesirable effects of undersampling begin to appear. They will be discussed in the next section.

In summary, the 'proper' sampling is in the range $1.2\psi_o<\Delta\psi<1.6\psi_o$ for the filter of equation 3-A15. Oversampling increases noise sensitivity at equivalent degree responses.

3.4.5 DATA SPACING: UNDERSAMPLING

A global data coverage can provide a good example. Filtered values, spaced by 15° of arc, are assumed to cover the whole earth, but the filter has a width ψ_o≈0.81°. Length scales between 15° and 1.1° (=$1.4\psi_o$) are then incorrectly sampled, because the width of the filter is much smaller than the distance between samples, a classical aliasing problem.

The resolution functions are of two kinds: those obtained right at a data position, and those far from one ('far' is more than 1° away). Far from a data point the resolution is zero, implying that there is no information about the geoid between data points (we did not use a model covariance function). At a data point the estimate is useless but interesting (Figure 3-13, curve for ν=10000). This curve is proportional to the degree response of the forward kernel ($A_n F_n/\sqrt{2n+1}$), and is precisely the resolution that would be obtained if the whole data set consisted of only one point (an answer that can be derived analytically because the problem has only one eigenvalue different from zero). These 191 data points act in isolation, each one providing information only in its immediate neighborhood.

The other curve in Figure 3-13 (ν = 30) is the degree response for the same distribution of data, but with a filter width $\psi_o \simeq 21°$. The result is the same everywhere on the sphere and has high noise sensitivity: 10 m/mgal. Because the resolution curves are similar whether the point is or not at a data position, the estimate of the field (not just one geoidal height) is a low-passed version of the real geoid. It is as if we knew the first 10 sets of coefficients for the spherical harmonic expansion, and from them had computed the geoid. (We note in passing that computation time can be saved for a uniform data distribution because AA^T acquires a Toeplitz structure [Colombo (1979)].

In summary, although undersampling decreases the sensitivity to data noise it increases the systematic error described by the degree response because of aliasing.

3.4.6 MODEL WEIGHTING FUNCTION

Undersampling failed because the formulation did not contain the information that gravity disturbances at different points are correlated; they were assumed to be independent. In spectral terms, the high degrees were assumed to be important (but they were not sampled properly). A model covariance function acts as an interpolator and decreases the importance of high degrees (see equations 2-44 and 2-46).

Consider point values of gravity disturbance (no filter is applied). Our first model covariance will be the same filtering function (3-A15). Figure 3-14, dashed line, gives an example. For $n < 180$ the response is the same as when the data are filtered. For $n > 180$ the response grows with n and goes to ∞. Growth occurs because the coefficients of A, in the absence of filtering, grow linearly with n; the inverse B manages to hold down the coefficients of BA only at degrees defined by the data distribution. The model estimate and its error remain bounded, a feature best seen if one considers again the definition of I. Disregarding data noise:

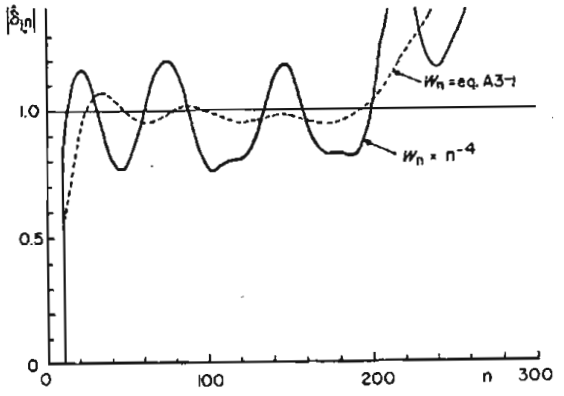

Fig. 14. Degree response for (lat, long) = (0°, 0°). Data are point values of gravity disturbance, minus a reference field up to degree 9. Distribution: 5°W to 5°E, 5°S to 5°N, every 1°. Model weighting was imposed, with Legendre coefficients equal to Kaula's rule, or equal to those of the filter (A15). Noise: 8 cm/mgal both.

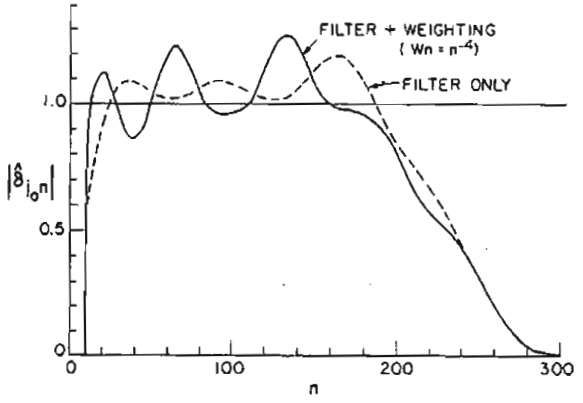

Fig. 15. Comparison between filtered data with and without model weighting by Kaula's rule. Dashed curve here is the same as the full curve of Figure 12. The full curve here has the same filtering, data distribution, and rank, but Kaula's rule was added as model weighting function. Noise: 184 and 170 cm/mgal.

$$\hat{m}_{p_o} = \int_p \hat{I}_{p_o p}\, m_p = \int_p \sum_i B_{p_o i}\, A_{ip}\, m_p$$

break up $\hat{I} = {}^1\hat{I} + {}^2\hat{I}$, where

$${}^2\hat{I}_{p_o p} = \sum_i B_{p_o i}\, {}^2A_{ip}$$

$${}^2A_{ip} = \sum_{n>n_2} A_n\, P_n(\cos \psi_{ip}) \,/\, \xi_n$$

i.e., ${}^2I$ and ${}^2A$ contain degrees higher than n_2 only. Therefore

$${}^2\varepsilon'' = \int_p {}^2\hat{I}_{p_o p} m_p = \sum_i B_{p_o i} \int {}^2A_{ip}\, m_p$$

$$= \sum_i B_{p_o i}\, {}^2d_i$$

Calling ${}^2d$ the rms value of ${}^2d_i$, which is the component of d_i with $n > n_2$, the rms value of ${}^2\varepsilon''_p$ is

$$({}^2d)^2 \Sigma_i B^2_{p_o i}$$

i.e., the component ${}^2d_i$ of the data, (which has $n>n_2$), acts as data noise in the inversion (equation (20)).

In summary, to use point values of gravity disturbance, a model weighting function must be used, and three sources of error can be recognized

$$\hat{m}_{p_o} = \int_p \hat{I}_{p_o p}\, m_p \pm \left[(({}^2d)^2 + \sigma_d{}^2) \sum_i B^2_{p_o i}\right]^{1/2}$$

The first term represents the band-pass filtering effect, and the second is the combined error due to degrees higher than n_2 in the data and to noise in the data

(When using filtered data, the error formula is similar, but ^2d is replaced by the error owing to lack of continuous data necessary to compute the filtered value.)

We emphasize that for any choice of weighting function, the resolution function describes its influence completely.

Figure 3-14 also shows the effect of changing the weighting function radically: Kaula's rule [Kaula, 1966] is used. Although both curves have broad similarities, this weighting strongly distorts the degree response. Intuitively, $W_n \sim n^{-4}$ tries very hard to assign energy to the lower degrees, whereas the data distribution cannot define them properly (but, if n^{-4} is indeed the spectrum of m, then the total rms error of m is minimum). The other weighting function (Figure 20) does not emphasize so much the low degrees. Weighting by n^{-4} produces a flat response when the size of the data region matches the lowest degrees included in the formulation. Sensitivity to noise is very small because high degrees in the data do not have to be amplified

When both an appropriate filter and weighting by $W_n = n^{-4}$ are used, the distortion of the degree response is also apparent (Figure 3-15).

In summary, any weighting function that makes (12) converge can be used, if its low and high degree 'cutoffs' are appropriate to the extent of the data and the distance between samples. In all cases, the error due to improperly sampled high degrees must be evaluated.

3.4.7 MORE DATA

Needless to say, the only real way to improve the quality of a geoidal estimate is to collect more and better data. For economic reasons, determining where to add data is important.

We have already seen that increasing the size of a data region improves the intermediate and low degree information. Obviously, increasing the number of data points inside the data region will improve the intermediate and high degrees. Figure 3-16 shows this effect explicitly. Notice that the influence of noise in the filtered values is much smaller when a narrower filter (relative to the sampling distance) is used.

These results provide the basis of a strategy for adding new information. If we wanted to use data with sufficiently close spacing to resolve all details of interest, and over a region large enough to match the low degree field provided by satellite orbit analysis, the associated matrices would be very large. If detail is needed only in a small region, it is convenient to break down the problem into two scales. Broad averages may be used to define the low degrees, whereas narrower averages, with closer spacing, can define the intermediate and high degrees to desired detail.

Consider the following example. The data region is

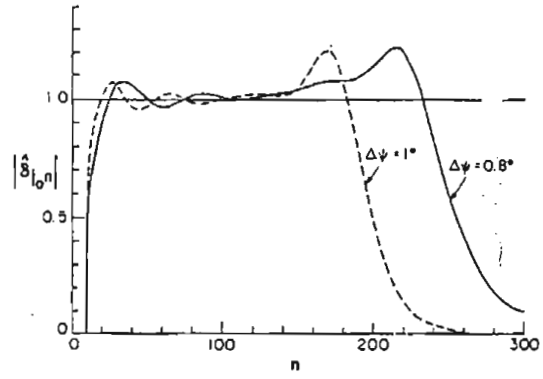

Fig. 16. Comparison between adding data in extent versus increasing data density. Dashed curve: data every 1° (ν = 10000), 6°W to 6°E, 6°S to 6°N. Full curve: data every 0.8° (ν = 20000), 4.8°W to 3.8°E. 4.8°S to 4.8°N. No model weighting. Both have 169 data points, and (lat, lon) = (0°, 0°). Ranks: 152, 154. Noise: 54, 24 cm/mgal.

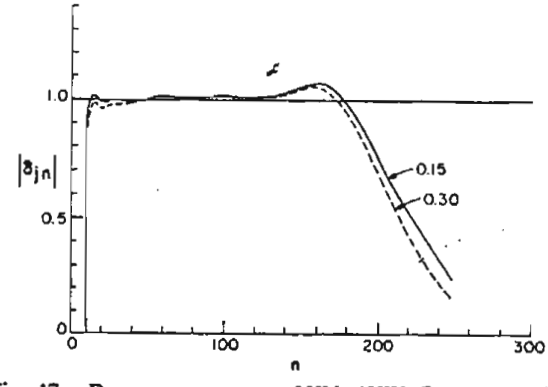

Fig. 17. Degree responses at 30°N, 40°W. Data: 147 values, in the North Atlantic, between 56°N and 0° every 4° in latitude, and every 4°/cos (lat) in longitude, filtered with ν = 800. To these, 121 values are added between 25°N to 35°N every 1°, 38°W to 49.5°W every 1.1°, filtered with ν = 12000, and have one-third the weight of the previous data. Both curves computed from unbiased TLS inverse, β/α = 0.15 mgal/cm (full curve) and 0.3 mgal/cm (dashed). Noise: 64 and 52 cm/mgal, respectively.

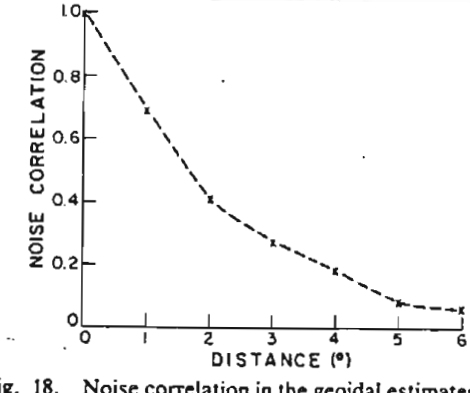

Fig. 18. Noise correlation in the geoidal estimates for the example of Figure 6. Point at the origin is (0°, 0°), and the distance is measured along the equator. Noise sensitivity at (0°, 0°) is 22 cm/mgal. See text for details.

the North Atlantic ocean, beween 0° and 56°N. Assume data are filtered with ψ_o=2.9° and sampled every 4° in latitude and every 4°/cos(lat) in longitude, yielding 147 values. We now add another 121 data values, each one obtained by filtering the original data with ψ_o=0.74°; these samples are spaced by 1° in latitude and 1°/cos(lat) in longitude, in the region 25°N to 35°N, 38°W to 49.5°W. Degrees 0-9 are removed from the formulation.

Each value filtered with ψ_o=0.74° was assigned a weight of one third (equivalent to the assumption that their standard error is 3 times larger than that of the broader averages). Different weights are needed because broader averages are always associated with small eigenvalues (fewer high degrees in the inner product function yield a smaller peak value, hence smaller elements in the corresponding matrix). In the region with detailed data (Fig. 3-17) degrees 13<n<140 are defined to within 2%. The high degree falloff occurs for n>180. Noise sensitivity is 64 cm/mgal. Increased damping in B (also in Figure 3-17) affects both the high and low degree ends of the response. If the broad averages are not at least 3 times more accurate than the narrow ones, damping affects only the low degrees.

It is not necessary for the region of interest to be in the center of the ocean. Detailed data in 30°N to 41°N every 1°, 56.8°W to 69°W every 1.2°, produce essentially similar degree responses for n>17. It is necessary,

however, that the width of the smaller region be at least twice the spacing betwen broader averages, so that their respective degree responses overlap.

In summary, simultaneous use of data with two different spatial scales gives adequate coverage for the definition of low degrees and adequate detail in small regions, while producing substantial computational savings. The only difference in the computation is that two partitions must be considered in the [A] matrix, and four in [AA^T]. Alternatively, the inversion can be done in two stages: the first uses only the broader averages over a large region, and the second stage uses the narrower averages over different small regions to improve the estimate [a technique discussed by Moritz, 1976].

3.4.8 DATA NOISE

All previous examples assumed that random errors in the filtered data values are uncorrelated. This yielded high sensitivity to noise in the geoidal estimates (around 50 cm/mgal in Figures 3-6 and 3-17). The main virtue of this assumption is its simplicity, but it is not very realistic. Because we start with point measurements of gravity, and take weighted averages of them (over small areas), any noise in the original data is correlated by the filter.

The noise in the original data is itself usually correlated, but this may be very difficult to describe. In a marine data set, for example, data collected during one

cruise, with the same instrument and sometimes without recalibration at the start and end of cruise, often show a constant offset relative to other cruises. In some ocean areas all the data may come from only one cruise, but in others the filtering process includes data from different cruises, each one with its own correlation. For lack of a quantitative description, we will now assume that noise in the original data is uncorrelated.

We return to the simple example of Figure 3-6, using the same inverse used for that figure. If noise in the original data were uncorrelated with constant variance, that in the filtered values would have a correlation $\simeq FF^T$ (a matrix whose elements are values of the inner product function of the filter). We recomputed the sensitivity to noise for the resolution curve of Figure 3-6 under these conditions and obtained 22 cm/mgal, less than half the 46 cm/mgal under the previous assumption. Equally important, model noise is itself strongly correlated (Figure 3-18), indicating that most of the error is at wavelengths large relative to the sampling distance.

The value 22 cm/mgal refers to 1 mgal correlated noise in the filtered values. To compute the sensitivity to noise in the original data, we need to know the average spacing $\Delta\psi$ between data points (if the filter is a degree average, for example, noise variance in the original data is reduced by a factor $(1 + 1°/\Delta\psi°)^{-2}$ in the average). Assuming the point

data are spaced, on average, $\Delta\psi$=0.25° (25 in a degree square), noise variance in the original data is attenuated by ~0.04 with the filter (A15) (ψ_0=0.81°). Each mgal of uncorrelated error in the original data would then produce only 4.4 cm error in the geoidal estimate, with a correlation approximated by that on Figure 19 (which is strictly correct for continuous data). The noise in marine gravity measurements is between 10 and 20 mgal (see chapter 4), hence they would produce errors between 44 and 88 cm in geoidal estimates.

Lack of continuous data is a source of large error in the filtered data values, an aliasing problem similar to that described under 'Model Weighting Function'. For practical purposes, we may consider all the energy at wavelengths shorter than twice the sampling distance as noise in the data. For the following discussion, the recent spectrum of Brammer and Sailor (1983) will be used at degrees>50. We assume a sampling every $\Delta\psi$ can accurately define wavelengths longer than $2\Delta\psi$. Then, the rms value of the undefined length scales in the gravity anomalies, which will act as noise in the point data, is

sampling $\Delta\psi$°	0.25	0.5	1.0	5.0
error (mgal)	21	30	35	38

It follows that any sampling coarser than once every 0.25° would produce larger geoid error than measurement noise would. Of course, in very 'rough' areas the omission

error will exceed these values. These errors are correlated because gravity anomalies are, but we may assume they are uncorrelated and add their variance to that produced by measurement noise. It is in this sense that undersampling produces similar effects to data noise.

3-5. SUMMARY

Because the problem is underdetermined, many 'reasonable' geoidal estimates can satisfy the same incomplete set of gravity data within its noise level. Different criteria of optimality lead to different inverse operators and different model estimates, no single one being the best. The SVD, TLS, and collocation inverses described in section 3 provide an efficient, data adaptive procedure for computing such estimates. The resolution function and its degree response, described in section 3 and used extensively in section 4, provide a clear description of the systematic error due to the incompleteness of any data set, and to the peculiarities of any inverse operator chosen. By assuming a spectrum for the unknown model, an rms value of this error can also be estimated.

The optimum inverse operators used here require inverting a matrix whose size is the number of data, an operation whose cost grows like the cube of this number (except for special cases). We have discussed in 3.4 (under 'More Data') a strategy for decreasing the size of these matrices

without losing significant information. Only data in a limited region are needed when low degree coefficients are available from satellite orbit analysis. This component must be removed from the data. The original data are filtered and the filtered field is sampled according to the filter used. Broader averages are sufficient to define the long and intermediate wavelengths (in such a way that the lowest degree defined by their distribution coincides with the highest degree available from satellite orbit analysis). Narrower averages, sampled more frequently, then provide the needed detail in a small region, whose size is at least twice the spacing between broader averages.

There are other valid but suboptimum inverses for computing geoidal estimates. The variants on Stokes integral discussed by Jekeli [1981] are excellent for large, uniform sets of point data, even if they are not optimum for discrete, usually filtered, and noisy data. These modified, discretized kernels are approximate linear inverse operators of the same forward problem posed here (independently of how they were derived), hence the error equations 3-18 and 3-25 apply also to them. The geoid presented in Chapter 4 was computed using one such suboptimal inverse, and its expected errors were computed using the error equations discussed in this chapter.

Emphasis was placed throughout this chapter on the limitations of incomplete data. The effect of data noise

was discussed briefly in section 3.4.8. It should be clear, however, that for any real data set both problems are equally important (section 3.4.8). High data noise poses additional questions, the most important of which is whether a range of length scales exists that is both well defined by the data distribution and sufficiently free fron noise.

Each point measurement of gravity can be thought of as having two components of error: one is due to measurement noise, the other one to unsampled short wavelengths in the neighbourhood of the point (aliasing error). This second component only appears when one needs to use data in a neighbourhood, e.g., when filtering gravity or when applying an integral transform that converts gravity into geoidal heights. Measurement noise is independent of sampling distance, but the aliasing error increases with distance. Assuming measurement noise to be about 20 mgals, the aliasing error will exceed measurement noise for sampling distances larger than 0.25°.

APPENDIX 3-1

INNER PRODUCTS AND CONVOLUTION ON A SPHERE

Various properties of integrals over the sphere used in previous sections are listed in this appendix. 'Fully normalized' Legendre functions and polynomials (H&M) are used throughout ; θ_p, λ_p are colatitude and longitude at point p. Integrals recover the mean value of the integrand over the sphere. Let

$$\Phi_{nmp} = P_{nm}(\cos\theta_p)\exp(im\lambda_p) / \sqrt{2} \qquad (3\text{-A1})$$

$$g_p = \sum_{n=o}^{\infty} \sum_{n=-n}^{n} g_{nm}\, \Phi_{nmp} \qquad (3\text{-A2})$$

$$F_{ip} = \sum_{n=o}^{\infty} \sum_{m=-n}^{n} F_{nmi}\, \Phi_{nmp} \qquad (3\text{-A3})$$

$$\int_p = \iint (\;)\, d\sigma_p$$

$$\xi_n = 1/\sqrt{2n+1}\;; \qquad \sum_n = \sum_{n=0}^{\infty} \qquad \sum_m = \sum_{m=-n}^{n}$$

The normalized Legendre polynomial $P_n = P_{no}$ satisfies

$$P_n(\cos\psi_{ip}) = \xi_n \sum_{m=-n}^{n} \Phi^*_{nmi}\, \Phi_{nmp}$$

and the inner product between F and g (complex for generality) becomes

$$\int_p F^*_{ip}\, g_p = \sum_n \sum_m F^*_{nmi}\, g_{nm} \qquad (3\text{-A4})$$

When F is a real-valued function of distance between i and p, like Stokes function, or the kernel (12), or the filter (3-A15) or a degree-averaging kernel, F_{nmi} becomes

$$F_{nmi} = F_n\, \Phi^*_{nmi} \qquad (3\text{-A5})$$

with F_n real. Then, (3-A3) becomes

$$F_{ip} = \sum_n F_n \frac{1}{\xi_n} P_n(\cos \psi_{ip}) \qquad (3\text{-A6})$$

When this filter is applied on g it yields

$$\int_p F_{ip}\, g_p = (Fg)_i = \sum_n F_n \sum_m g_{nm}\, \Phi_{nmp} \qquad (3\text{-A7})$$

An inner product of F with itself yields

$$\int_p F_{ip}\, F_{i'p} = (FF^T)_{ii'} = \sum_n F_n{}^2 \frac{1}{\xi_n} P_n(\cos \psi_{ii'}) \qquad (3\text{-A8})$$

The identity operator $I_{p_o p}$ on the sphere can be defined by its essential properties, sifting and unimodularity;

$$\int_p I_{p_o p}\; g_p = g_{p_o} \qquad (3\text{-A9})$$

$$\int_p I_{p_o p} = 1 \qquad (3\text{-A10})$$

It is easy to see that

$$I_{p_o p} = \sum_n \sum_m \Phi^*_{nmp_o}\, \Phi_{nmp} \qquad (3\text{-A11})$$

formally satisfies the requirements (because this series does not converge for $p=p_o$, a proof requires use of distributions or generalized functions).

Furthermore, I is only a function of distance, hence it can be written (formally) as

$$I_{p_o p} = \sum_n \frac{1}{\xi_n} P_n(\cos \psi_{p_o p}) \qquad (3\text{-A12})$$

APPENDIX 3-2

RESOLUTION ON A SPHERE: SPHERICAL HARMONIC EXPANSION

The resolution function at point $_{po}$ is

$$I_{p_o p} = \sum_i B_{p_o i} \; A_{ip}$$

where the summation is performed over the data positions i.

The forward kernel can always be expanded as

$$A_{ip} = \sum_n \sum_m A_{nmi} \; \Phi_{nmp}$$

therefore, $\hat{I}$ can be expanded as

$$\hat{I}_{p_o p} = \sum_n \sum_m \hat{I}_{p_o nm} \; \Phi_{nmp} \qquad (3\text{-A13})$$

When A_{ip} is only a function of distance betwen i and $_p$

$$A_{nmi} = A_n \; \Phi_{nmi}$$

In this case, the degree variances of δ are

$$|\hat{\delta}_{p_o n}|^2 = \sum_{m=-n}^{n} |\hat{I}_{p_o nm}|^2 \qquad (3\text{-A14})$$

$$= \xi_n^2 A_n^2 \sum_i \sum_{i'} B_{p_o i} \; B_{p_o i'} \; P_n(\cos \psi_{ii'})$$

APPENDIX 3-3

AN APPROXIMATELY GAUSSIAN SPHERICAL FILTER

We wish to find a filter F with these properties: (1) $F_n << n^{-2}$ as $n \to \infty$, so that (12) would converge rapidly even if $W_n = 1$, or $n \to \infty$; (2) $F(\psi) \simeq 0$ for $\psi > \overline{\psi}$, so that the integral over the sphere can be replaced by an integral over a small spherical cap; (3) a low-pass filter, which integrates to 1. Some searching with the above criteria in mind led us to

$$F(\psi) = (\nu+1)\ (\cos\ \psi/2)^{2\nu} = \sum_{n=0}^{\nu} F_n\ \xi_n^{-1}\ P_n(\cos\ \psi) \qquad (3\text{-A15})$$

$$F_n = \frac{(\nu+1)\ (\nu!)^2}{(\nu+n+1)!\ (\nu-n)!} \qquad (3\text{-A16})$$

(Gradsztheyn and Ryzhik [1965], equation (7-127), with $t = 2\ (\cos^2(\psi/2) - 1)$). Figures 19 and 20 show examples of this filter. A quick characterization of the filter can be given by a half-width ψ_o, defined as the distance to the inflexion point

$$d^2F/d\psi^2\Big|_{\psi=\psi_o} = 0$$

that yields

$$\tan^2(\psi_o/2) = 1/(2\nu-1) \qquad (3\text{-A17})$$

Another interesting property of this filter, for large ν and small ψ is its similarity to a gaussian

$$F(\psi) \simeq (\nu+1)\ \exp(-\psi^2/2\psi_1^{\ 2}) \qquad (3\text{-A18})$$

$$\sin^2(\psi_1/2) = 1/2\nu \simeq \psi_1^{\ 2}/4 \simeq \psi_o^{\ 2}/4 \qquad (3\text{-A19})$$

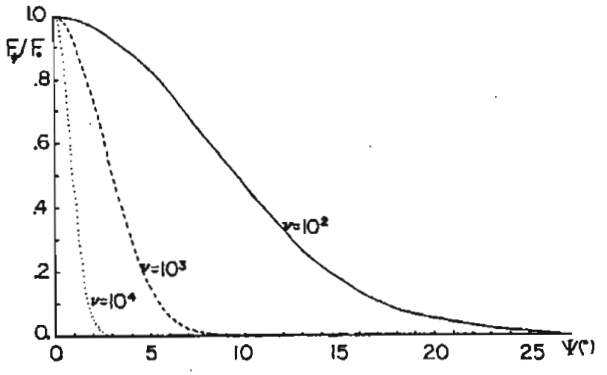

Fig. 19. Examples of the spherical filter given by equation (A15).

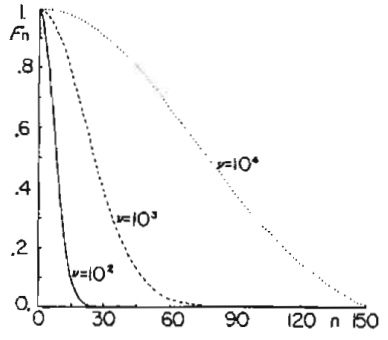

Fig. 20. Legendre coefficients of the spherical filter given by equation (A16).

This approximation if useful because a Gaussian is a commonly used filter for plane geometries, precisely the range of ν and ψ for which the approximation is valid.

CHAPTER 4

THE ACCURACY AND COVERAGE OF MARINE GRAVITY DATA IN THE NORTH ATLANTIC: CONSEQUENCES FOR GEOID ESTIMATION.

4.1 INTRODUCTION

The construction of geoids from gravity was analyzed in the previous chapter. We now discuss the quality of gravity acceleration data, estimate a geoid, and compare it to an altimeter-derived surface. In section 2 the quality of publicly available marine gravity data for the North Atlantic is assesed by looking at the discrepancies between measurements at the points where cruise tracks cross; summary statistics per cruise and geographical distribution are given. The method by which geoidal heights are computed is described in section 3, together with the relevant equations for error computation (the optimum inverses of the previous chapter were not used, for reasons discussed in section 3). Comparison between geoidal estimates and a SEASAT altimeter surface, together with the error estimates of both, are shown in section 4. It is shown there that, although the geoidal estimate accounts for much of the variance in the SEASAT surface, the remaining variance is much larger than the amplitude of the oceanographic signals we seek, except in a small area, off the U.S. Coast, and is usually consistent with the estimated errors of the geoid.

4.2 DATA SET. CROSSOVER ANALYSIS.

The marine gravity data set consisted of cruises from various institutions, obtained through the National Geophysical and Solar Terrestrial Data Center, supplemented by cruises provided directly by the Lamont-Doherty Geological Observatory and the Woods Hole Oceanographic Institution. All gravity anomalies were converted to the GRS-67 reference field (see, for example, Bomford 1980). These cruises are listed in Appendix 4-2.

The cruises were filtered alongtrack, with a Gaussian filter, halfwidth 50 km, after deleting abnormally large values. The purpose of the filtering was to decrease the number of data values to be handled. The places where two cruises met were found by a search routine, and the values at the crossover point linearly interpolated from the two neighboring filtered values. A crossover discrepancy was computed as the difference between the interpolated values. Computing the discrepancies from filtered data, rather than using the original "point" values (themselves averages over 1 minute of time alongtrack), introduces an aliasing error into the crossover discrepancies, schematically explained in Figure 4-1. The difference between the point and filtered values places an upper bound on the aliasing error. This difference, for cruise SS009, has an RMS value of 15 mgal, and reaches peak values of 80 mgals; these two figures will be considered

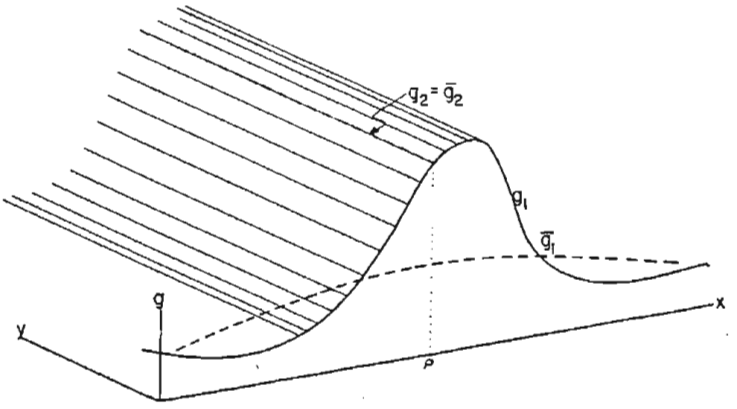

Figure 4-1: Sketch of aliasing error in crossover discrepancies computed from filtered cruises. Suppose the field is locally two-dimensional —lines of constant g run parallel to the y axis. A cruise "2" moving parallel to y measures g_2; after alongtrack filtering, the filtered values are g_2, but here $g_2 = g_2$. A cruise "1" parallel to the x axis measures g_1; after filtering the result is g_1. $g_1(P) \neq g_2(P)$ because of the filtering, not because of measurement errors.

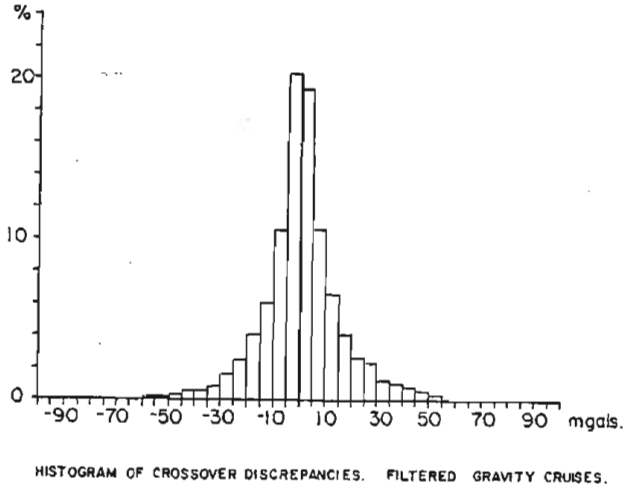

Figure 4-2. Histogram of crossover discrepancies. This histogram includes all 13518 crossovers, including those from cruises rejected for the geoid computation. 95% of the values are smaller than 45 mgals. About 0.3% (42 crossovers) exceed 200 mgals.

representative upper bounds on the RMS and peak values of the aliasing error, because cruise SS009 contains a mix of smooth midocean gravity anomalies and the rougher and much larger anomalies associated with the edge of the US shelf, Canary islands (300 mgal amplitude), and other features rich in short wavelengths. Of course, filtering alongtrack also attenuates uncorrelated errors, which has the effect of decreasing the total error variance.

The mean and standard crossover discrepancies are listed in Appendix 4-2 for each cruise. Only those cruises labelled 'T' were used in the final geoid computation. A histogram of all the errors is presented in Figure 4-2; 60% of the errors have magnitude smaller than 10 mgals and 95% are under 45 mgals. A student 't' test, at 95% confidence, showed that the mean crossover for many cruises had a systematic origin (gravimeter drift and lack of calibration at the beginning and end of a cruise can produce such an error), therefore least squares estimates of the mean crossovers were computed for each cruise, and removed before the geoid computation†. Figure 4-3 shows the distribution of discrepancies over the North Atlantic. They are larger in areas of high gravity

† In retrospect, only crossovers in areas with smooth anomalies should have been used. This, however, would only affect cruises with a majority of crossovers in rough areas.

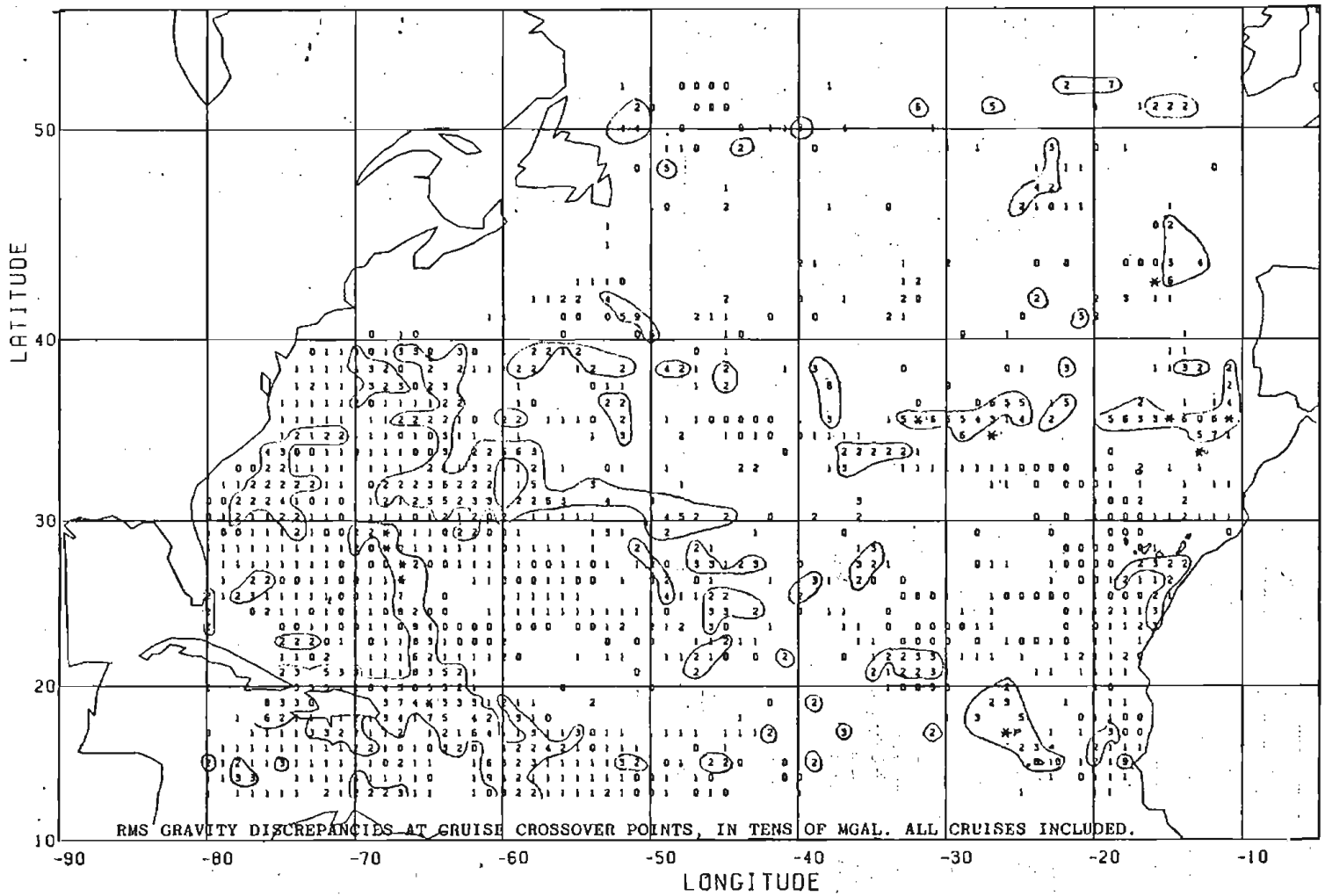

FIGURE 4-3: RMS gravity discrepancies at cruise crossover points. Averaged over 1° bins. Annotations in tens of mgals. Empty bins indicate that no crossovers occurred inside the bin. '*' indicates discrepancies exceeding 100 mgal.

gradients, such as the Caribbean, Puerto Rico Trench and continental edges, both because of errors in positioning at sea, and because of the previously discussed aliasing error.

To compute geoidal estimates we needed gravity data up to 10° away from the point of interest. Land gravity data for the continental U.S. was provided by the National Geophysical and Solar Terrestrial Data Center, updated through 1980. Land gravity over Canada was provided by the Gravity and Geodynamics Division of the Canadian Department of Energy, Mines and Resources, updated through November 1980. These data sets include accuracy estimates, and they were accepted at face value. Gravity data in the Caribbean were digitized from Figure 3 of Bowin (1976), and assigned errors of 25 mgal. Data on the Bahamas and Bermuda islands were also digitized from Bowin et al. (1982), and assigned the same error. The 'point' gravity data, both on land and at sea, were gridded every 0.5° of latitude and every 0.5°/cos (lat) in longitude, starting at 70°W. A gridded value at point p was computed as $\sum_i (F_{ip}\, \tilde{g}_i)/\sum_i (F_{ip})$, where $\sum_i$ indicates a sum over all gravity values $\tilde{g}_i$ within a 2° radius from point p. The filter F_{ip} is the one described in Appendix 3, chapter 3, with ν = 50,000, which is equivalent to a Gaussian with half width ≅ 40 km. Grid nodes with scarce or no data were declared empty. Figure 4-4 shows the distribution of occupied grid nodes belonging to cruises with no more than 20 mgal rms crossover discrepancies; this data set will be called DATA-1. The data in Figure 4-5 (DATA-2) have less than 30 mgal errors.

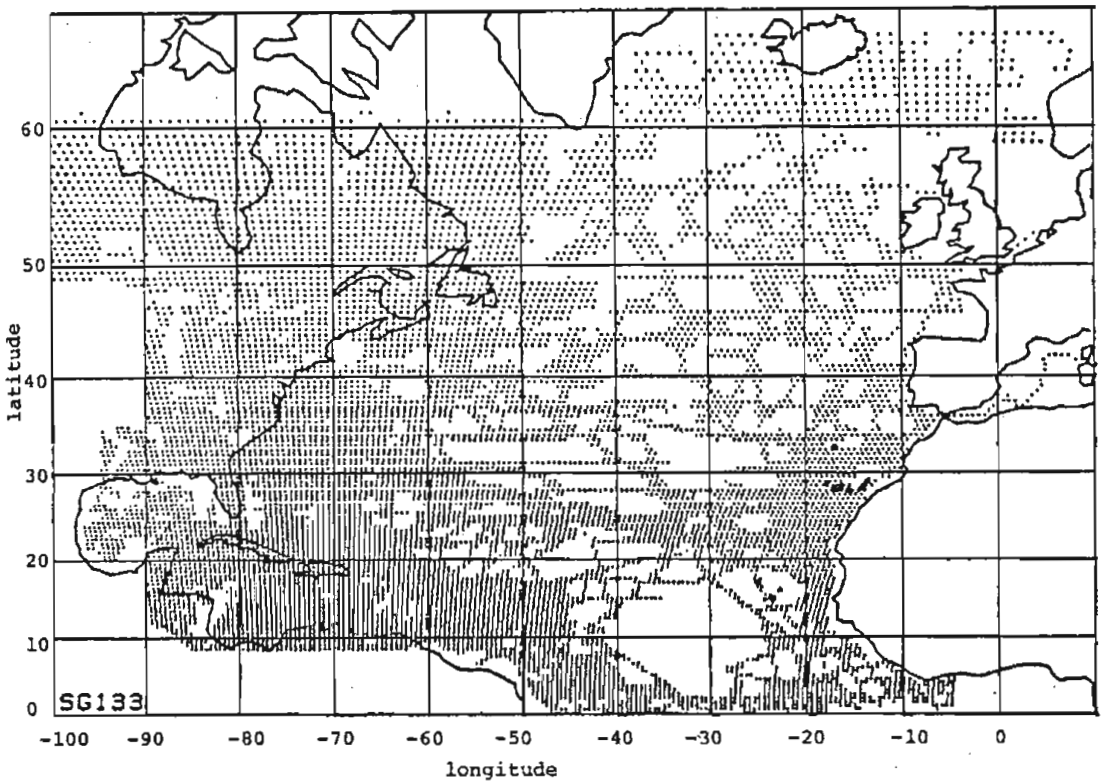

FIGURE 4-4. Distribution of gridded gravity data from cruises whose total error does not exceed 20 mgals.

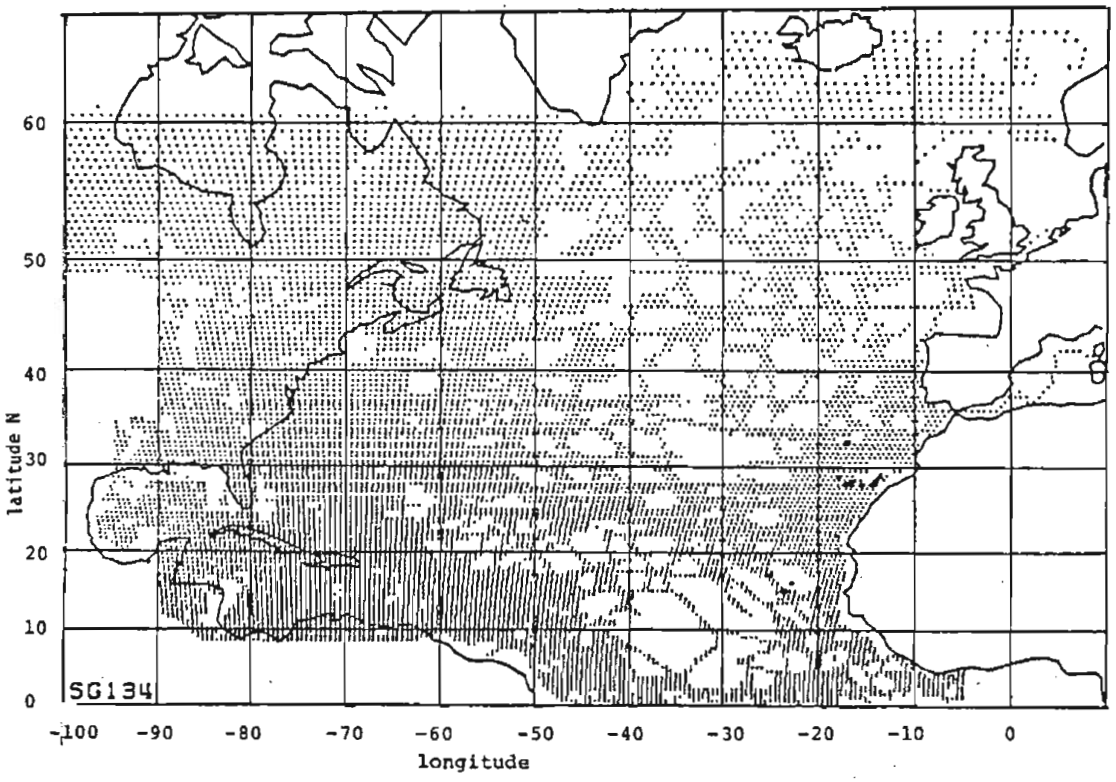

FIGURE 4-5. Distribution of gridded gravity data from cruises whose total error does not exceed 30 mgals.

4.3 METHOD OF COMPUTATION

Three geoidal estimates were computed: $\gamma[1]$ and $\gamma[2]$ differ in the amount of basic data used; $\gamma[2]$ and $\gamma[3]$ differ only in one computational step - an optimum interpolation of empty grid nodes for geoid 3. The optimum inverses described in chapter 3 were not used for this geoid estimation because the author considered the amount of computer time and storage that they would have required on the available computer (an IBM 370) to be excessive, even after partitioning the problem suitably. The method used is as follows: 1) - only for $\gamma[3]$- all empty nodes of the grid depicted in figure 4-5, over the ocean, were filled with optimally interpolated gravity estimates; 2) the product of gridded gravity values (minus their GEM-9 component) times a modified Stokes function were numerically integrated over all grid nodes within 10° from the point where the geoid was desired. The type of numerical integration, and the required equations for the modified Stokes function are detailed in Appendix 4-1. Only an overview of the method and a brief justification are given in this section.

Optimum estimation is needed so that both the absence of necessary data and errors in the available data can be taken into account. A geoidal estimate at a point P is sensitive to gravity data far away from P (see 'Data Coverage" in chapter 3), but an estimate of gravity itself at P is only sensitive

to gravity in its immediate neighbourhood - hence it can be computed from far fewer data values. By filling the empty grid nodes with optimally interpolated values one greatly reduces the problem of missing data. If gravity is known at all grid nodes within a certain radius ρ from P, and a long wavelength reference field is removed, then one can compute accurate geoidal heights with a numerical integration, so long as the kernel compensates for the absence of data both outside the cap of radius ρ around P, and in-between grid nodes. Molodenskii's modification of Stokes kernel is optimum for continuous data inside a cap of radius ρ, and its Legendre series was truncated at degree 360 to account for the discrete sampling (other options could have been used).

The numerical integration is simply a linear combination of gravity data. It follows that the error equation 2-17 (ch. 2) can be applied to the resulting geoid, and a realistic error estimate can be computed. Such a computation requires that we know the error structure of the gridded data and the power spectrum (or covariance function) of the desired geoid. The power spectrum computed by Wagner and Colombo (1978) - based on GEOS-3 altimetric data, satellite orbit perturbations, and 1° averaged surface gravity - was used to describe the average behaviour of the geoid.

The computational method also decreased the influence of error in the GEM-9 coefficients. Jekeli (1981) analyzed modified Stokes integrals assuming continuous gravity data;

his error curves show that the influence of GEM-9 noise in equation (4-1-2) is some 30 cm RMS, whereas the accumulated error in the GEM9 coefficients is 190 cm rms. The difference is information provided by the gravity data, hence this $(30\ cm)^2$ variance must be added to the error variance computed from equation (4-1-13) .

4.4 GEOIDAL ESTIMATES: PART 1

In this section two geoidal estimates (γ) are compared to a filtered version of the Seasat altimeter data, adjusted by Rapp (1982). All 3 months of Seasat altimetry were used. The Seasat heights were gridded simply by averaging all data inside a box defined by two parallels separated by 1°, and two meridians separated by 1° (hence the areal extent of the 1° average changes with latitude). On the other hand, the geoids γ are approximately gaussian averages - those used to grid the gravity data. The expected discrepancy between a 1° averaged geoid and the gaussian-averaged geoid (Appendix 3-3, ν = 50,000) is about 10 cm, again assuming the Wagner and Colombo (1978) spectrum describes the geoid. These 1° averaged altimeter heights will be referred to as 's'. The resolution error estimates shown later, are also relative to an errorless geoid filtered in this manner.

For the examples shown in this section, no optimum interpolation of either DATA-1 or DATA-2 was performed prior to numerical integration.

Figure 4-6 shows profiles of the difference between the geoidal estimates and s. Clearly, both geoidal estimates contain more information about s than GEM 9 does: the standard deviation of the differences between GEM 9 and s is 172 cm whereas s-γ[1] and s-γ[2] have 100 and 82 cm s.d., respectively. The second, somewhat surprising, observation is that in this case the additional data, although much noisier than the existing ones, provide substantial information. Possible reasons: 1) the global rms value of gravity anomalies is about 42 mgal, and about 39 mgal if the first 20 degrees are removed (in locally 'rough' regions, e.g., across a trench, it can be much higher). This is about the error of implicitly interpolated values where data are very scarce, whereas the noisy data being added have smaller errors. 2) Data errors have most of their energy at wavelengths shorter than the 100 km cutoff of these maps (P. Malanotte-Rizzoli, 1983, personal communication); 3) the crossover analysis based on filtered cruises overestimates errors.

In Figure 4-6, the mean difference s - GEM 9 is -110 cm, whereas the means of s-γ[1] and s-γ[2] are +107 and +52 cm respectively. Rapp (1981) has argued that the systematic difference between his adjusted heights (referred to the GRS 1980 ellipsoid) and GEM 9 (which implicitly refers to the best fitting ellipsoid, of unknown parameters) is due to a difference between the radii of the two ellipsoids. The bias in geoids 1 and 2 is most likely error: the mean value of the

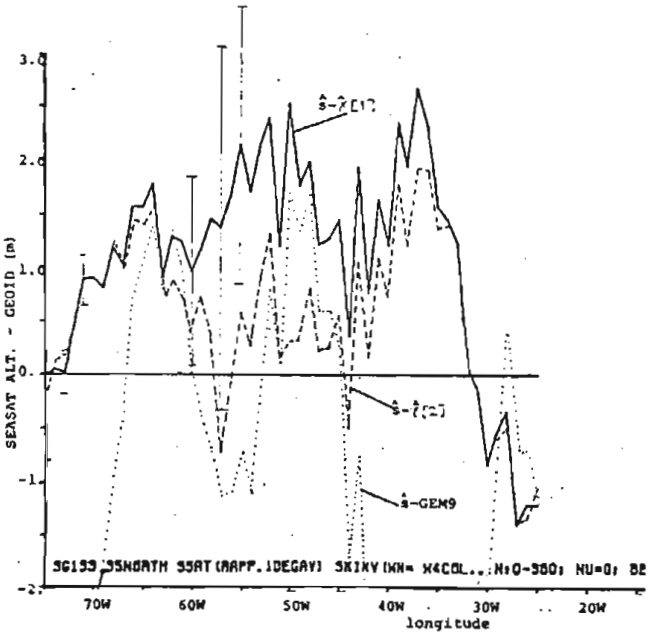

FIGURE 4-6: Profiles along 35°N. Differences s-GEM9, s-γ[1] and s-γ[2]. The mean and standard deviations are: (-110,172), (107,100) and (52,82) cm respectively. The bias is most likely error (see text discussion). γ[2] accounts for much of the variance in s unaccounted for by GEM9, but the difference is still larger than expected ζ. Even the difference between γ[1] and γ[2] -east of 60°W-exceeds expected ζ. Expected noise in γ[2] varies from 15 cm -west of 70°W-to 62 cm at 54°W. Resolution errors vary between 15 and 35 cm.

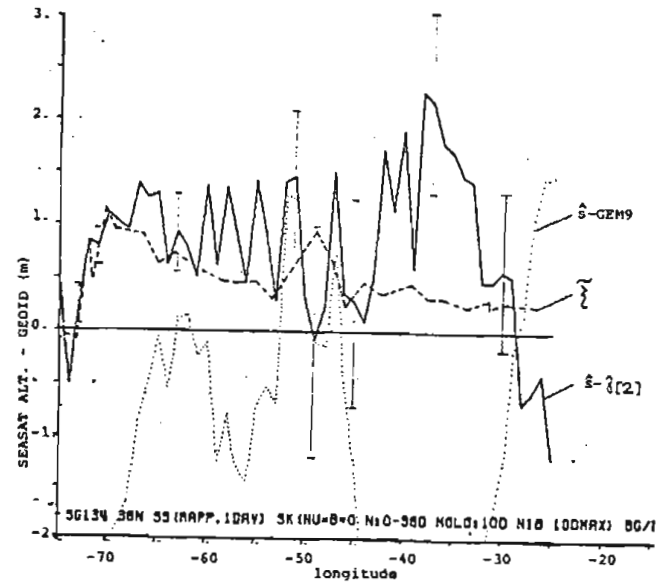

FIGURE 4-7: Profiles at 36°N. The mean and s.d. of s-GEM9 are (-142,176)cm; those of s-γ[2] are (80,74)cm. ζ, as inferred by Wunsch(1981) from hydrographic data only, accounts for most of the difference s-γ[2] west of 60°W. Expected geoid noise varies between 11 cm -at 69°W-and 46 cm. Expected resolution error varies between 10 cm -at 73°W- and 105 cm, assuming the Wagner and Colombo spectrum is an accurate description of the average geoid spectrum.

modified Stokes function over a 10° cap is about 52 cm/mgal, therefore a constant residual error in the crossover adjustment of only 2 mgal would produce a bias of more than 100 cm.

Figure 4-7 compares s-γ[2], at 36°N with the component of sea-surface topography associated with the general circulation at that latitude, ζ, as estimated by Wunsch (1981) from hydrographic data only. The mean of ζ over the whole ocean, 70 cm, was removed because the geostrophic equation does not define it. Clearly, ζ accounts for much of the variance in the profile, but it is difficult to infer ζ from s-γ. The reason is simply that the errors in the geoid are not of the same magnitude along the whole profile (this point is expanded in chapter 5). The error bars at points of Figure 7 indicate total RMS error estimate. The influence of noise in the gravity data is less than 15 cm west of 67°W, and varies between 30 and 40 cm elsewhere (highest: 63 cm at 30°W). The resolution error, however, is about 15 cm (RMS) west of 65°W, but climbs to 120 cm at 43°W (the resolution error was computed at a few selected points only due to its high computational expense). The relative sizes of ζ and the total geoid error (the error in s is negligible) imply that only features significantly larger than 20-30 cm can be recovered between 75°W and 65°W. Only the Gulf Stream itself is assured from Figure 4-7. East of 65°W at this latitude, ζ cannot be recovered because it is not significantly larger than the 120 cm rms error of the geoidal estimate. (Furthermore, the

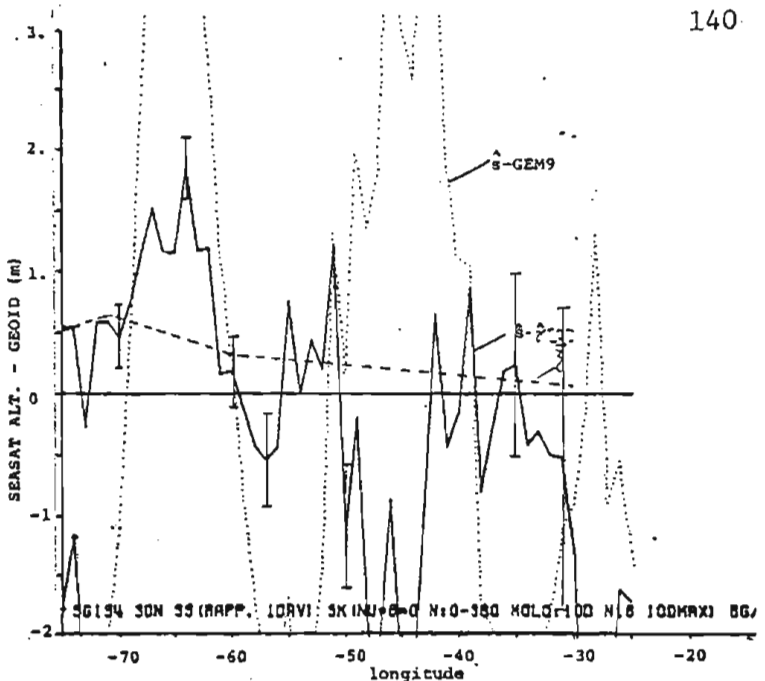

FIGURE 4-8: Profiles at 30°N. The mean and s.d. of s-GEM9 and s-γ[2] are (22,240) and (-33,144)cm respectively. The areas around 45°W and 27°W are covered only by high noise cruises. West of 55°W expected geoid noise alone is between 40 and 55 cm.

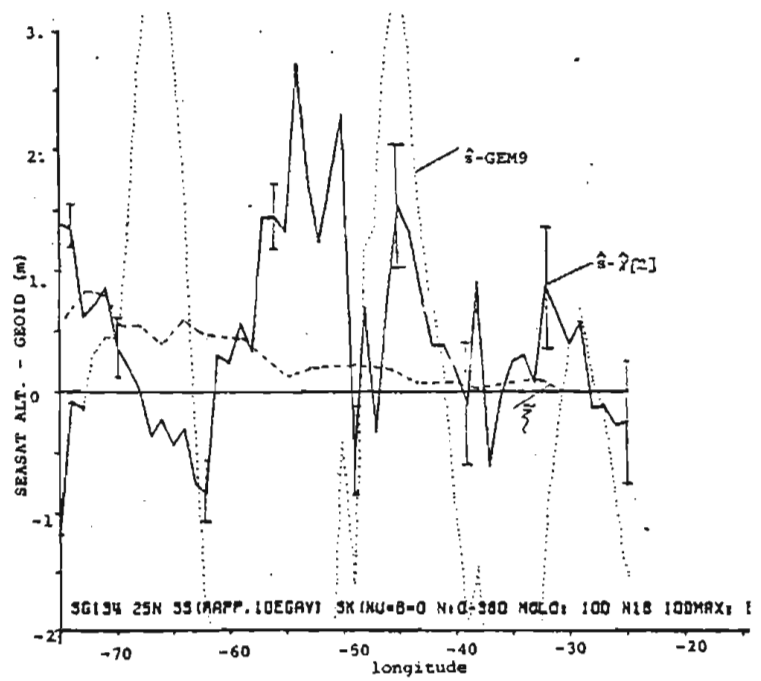

FIGURE 4-9: Profiles along 25°N. Means and s.d. of s-GEM9 and s-γ[2] are (-75,244) and (50,70) cm. Although the total variance decreases substantially when γ[2] is substracted, the difference s-γ[2] bears no oceanographic resemblance.

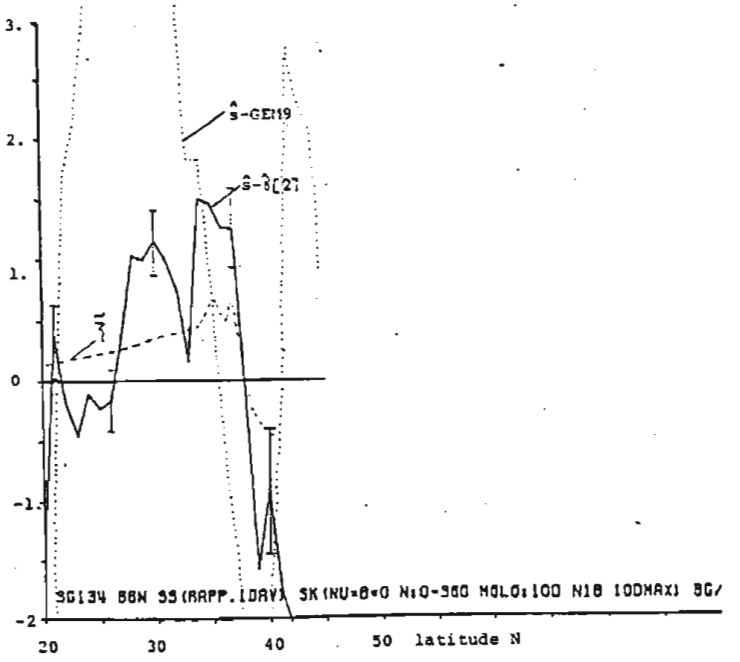

FIGURE 4-10: Profile along 66°W. Mean and s.d. of s-GEM9 and s-γ[2] are (116,379) and (-24,140) cm respectively. The Gulf Stream slope -at 38°N-is completely masked by a large error in γ[2]. Expected total errors underestimate the actual discrepancies.

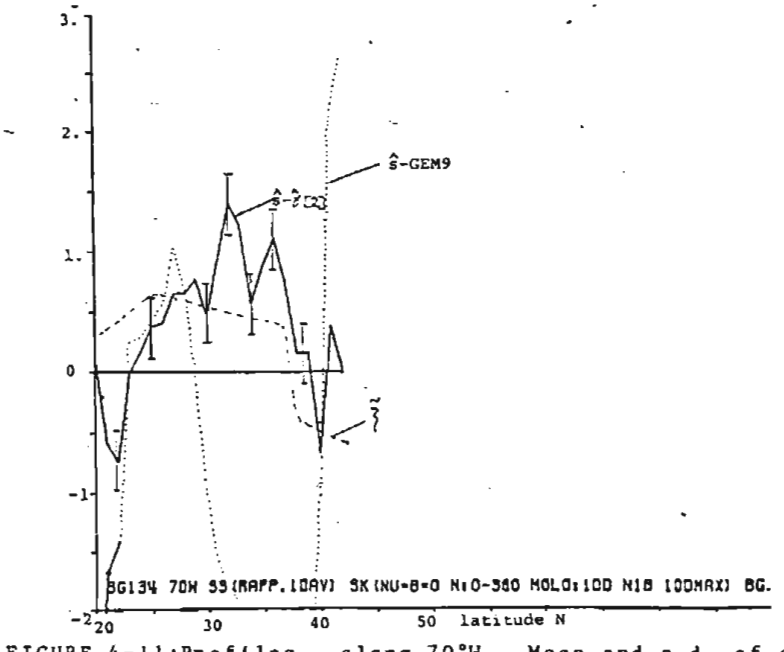

FIGURE 4-11: Profiles along 70°W. Mean and s.d. of s-GEM9 and s-γ[2] are (-96,170) and (40,57) cm respectively. This profile is somewhat puzzling: the Gulf Stream slope -at 38°N- is clearly defined, but the gentle southward slope in s-γ[2] -south of 35°N- is an artifact of the Puerto Rico trench -at 20°N-. The trench first enters the computations at 30°N, because of the 10° radius of integration.

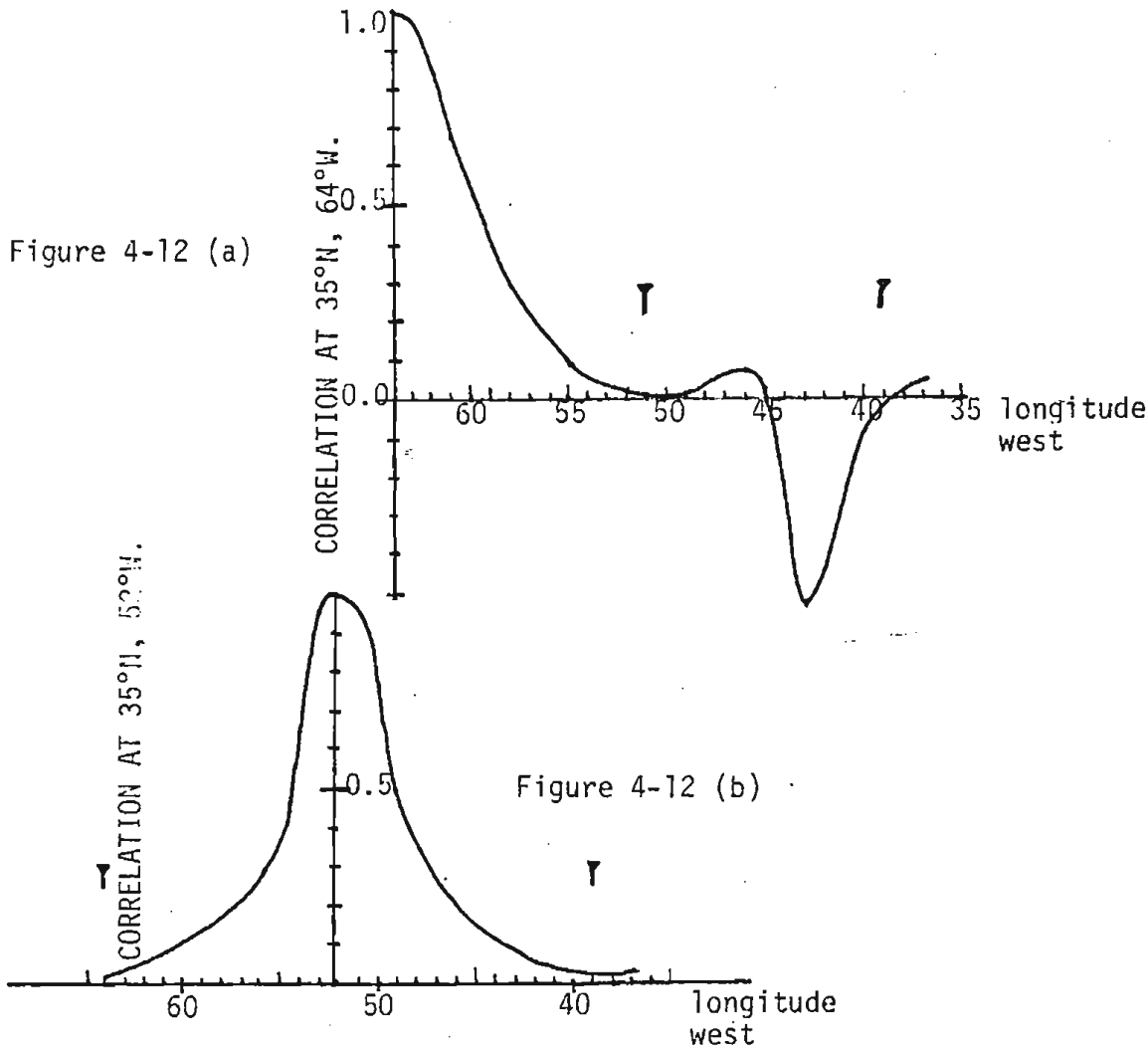

Figure 4-12 (a)

Figure 4-12 (b)

FIGURE 4-12. Examples of expected geoid error correlation, at 35°N. At this latitude, 1° of longitude ≈ 0.82° of distance ; ▼ indicate distances of 10°, the radius ofintegration.

The correlations a) are not the same in all directions around a point; b) change as the availability of data changes. c) if properly established, can help distinguish between ζ and geoid error of comparable size, in the same sense as two sinewaves of sufficiently different wavelength can be isolated from a time series.

systematic error is strongly correlated, over distances of up to 10°, and so is ζ).

Figures 4-8 through 4-11 show profiles along 25°N, 30°N, 66°W and 70°W; they display characteristics similar to those already discussed and they are pointed out in their captions. The spatial correlation between total errors also varies with position and direction: figure 4-12 shows two examples.

4.5 GEOIDAL ESTIMATE: PART 2

Another geoidal estimate, $\gamma[3]$, was computed using DATA-2. The only difference from $\gamma[2]$ is that gridded gravity was optimally interpolated at empty nodes prior to numerical integration (see Appendix 4-1).

Figure 4-13 shows a map of s-GEM9 and figure 4-14 shows a map of s-$\gamma[3]$. The main points to note about $\gamma[3]$ are:

- -the mean difference s-$\gamma[3]$ is -1 cm.
- -the rms difference is 184 cm; prior to optimal interpolation it was 197 cm; s-GEM9 has 292 cm rms.
- -the maximum discrepancies are ± 800 cm (prior to interpolation: -1200, +900 cm)

Figure 4-15 shows expected rms errors for $\gamma[3]$; a few points about this map. 1) The errors for $\gamma[3]$ were computed as if no optimum interpolation of empty nodes had been performed (i.e., they are the same expected errors of $\gamma[2]$). The reason was computational economy: with this simplification, the strong correlation between the error of an unoccupied node and the error of a neighbouring occupied node is

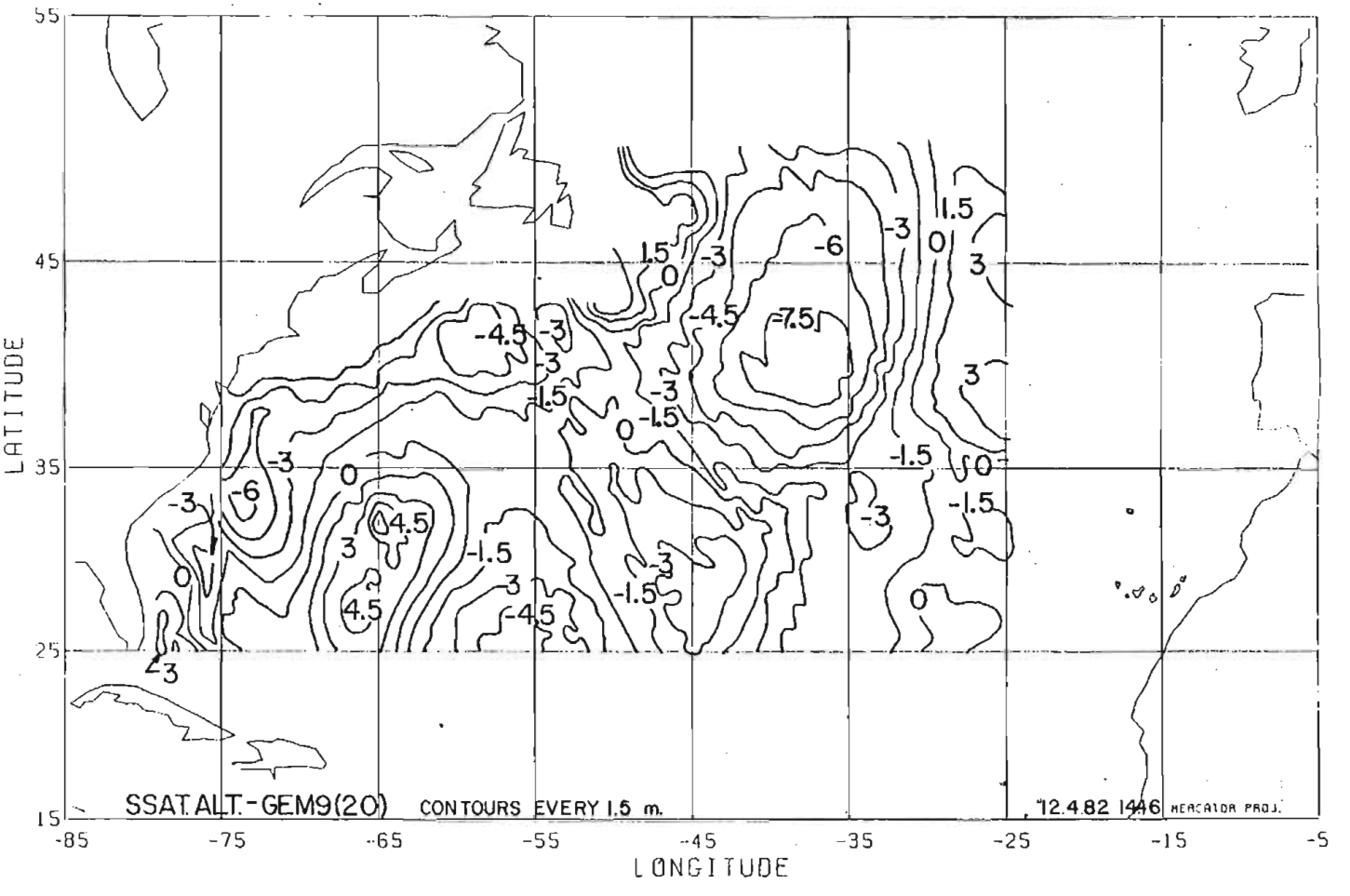

Figure 4-13. Difference between altimetric surface (s) and the GEM-9 geoid to degree 20, averaged over 1°x1° area bins. Labels in meters. The mean of this surface is -1.30 m, the standard deviation is 2.92 m. Rapp (1982) has argued that the bias indicates a difference between the radius of the Geodetic Reference System 80, to which s is referred, and the radius of the best fitting ellipsoid. Most of the energy in this map is the gravity field at length scales between 2000 and 100 km.

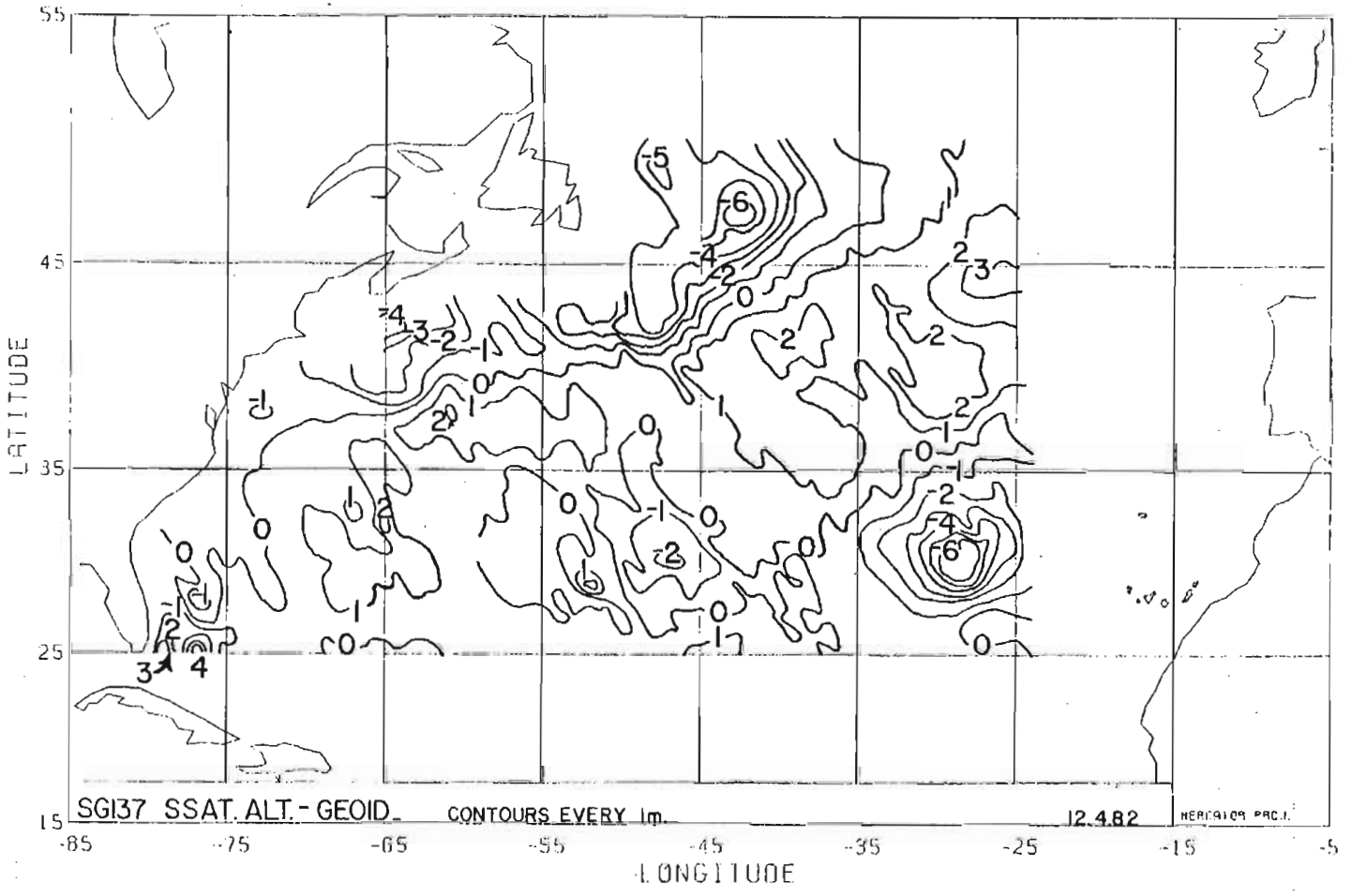

Figure 4-14: Difference between altimetric surface (s) and the geoid computed in this paper (γ). Contours every 1 m. The mean of $s-\gamma$ is -0.01 m, the standard deviation is 1.84 m. The power in this surface is much smaller than that in Figure 4, implying that the gravity data provide significant information about the geoid. However, the power in $s-\gamma$ is still much larger than expected oceanographic signals.

automatically taken into account; the more precise error computation would have required storing a huge number of correlation coefficients. This simplification tends to overestimate expected errors. 2) The Wagner and Colombo spectrum underestimates power at short wavelengths (compare with Brammer and Sailor (1982)). This feature tends to underestimate expected errors. 3) The approximately (30 cm)^2 residual variance due to errors in GEM9 was not added, because we intend to use this map in Chapter 5 as if the errors were uncorrelated over distances greater than 100 km (see, however, Fig. 4-12), but GEM-9 errors are strongly correlated over distances less than 2000 km. 4) The errors were not computed at all grid nodes, because of their computational expense. Previous to contouring, a simple interpolation using a weight = 1/distance^2 was performed.

The expected errors computed in this chapter are needed to estimate ζ, hence it is necessary to assess how good they are. Let σ_γ denote the expected errors of figure 4-15. Figure 4-16 shows the relative differences $(s-\gamma[3])/\sigma_\gamma$. If the σ_γ were reasonably accurate, if ζ were negligible, and if the differences were approximately gaussian, then the rms of figure 4-16 would be 1, and 95% of the values would be under 2. In fact the rms is 2.1, and 95% of the values are under 5. When σ_γ is replaed by $\sqrt{\sigma\gamma^2+(30\text{cm})^2}$, the results change negligibly (30 cm is the rms of ζ, a point argued in Chapter 5). The apparent conclusion is that σ_γ underestimates the actual errors by a factor of about 2, but follows their

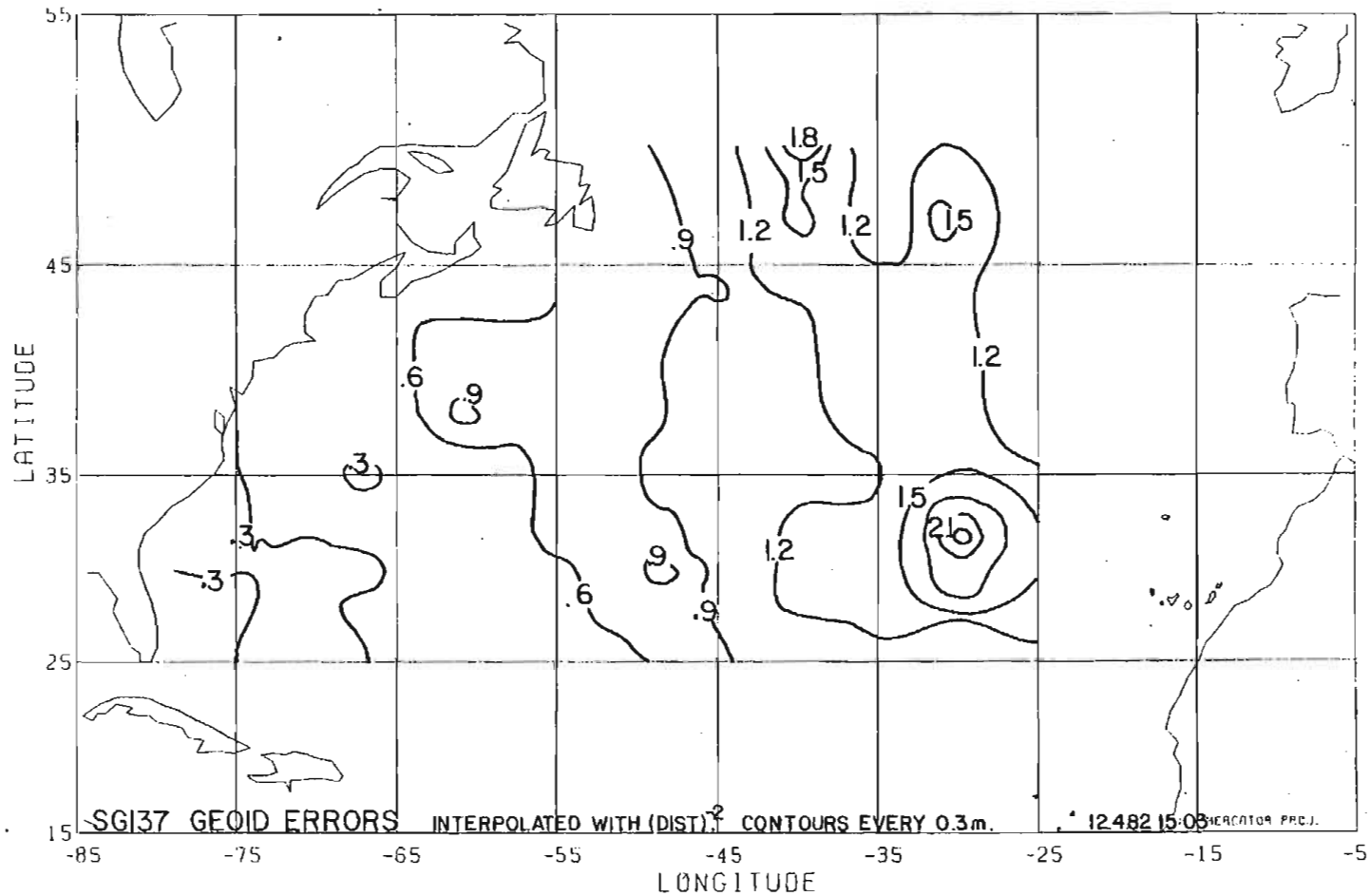

Figure 4-15. Expected errors σ_γ in the geoid γ, using the Wagner and Colombo (1979) spectrum to describe the average geoid behaviour. Contours every 0.3 m.

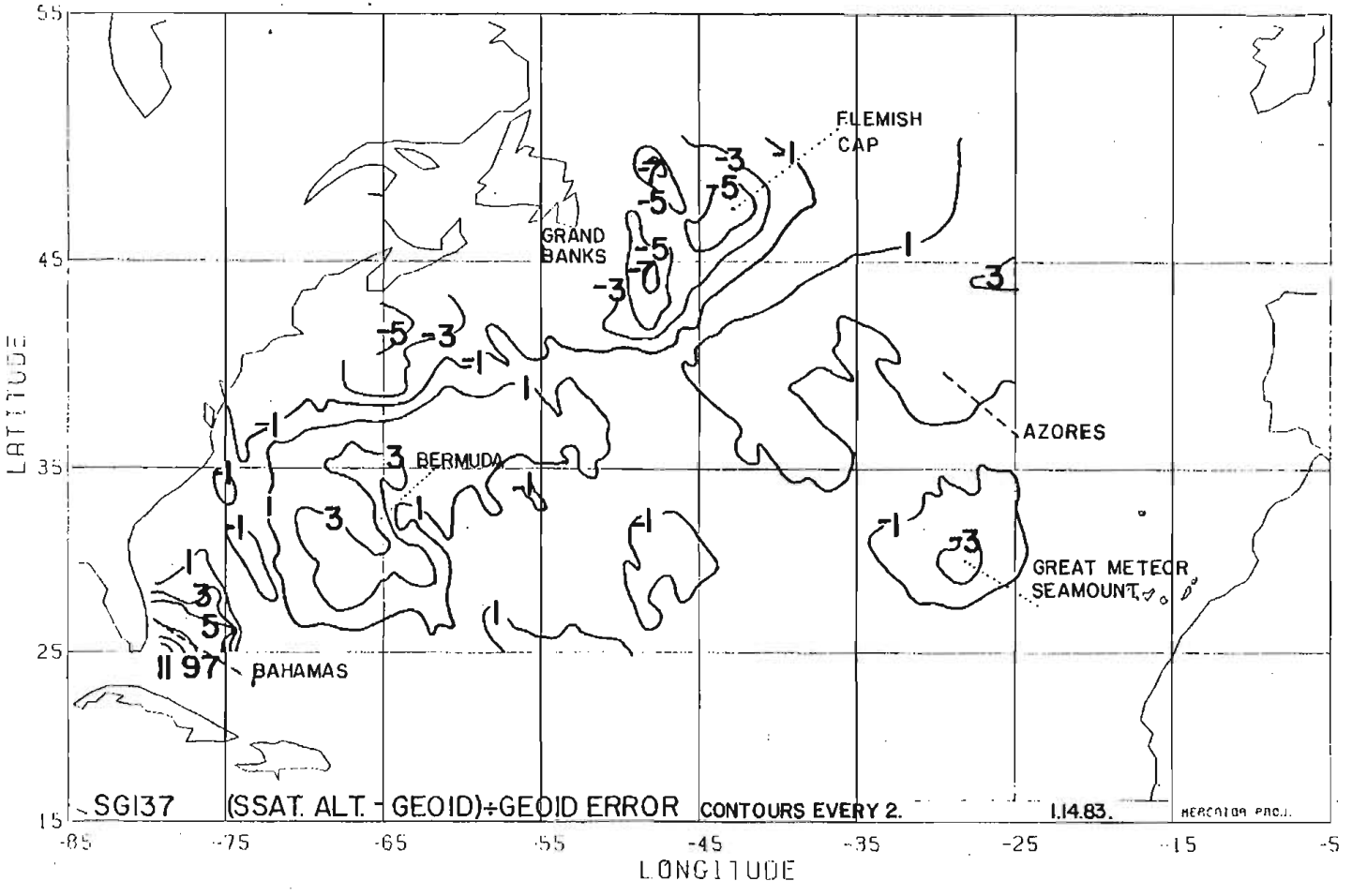

Figure 4-16: $(s-\gamma)/\sigma_\gamma$. Contours every factor of 2. If σ_γ accurately described the errors in γ, if ζ and errors in s were negligible, and if $(s-\gamma)/\sigma_\gamma$ were gaussian, then the rms of this surface would be 1. The rms of this surface is 2.1. The largest discrepancies are associated with obvious sources of short wavelength power in the gravity field; at these sites, $(s-\gamma)/\sigma_\gamma$ can be expected to exceed 1 even if the σ_γ had been computed using the exact spectrum of γ, but the Wagner and Colombo spectrum also underestimates power at short wavelengths.

change with position closely. The large discrepancies in $(s-\delta)/\sigma_\delta$ around the Bahamas and Bermuda (figure 4-16) could be ascribed to a) the expected geoid spectrum, that underestimate power at short wavelengths, b) the fact that no 'indirect effect' corrections (Chapter 2, section 1) were performed for the island masses above the geoid. The fact that the area around the Great Meteor seamount, with no masses above the geoid, also has large errors, suggests that a) is the dominant error, even though b) may not be negligible.

4.6 SUMMARY

The accuracy of marine gravity data is such that 60% of all errors are under 10 mgals and 95% are under 45 mgals. When such data are averaged over areas 0.5° in diameter, and subsequently used to estimate an equally averaged geoid, the influence of data noise produces geoid noise between 20 and 40 cm, occasionally 60 cm (all errors refer to the wavelength band 2000 km to 100 km).

The coverage of publicly available marine gravity data in the North Atlantic is such that small areas, a few degrees in diameter, and void of data, are scattered throughout. When the inverse discussed in Appendix 4-1 is used to estimate a filtered geoid, and the filtered gravity data are only given at grid spacings of 0.5°, no more than 11 cm rms resolution error is committed if all grid nodes inside a 10° radius are occupied. Due to the empty areas, however, this error climbs

in places to 250 cm RMS. Because in regions of scarce data less attenuation of gravity errors occurs, this high systematic error is compounded by the presence of noise of 30 or 40 cm RMS due to gravity measurement noise. Although optimum interpolation of filtered gravity in empty nodes reduces the systematic component of error, the total error is not reduced drastically. This implies that this data set, which represents almost all publicly available gravity over the North Atlantic up to 1981, is unable to recover oceanographic signals over a large part of the ocean. The discrepancies between altimetric and geoidal surfaces agree with the computed expected errors if the latter are doubled. The likely reason for this underestimate is the power spectrum used to describe the average geoid behaviour, which underestimates power at short wavelengths.

APPENDIX 4-1: METHOD OF COMPUTATION

This appendix complements the second section of the chapter by presenting the details of the geoid and geoid error computations.

(I)-Empty nodes in the grid shown in figure 4-4 and lying in oceanic regions were optimally interpolated from neighbouring data. The covariance of filtered gravity anomalies $D_{ii'}$ was computed by:

$$D_{ii'} = \sum_{n=2}^{360} (2n+1)\, F_n^2\, A_n^2\, M_n\, Pn(\cos_{ii'}) \qquad (4\text{-}1\text{-}1)$$

$A_n = \Gamma\ (n-1)$; $\Gamma = 981$ gals

F_n = coefficients of the spherical filter described in appendix 3-3, with =50,000

M_n = degree variances of the geopotential, as estimated by Wagner and Colombo (1980).

$_{ii'}$ = spherical distance between points i, i'.

P_n = unnormalized Legendre polynomial, degree n.

The interpolation formula is

$$g_p = \sum_i \sum_{i'} Dip([D+E]^{-1})_{ii'}\, g_i$$

where g_p: interpolated gravity; g_i: gridded gravity data

E: error covariance of g

D: given by 4-1-1

Error estimates for the interpolated values were computed (equation 2-17), but their correlations were not.

(II)-Each geoidal height $\hat{\gamma}_p$ was computed as a weighted average of gridded filtered gravity d_i available within 10° of position p.

$$\hat{\gamma}_p = \gamma_p^\circ + \frac{\sum_i S'_{pi}\,(\tilde{d}_i - d_i^\circ)}{\sum_i S'_{pi}} \cdot \int_{i'} S'_{pi'}\, d\sigma_{i'} \qquad \begin{matrix}\psi_{ip} < \bar{\psi} = 10^\circ \\ \psi_{i'p} < \bar{\psi} = 10^\circ\end{matrix} \qquad (4\text{-}1\text{-}2)$$

The kernel function S' is a modified Stokes function detailed in point III; the integral in 4-1-2 is the mean value of S' inside a cap of given radius. The values δ° and d° are geoidal heights and gravity anomalies, respectively, computed from the GEM-9 coefficients (Lerch et al., 1979) up to degree 20.

(III)-Molodenskii's modification of Stokes function (Molodenskii et al., 1962, section VII-4; see also Jekeli, 1981) satisfies the following property:

$$\text{minimize} \left| \int_{\text{sphere}} S_{pi}\, g_i\, da_i - \int_{\text{cap}} (S_{pi}=S_{pi}) g_i\, da_i \right|^2 \qquad (4\text{-}1\text{-}3)$$

here S_{pi} is Stokes function (equation 4-1-5, but the upper limit of Σ is ∞) and g_i is the errorless gravity anomaly at point i, hence the first integral is the exact value of the geoid at point p (in spherical approximation). The second integral is only performed over a cap around p, i.e., $\psi_{ip} < \bar{\psi}$; the function S is chosen to satisfy 4-1-3. If our data are errorless, continuous data around p, then such a choice minimizes the error of γ.

Let:

$$S'_{pi} = S_{pi} - \tilde{S}_{pi}$$

where S is the usual Stokes function, here truncated at degree 360 to take into account the discrete sampling every 0.5°

$$S_{pi} = (R/\gamma) \sum_{n=2}^{360} (2n+1)\ (n-1)\ P_n(\cos(\psi_{pi})) \qquad (4\text{-}1\text{-}4)$$

and Molodenskii's correction term is of the form

$$S_{pi} = (R/\gamma) \sum_{n=2}^{n_o} (2n+1)\ S_n\ P_n(\cos\psi_{pi}) \qquad (4\text{-}1\text{-}5)$$

n_o=10 was used, in conjunction with GEM-9 up to degree 20. There is no best n_o; if it is too low the influence of remote gravity is not compensated sufficiently; if n_o is too large, the modified kernel tends to amplify data noise in its attempts to completely remove the influence of remote gravity. The value n_o=10 was chosen on the basis of Jekeli's (1981) analysis.

(IV)- The coefficients $\tilde{S}_n$ were computed with the following algorithm, discussed by Jekeli (1981). The only reason I reproduce the equations used in the computations, is that equation 33 of that article is incorrect.

$$\bar{x} = \cos\bar{\psi}$$
$$k = (1/2)(1+\cos\bar{\psi})$$

$$S_n = (1/2) \sum_{l=n}^{n_o} (2l+1)\ u_l\ h_{ln} \qquad (4\text{-}1\text{-}6)$$

$$u_l = (1/k) \sum_{m=0}^{n_o} (2m+1)/2\ h_{lm}\ Q_m(\bar{\psi}) \qquad (4\text{-}1\text{-}7)$$

$$h_{jn} = \begin{cases} 0 & j<n \\ (2n+1)\, k^{-j} \sum_{i=0}^{p} \binom{p}{i} \binom{q}{i} (1-k)^{i+1} & j>n\geq 0 \\ 2/(2j+1)\, k^{-j} & j=n\geq 0 \end{cases} \qquad (4\text{-}1\text{-}8)$$

$$p = j-n-1; \quad q = j+n$$

The Q_m (called Molodenskii's truncation coefficients) were computed with Hagiwara's algorithm (1976).

(V)-Equation 4-1-2 can be written in a form more suitable for error computations as

$$\gamma_p - \gamma_p^o = \sum_i B_{pi} (d_i - d_i^o) \qquad (4\text{-}1\text{-}8)$$

In this equation the sum was assumed to occur only over available data, and interpolated values were not considered data for the purpose of error computation. The error covariance of the right hand side of 4-1-8 can be estimated by:

$$E'' = B(D+E_D)B^T + M - BC - (BC)^T \qquad (4\text{-}1\text{-}9)$$

(e.g., Moritz, 1976), where D is the covariance of $(d-d^o)$ and E_D its error covariance, M is the covariance of the γ_p, and C the crosscovariance between $(d-d^o)$ and m. These were computed as:

$$M_{pp'} = \sum_{n=2}^{360} (2n+1)\, M_n\, F_n^2\, P_n(\cos\psi_{pp'})$$

$$D_{ii'} = \sum_{n=2}^{360} (2n+1)\, F_n^2\, A_n^2\, M_n\, Pn(\cos\psi_{ii'}) \qquad (4\text{-}1\text{-}10)$$

$$C_{ii'} = \sum_{n=2}^{360} (2n+1)\ F_n^2\ A_n\ M_n\ Pn(\cos\psi_{ii'})$$

Equation 4-1-9 only describes the error in $\delta - \delta_o$. To obtain the error in the geoidal estimate, the error covariance of the reference field (GEM-9) must be added. The influence of GEM-9 errors in equation 4-1-8 is less than the error due to the coefficients, if there are indeed gravity data supplying additional information. Jekeli (1981, figure 2) estimated the residual error for the modified Stokes function used here to be 30 cm. This error has a correlation length larger than the 10^o radius of integration.

APPENDIX 4-2: CRUISE CROSSOVERS

This appendix summarizes the results of the crossover analysis between gravity cruises. The following pages are a computer printout, and the meaning of the columns is as follows:

Column 1 (CR #): cruise number, an identifier with meaning only within this work.

Column 2 (CR NAME): cruise name, as specified by originating institution .

Column 3 (NX'S): number of crossovers found for this cruise, including self-crossings.

Column 4 (MEAN): arithmetic mean of all crossovers for this cruise, in milligals.

Column 5 (S.D.): standard deviation of all crossover discrepancies for this cruise.

Column 6 (SELFNX): number of crossings with itself.

Column 7 (MEAN) : mean of self-crossings.

Column 8 (S.D.) : standard deviation of self crossings.

Column 9 ( ) : number of the cruise with which the current cruise has the largest negative crossover discrepancy.

Column 10 (MIN): value of largest negative discrepancy

Column 11 ( ) : same as 9 p or the maximum discrepancy.

Column 12 (MAX): value of largest positive discrepancy.

Column 13 : logical flag. If 'F', cruise was not used for geoid computations.

XOVMSD2 7JUN82 22:40
CROSSOVER MEAN & SD, EXCLUDING CRUISES LABELLED 'F'. FCR1 AND FCR2 JOINED./
-- 06/08/82 01:03:51 (EDT)--

CR#	CR NAME	NX'S	MEAN	S.D.	SELFNX	MEAN	S.D.		MIN		MAX	
(1)	V2305	34	-5.0	13.9	3	-2.1	4.4	(184)	-37.	(60)	37.	T
(2)	V2306	189	-0.7	14.8	32	-0.0	0.6	(203)	-36.	(37)	49.	T
(3)	V2307	65	-8.7	13.3	0	0.0	0.0	(67)	-55.	(53)	32.	T
(4)	V2501	224	7.1	12.7	31	8.9	13.4	(162)	-44.	(130)	50.	T
(5)	V2502	139	1.6	14.1	1	3.2	0.0	(177)	-39.	(177)	48.	T
(6)	V2503	176	-12.4	12.7	2	-0.6	0.5	(185)	-47.	(129)	24.	T
(7)	V2504	95	-8.5	8.8	3	-2.5	5.2	(123)	-31.	(39)	19.	T
(8)	V2608	210	2.0	12.8	84	0.4	2.0	(159)	-64.	(142)	84.	T
(9)	V2609	240	-2.5	7.6	36	0.3	1.7	(141)	-28.	(23)	23.	T
(10)	V2610	12	-9.5	10.6	0	0.0	0.0	(49)	-21.	(73)	15.	T
(11)	V2713	44	5.4	10.8	3	1.9	2.3	(215)	-23.	(68)	36.	T
(12)	V2714	91	-1.4	10.9	3	0.6	0.4	(162)	-38.	(43)	21.	T
(13)	V2801	63	-4.7	10.4	0	0.0	0.0	(162)	-34.	(36)	27.	T
(14)	V2804	27	10.9	8.0	0	0.0	0.0	(82)	-8.	(20)	25.	T
(15)	V2805	35	5.7	18.1	1	0.1	0.0	(184)	-24.	(68)	63.	T
(16)	V2806	50	-1.0	20.3	0	0.0	0.0	(162)	-92.	(130)	50.	T
(17)	V2807	204	1.9	17.6	18	-3.7	8.9	(40)	-42.	(130)	97.	T
(18)	V2908	111	-2.2	11.0	8	-1.1	2.5	(83)	-31.	(68)	47.	T
(19)	V2909	77	-1.9	7.7	0	0.0	0.0	(27)	-35.	(60)	19.	T
(20)	V2911	81	-1.7	8.5	3	-4.7	7.3	(14)	-25.	(93)	29.	T
(21)	V2912	37	0.3	16.0	0	0.0	0.0	(162)	-61.	(37)	38.	T
(22)	V3001	96	-0.0	11.1	2	2.9	6.4	(182)	-24.	(3)	29.	T
(23)	V3002	168	1.9	11.8	0	0.0	0.0	(9)	-23.	(132)	41.	T
(24)	V3003	132	2.2	13.3	0	0.0	0.0	(75)	-45.	(85)	56.	T
(25)	V3004	53	1.0	10.2	3	-2.2	1.4	(94)	-24.	(62)	26.	T
(26)	V3005	138	3.8	11.6	0	0.0	0.0	(203)	-43.	(215)	33.	T
(27)	V3006	209	3.6	15.2	53	0.0	9.8	(28)	-29.	(200)	51.	T
(28)	V3007	115	7.9	13.0	31	-2.0	7.5	(184)	-24.	(200)	54.	T

(29)	V3008			41	2.4	6.2	3	1.4	1.5	(184)	-20.	(13)	17.	T
(30)	V3009			66	0.6	5.4	24	-0.6	4.9	(30)	-14.	(20)	17.	T
(31)	V3012			36	-5.8	12.2	2	-2.5	13.8	(90)	-34.	(60)	16.	T
(32)	V3013			47	4.4	15.6	19	10.2	11.0	(213)	-37.	(32)	33.	T
(33)	A2	73	2	25	-10.3	15.3	0	0.0	0.0	(162)	-45.	(67)	24.	T
(34)	CH	36		212	-2.0	17.3	18	-1.2	8.1	(40)	-43.	(68)	113.	T
(35)	CH	43	1	19	2.2	33.5	0	0.0	0.0	(184)	-40.	(43)	83.	T
(36)	CH	43	3	15	-13.8	14.6	0	0.0	0.0	(209)	-38.	(79)	11.	T
(37)	CH	44		306	-0.5	27.6	34	-4.4	23.2	(37)	-65.	(171)	102.	T
(38)	CH	46		383	-2.9	23.7	39	2.6	30.3	(158)	-127.	(38)	75.	F
(39)	CH	55		194	-1.1	16.0	0	0.0	0.0	(162)	-77.	(176)	40.	T
(40)	CH	57		330	8.3	25.3	45	-10.2	13.8	(61)	-116.	(175)	110.	T
(41)	CH	61	1	13	8.3	11.0	0	0.0	0.0	(80)	-6.	(63)	29.	T
(42)	CH	34		431	4.6	35.4	28	4.0	41.5	(45)	-92.	(142)	197.	F
(43)	CH	96	5	48	-25.6	29.3	1	-0.2	0.0	(162)	-148.	(67)	14.	T
(44)	A2	54	3	136	-1.9	16.6	0	0.0	0.0	(162)	-64.	(161)	67.	T
(45)	CH	99	1	122	0.9	16.1	0	0.0	0.0	(17)	-35.	(142)	90.	T
(46)	CH	61	2	19	5.3	24.3	0	0.0	0.0	(35)	-71.	(43)	31.	T
(47)	A2	75	1	68	-4.5	14.9	0	0.0	0.0	(162)	-65.	(161)	38.	T
(48)	CH	115	1	54	15.1	21.7	0	0.0	0.0	(184)	-24.	(88)	67.	T
(49)	CH	115	9	116	3.7	25.6	0	0.0	0.0	(94)	-72.	(161)	82.	T
(50)	A2	92	1	171	1.2	7.2	95	0.0	3.4	(51)	-23.	(34)	32.	T
(51)	A2	93	1	129	4.3	10.2	0	0.0	0.0	(69)	-29.	(6)	30.	T
(52)	A2	92	2	161	5.3	12.9	58	-2.3	4.8	(124)	-32.	(72)	48.	T
(53)	CH	39	1	127	-6.4	20.1	16	1.4	2.8	(69)	-91.	(130)	43.	T
(54)	V1712			27	-16.0	11.8	3	-0.1	0.1	(87)	-32.	(62)	4.	T
(55)	V1713			43	-7.8	11.2	0	0.0	0.0	(184)	-34.	(215)	31.	T
(56)	C0801			146	-4.1	19.5	5	-2.6	7.3	(177)	-74.	(171)	58.	T
(57)	C0809			239	3.1	39.5	71	8.9	49.0	(57)	-115.	(161)	158.	F
(58)	C0812			130	-5.4	12.8	1	1.0	0.0	(40)	-45.	(44)	41.	T
(59)	C0912			39	-1.1	19.4	2	1.1	3.3	(91)	-51.	(91)	69.	T
(60)	C0913			87	-7.2	13.5	1	0.1	0.0	(162)	-66.	(63)	26.	T
(61)	C1001			196	2.6	16.4	2	0.2	0.1	(45)	-48.	(40)	116.	T

(62) C1311	151	-10.7	11.1	14	0.4	2.2	(94)	-50.	(65)	12.	T
(63) V1717	31	-4.3	16.3	0	0.0	0.0	(88)	-40.	(40)	29.	T
(64) V1802	100	0.4	10.9	5	-0.6	1.2	(40)	-47.	(142)	30.	T
(65) V1803	117	-3.7	21.6	0	0.0	0.0	(94)	-83.	(95)	46.	T
(66) V1818	98	-1.2	16.3	7	0.0	0.2	(95)	-53.	(175)	52.	T
(67) V1819	72	-5.9	20.7	1	-0.1	0.0	(162)	-79.	(3)	55.	T
(68) V1913	46	-13.8	23.8	1	0.4	0.0	(34)	-113.	(67)	55.	T
(69) V2012	28	23.2	23.4	0	0.0	0.0	(203)	-24.	(53)	91.	T
(70) V2013	11	18.8	21.9	0	0.0	0.0	(182)	-5.	(53)	73.	T
(71) C1509	100	3.0	12.4	0	0.0	0.0	(94)	-41.	(65)	28.	T
(72) C1510	75	-3.5	13.4	19	0.9	3.0	(52)	-48.	(6)	14.	T
(73) C1601	65	-13.9	23.3	3	-10.3	17.9	(40)	-45.	(130)	128.	T
(74) C1612	138	7.8	19.3	1	-0.8	0.0	(94)	-45.	(161)	115.	T
(75) C1613	207	-0.1	14.1	22	0.6	1.6	(94)	-111.	(160)	58.	T
(76) C1701	137	2.9	12.0	1	0.0	0.0	(37)	-58.	(161)	47.	T
(77) C1702	123	-2.1	19.6	16	15.1	19.6	(74)	-51.	(77)	41.	T
(78) V2207	87	9.9	20.4	0	0.0	0.0	(162)	-105.	(130)	56.	T
(79) V2301	53	-5.6	11.0	0	0.0	0.0	(209)	-25.	(37)	41.	T
(80) V2302	38	8.8	10.1	0	0.0	0.0	(81)	-11.	(60)	30.	T
(81) V2303	116	-0.1	6.8	77	0.5	6.8	(81)	-21.	(81)	15.	T
(82) V2304	22	-0.8	8.8	6	0.0	3.2	(14)	-17.	(89)	18.	T
(83) V3014	100	-3.7	10.6	0	0.0	0.0	(203)	-34.	(18)	31.	T
(84) V2201	218	0.3	12.2	6	1.2	1.0	(40)	-50.	(73)	40.	T
(85) V2202	119	-3.5	20.5	0	0.0	0.0	(37)	-56.	(171)	93.	T
(86) V2401	163	3.6	11.2	10	-0.1	0.2	(39)	-33.	(175)	53.	T
(87) V2604	125	1.1	6.0	111	0.6	3.5	(25)	-13.	(54)	32.	T
(88) V2701	27	4.3	19.0	0	0.0	0.0	(48)	-67.	(63)	40.	T
(89) V2706	25	-2.7	8.1	0	0.0	0.0	(14)	-19.	(19)	14.	T
(90) V2707	44	2.5	13.1	8	-0.7	1.5	(208)	-27.	(31)	34.	T
(91) V2708	19	-9.8	25.9	1	-0.8	0.0	(59)	-69.	(59)	51.	T
(92) V2709	77	1.4	6.5	1	1.3	0.0	(205)	-7.	(165)	39.	T
(93) V2702	40	0.2	9.0	0	0.0	0.0	(20)	-29.	(13)	21.	T
(94) V2607	233	23.6	25.8	61	-1.6	3.2	(71)	-18.	(171)	128.	T

(95) V2413	220	-0.7	19.9	0	0.0	0.0	(40)	-67.	(40)	68.	T
(96) C0901	215	-5.8	13.6	5	6.6	7.2	(95)	-56.	(153)	38.	T
(97) C0902	86	-4.1	11.3	33	0.8	7.3	(159)	-38.	(147)	20.	T
(98) V2101	92	6.6	22.5	1	-0.4	0.0	(95)	-24.	(130)	159.	T
(99) UNGEOIL3	242	-55.2	97.6	14	3.1	32.4	(44)	-352.	(202)	132.	F
(100) UNGEOIL4	141	-5.7	25.4	24	11.0	28.1	(142)	-82.	(100)	76.	F
(101) UNGEOIL6	123	-8.7	28.5	0	0.0	0.0	(56)	-137.	(155)	47.	F
(102) ME 2T	100	2.9	7.8	2	-1.6	0.6	(208)	-21.	(215)	33.	T
(103) 69006011	17	-35.0	30.0	0	0.0	0.0	(208)	-102.	(59)	5.	F
(104) 69005611	19	-28.6	29.7	4	0.0	0.1	(60)	-105.	(1)	20.	F
(105) 69005621	23	-37.2	26.3	3	0.2	0.3	(29)	-75.	(105)	0.	F
(106) 69005631	21	-49.1	42.3	0	0.0	0.0	(208)	-130.	(60)	15.	F
(107) 69005641	30	-26.6	37.7	4	-0.3	0.4	(90)	-179.	(59)	14.	F
(108) 69005651	8	-3.9	6.9	4	0.8	0.4	(213)	-16.	(213)	1.	F
(109) 72001011	20	-52.9	87.4	1	-12.6	0.0	(59)	-335.	(59)	15.	F
(110) 72001021	30	-58.4	42.3	3	0.3	0.8	(208)	-156.	(110)	1.	F
(111) 72004811	26	-45.1	40.9	14	-12.2	23.4	(200)	-100.	(111)	19.	F
(112) 72001111	16	-45.8	35.2	2	-0.5	0.1	(90)	-112.	(112)	-0.	F
(113) DICPVERD	18	-33.9	43.3	2	7.1	29.1	(78)	-115.	(113)	28.	F
(114) DIATLNFZ	40	0.9	5.7	0	0.0	0.0	(204)	-12.	(205)	11.	T
(115) DIATTRAV	64	-2.1	13.8	0	0.0	0.0	(78)	-38.	(131)	23.	T
(116) DIVENCZ	19	-3.3	20.9	12	-1.8	8.6	(142)	-65.	(155)	30.	T
(117) DILANTLS	36	-39.4	48.6	5	1.9	4.2	(65)	-157.	(160)	24.	F
(118) KEA06-69	106	6.9	12.4	26	-3.0	4.3	(157)	-20.	(153)	37.	T
(119) KEA07-69	4	2.9	5.8	0	0.0	0.0	(138)	-4.	(135)	10.	T
(120) KEA08-69	27	-7.1	24.7	0	0.0	0.0	(134)	-40.	(129)	65.	T
(121) KEA09-69	54	1.5	14.6	11	12.2	12.2	(96)	-30.	(121)	31.	T
(122) KEA10-69	79	3.0	8.3	3	10.2	6.8	(123)	-14.	(60)	24.	F
(123) KEA11-69	87	14.0	10.4	0	0.0	0.0	(76)	-6.	(6)	44.	T
(124) KEA01-70	34	3.9	17.0	9	0.9	3.4	(48)	-58.	(67)	44.	T
(125) KEA04-70	18	-3.7	18.9	0	0.0	0.0	(48)	-66.	(67)	22.	T
(126) KEA05-70	89	-2.0	14.1	0	0.0	0.0	(162)	-47.	(43)	25.	T
(127) KA343618	0	0.0	0.0	0	0.0	0.0	(0)	999.	(0)	-999.	T

(128)	WI933014	0	0.0	0.0	0	0.0	0.0	(0)	999.	(0)	-999.	T
(129)	TAG70	141	-9.7	17.3	1	-5.4	0.0	(185)	-73.	(73)	22.	T
(130)	TAG71IDO	287	-7.0	22.3	64	0.5	2.0	(98)	-159.	(2)	26.	T
(131)	BOMEXDI	151	-2.1	13.9	17	5.4	19.1	(74)	-46.	(131)	33.	T
(132)	KEA01-67	106	-2.8	19.8	11	-9.2	14.0	(135)	-67.	(133)	65.	T
(133)	KEA02-67	12	-26.4	20.1	0	0.0	0.0	(132)	-65.	(169)	1.	T
(134)	KEA06-68	56	-3.6	22.3	1	-4.1	0.0	(135)	-60.	(120)	40.	T
(135)	KEA07-68	66	14.5	19.7	3	6.2	0.7	(119)	-10.	(130)	82.	T
(136)	KEA08-68	118	-0.3	6.0	16	-1.0	4.5	(139)	-15.	(140)	12.	T
(137)	KEA01-69	156	2.8	14.6	23	-1.3	9.0	(135)	-42.	(153)	54.	T
(138)	KEA02-69	89	-0.7	5.9	24	-1.4	4.5	(137)	-24.	(139)	11.	T
(139)	KEA03-69	141	1.1	6.3	11	0.5	7.4	(169)	-14.	(96)	24.	T
(140)	KEA04-69	152	1.3	7.5	38	0.2	4.0	(141)	-16.	(9)	21.	T
(141)	KEA05-69	27	10.9	9.3	0	0.0	0.0	(175)	-3.	(9)	28.	T
(142)	C0903	203	-4.9	27.1	11	4.8	7.9	(172)	-94.	(161)	84.	T
(143)	WEL7502	125	-0.6	10.7	0	0.0	0.0	(147)	-26.	(6)	30.	T
(144)	WI343503	0	0.0	0.0	0	0.0	0.0	(0)	999.	(0)	-999.	T
(145)	LIS-BIS	38	1.7	3.1	38	1.7	3.1	(145)	-1.	(145)	16.	T
(146)	C0807	49	-0.2	25.4	5	0.3	0.4	(40)	-41.	(142)	76.	T
(147)	C1003	134	0.3	18.8	19	2.9	5.0	(142)	-56.	(142)	59.	T
(148)	C1012	249	-0.1	11.0	8	-1.0	0.6	(147)	-48.	(153)	51.	T
(149)	C1112	169	6.4	12.1	0	0.0	0.0	(147)	-35.	(155)	51.	T
(150)	C1201	154	0.3	15.4	1	-0.7	0.0	(153)	-61.	(153)	59.	T
(151)	C1309	4	-0.9	10.6	0	0.0	0.0	(152)	-16.	(142)	6.	T
(152)	C1310	121	0.6	13.1	3	4.4	5.6	(155)	-37.	(142)	28.	T
(153)	V1817	115	-13.9	23.8	1	30.1	0.0	(159)	-65.	(150)	61.	T
(154)	V1903	58	1.3	11.9	1	11.0	0.0	(8)	-22.	(176)	24.	T
(155)	V2002	143	0.7	27.5	0	0.0	0.0	(37)	-65.	(175)	84.	T
(156)	V2808	48	2.8	8.7	6	-0.2	0.3	(157)	-21.	(155)	21.	T
(157)	V2103	55	9.6	18.2	1	29.5	0.0	(150)	-33.	(153)	45.	T
(158)	V2114	223	-2.0	10.7	28	-0.3	3.2	(159)	-51.	(142)	29.	T
(159)	V2402	96	10.4	16.7	4	1.7	3.7	(150)	-25.	(153)	65.	T
(160)	EQUAP72	***	-4.9	15.2	143	-3.1	10.0	(37)	-88.	(40)	71.	T

(161)	CAG71ID0			696	-5.5	23.8	75	-4.0	14.9	(74)	-115.	(171)	83. T
(162)	A2	67	1	47	33.8	37.1	0	0.0	0.0	(184)	-23.	(43)	148. T
(163)	A2	73	1	32	1.7	13.1	0	0.0	0.0	(48)	-29.	(53)	26. T
(164)	V3205			115	0.9	6.5	6	-0.4	0.6	(204)	-15.	(2)	26. T
(165)	V3204			80	-7.5	12.0	2	5.2	9.9	(205)	-49.	(18)	28. T
(166)	C1603			129	2.1	10.9	5	0.0	0.0	(195)	-35.	(62)	29. T
(167)	C1604			44	0.2	8.2	36	-1.0	2.4	(167)	-7.	(56)	16. T
(168)	C1801			58	7.0	30.3	16	-10.6	25.5	(193)	-94.	(187)	113. T
(169)	C1902			93	4.1	14.7	2	20.6	18.3	(134)	-29.	(129)	38. T
(170)	C1903			177	-1.2	13.2	4	0.6	2.4	(175)	-33.	(175)	53. T
(171)	C1904			217	-3.3	24.7	3	22.0	11.5	(94)	-128.	(142)	85. T
(172)	C1906			196	-1.3	12.2	17	-0.2	0.7	(146)	-53.	(142)	94. T
(173)	C1907			117	-2.0	15.4	10	-0.0	0.1	(171)	-54.	(44)	46. T
(174)	C2101			166	3.0	11.8	11	6.6	9.1	(123)	-23.	(6)	39. T
(175)	C2102			244	-7.7	21.8	13	-12.3	16.7	(94)	-112.	(134)	42. T
(176)	C2103			147	-4.2	11.5	15	-2.7	6.1	(39)	-40.	(155)	28. T
(177)	C2104			203	3.1	17.8	81	8.9	11.0	(5)	-48.	(56)	74. T
(178)	C2105			6	9.0	10.5	0	0.0	0.0	(195)	-3.	(54)	28. T
(179)	C2106			8	3.6	18.5	1	-0.0	0.0	(193)	-24.	(188)	37. T
(180)	C2109			40	6.0	8.2	0	0.0	0.0	(171)	-14.	(146)	20. T
(181)	C2110			88	9.5	14.5	7	-0.0	0.1	(40)	-28.	(66)	35. T
(182)	C2111			113	7.5	10.7	35	0.3	1.7	(20)	-18.	(37)	49. T
(183)	C2112			20	-1.8	9.0	0	0.0	0.0	(14)	-24.	(93)	11. T
(184)	C2115			67	23.9	15.0	0	0.0	0.0	(35)	-32.	(43)	71. T
(185)	C2116			154	4.2	13.3	21	-0.3	0.6	(40)	-41.	(129)	73. T
(186)	I0775*			26	-4.4	14.9	2	-0.3	0.2	(189)	-30.	(187)	34. T
(187)	I1377*			66	-13.1	28.7	6	3.0	53.0	(168)	-113.	(187)	106. T
(188)	I1578*			41	-11.4	14.1	11	-3.8	5.2	(168)	-43.	(187)	15. T
(189)	I1678*			39	-2.5	13.2	3	-10.1	17.8	(189)	-31.	(186)	30. T
(190)	V2709			132	2.5	8.0	1	1.0	0.0	(94)	-23.	(165)	39. T
(191)	V3101			27	-0.1	8.2	0	0.0	0.0	(102)	-11.	(56)	24. T
(192)	V3102			46	2.6	6.0	33	0.9	2.8	(193)	-7.	(188)	22. T
(193)	V3103			47	16.0	25.0	4	10.6	6.2	(168)	-20.	(187)	111. T

(194)	V3104	37	5.0	10.4	19	-0.0	1.6	(178)	-9.	(54)	32. T
(195)	V3105	29	6.3	12.0	3	-1.7	1.7	(94)	-18.	(166)	35. T
(196)	V3106	126	1.3	12.4	7	1.7	4.9	(94)	-34.	(6)	32. T
(197)	V3107	216	-0.3	16.7	29	0.1	1.1	(94)	-81.	(161)	61. T
(198)	V3108	78	2.3	7.3	14	-0.1	1.3	(40)	-19.	(67)	38. T
(199)	V3201	80	-6.1	10.3	9	-0.0	1.3	(123)	-28.	(3)	16. T
(200)	V3203	209	-8.5	15.0	63	1.0	8.2	(28)	-54.	(91)	37. T
(201)	V3206	171	6.3	13.4	67	3.2	7.5	(5)	-27.	(130)	54. T
(202)	V3207	197	0.5	15.6	57	-1.8	3.5	(94)	-58.	(40)	106. T
(203)	SS005	111	8.6	19.1	0	0.0	0.0	(45)	-52.	(171)	56. T
(204)	SS006	111	6.6	10.5	0	0.0	0.0	(203)	-29.	(130)	49. T
(205)	SS007	100	-0.8	13.0	0	0.0	0.0	(184)	-32.	(165)	49. T
(206)	SS008	104	2.8	10.7	1	0.1	0.0	(134)	-34.	(72)	34. T
(207)	SS009	102	4.2	11.2	0	0.0	0.0	(184)	-30.	(129)	33. T
(208)	SS010	73	6.9	12.6	0	0.0	0.0	(59)	-41.	(32)	34. T
(209)	SS011	77	-0.9	16.9	0	0.0	0.0	(162)	-64.	(37)	40. T
(210)	SS012	6	-4.5	12.3	0	0.0	0.0	(184)	-23.	(60)	13. T
(211)	SS013	13	-5.7	13.9	0	0.0	0.0	(102)	-28.	(60)	26. T
(212)	SS014	14	-1.1	10.5	0	0.0	0.0	(80)	-23.	(60)	11. T
(213)	SS020	28	16.0	12.5	0	0.0	0.0	(91)	-8.	(91)	45. T
(214)	SS12A	9	0.2	5.9	0	0.0	0.0	(93)	-9.	(79)	7. T
(215)	IVK01	56	-3.0	12.5	1	1.1	0.0	(102)	-33.	(62)	38. T

* : indicates cruise is wholly in the southern hemisphere.

CHAPTER 5

ESTIMATION OF TIME AVERAGED CIRCULATION AND GEOID IMPROVEMENT

5.1 INTRODUCTION

This chapter describes a first approximation to the recovery of time averaged oceanographic differences between the altimetric and geoidal surfaces; it also discusses how to remove these features from the altimetric surface in order to obtain a closer estimate of the marine geoid.

The main simplification introduced is the neglect of the correlation between geoidal errors at different locations, a simplification made owing to computer limitations. The geoidal estimate labelled $\hat{\gamma}[3]$ and described in chaper 4 (section 4.5) is used here (see figures 4-14 and 4-15). The altimetric surface ($\hat{s}$) was also described in chapter 4, section 4.3.

In order to introduce as little oceanographic information as possible in the computations, a first estimate is obtained assuming that the geostrophic component of sea-surface topography is a spatially uncorrelated quantity with uniform variance throughout the ocean. This is not a very good description, but it allows us to check the resulting estimate of the circulation with one derived exclusively from hydrographic data.

A second estimate of the circulation is computed by combining the gravimetric geoid, the altimetric surface, and

an estimate of the geostrophic component based only on hydrographic data. This second estimate of the circulation is then removed from the altimetric data to yield an improved geoid.

5.2 ASSUMING SPATIALLY UNCORRELATED ζ

Most features of figure 4-14 would strike a physical oceanographer as artifacts of geoid error, particularly the 5 m low north of 40°. The simplest oceanographic information one has in mind is a statement that $s - \gamma$ has some 'reasonable' rms value, say 30 or 40 cm, implying reasonable maximum values some 3 to 5 times larger. In this section that is precisely the assumption made, with the further assumption that geostrophic heights separated by 100 km or more are uncorrelated.

With this particularly simple framework, the optimum estimation of ζ, the geostrophic discrepancy in $s-\gamma$, simplifies (from equation 2-29) to

$$\hat{\zeta}_p = (\hat{s}_p - \hat{\gamma}_p).[Z_{pp}/(Z_{pp}+E_{pp})] \qquad (5\text{-}1)$$

where Z_{pp} is the expected variance of ζ_p, and E_{pp} are the error variances of $(\hat{s}-\hat{\gamma})$, dominated by geoid errors. The expected geoid errors used for this computation are those depicted in figure 4-15. The altimetric data were assigned 25 cm error, based on Rapp's (1982) analysis. If the above assumptions gave an accurate description of ζ, then the expected error of the estimate computed with equation 5-1 would be :

$$\sigma^2_p = Z_{pp} \{ 1 - Z_{pp}/(Z_{pp}+E_{pp}) \} = Z_{pp} (1-w_p) \qquad (5\text{-}2)$$

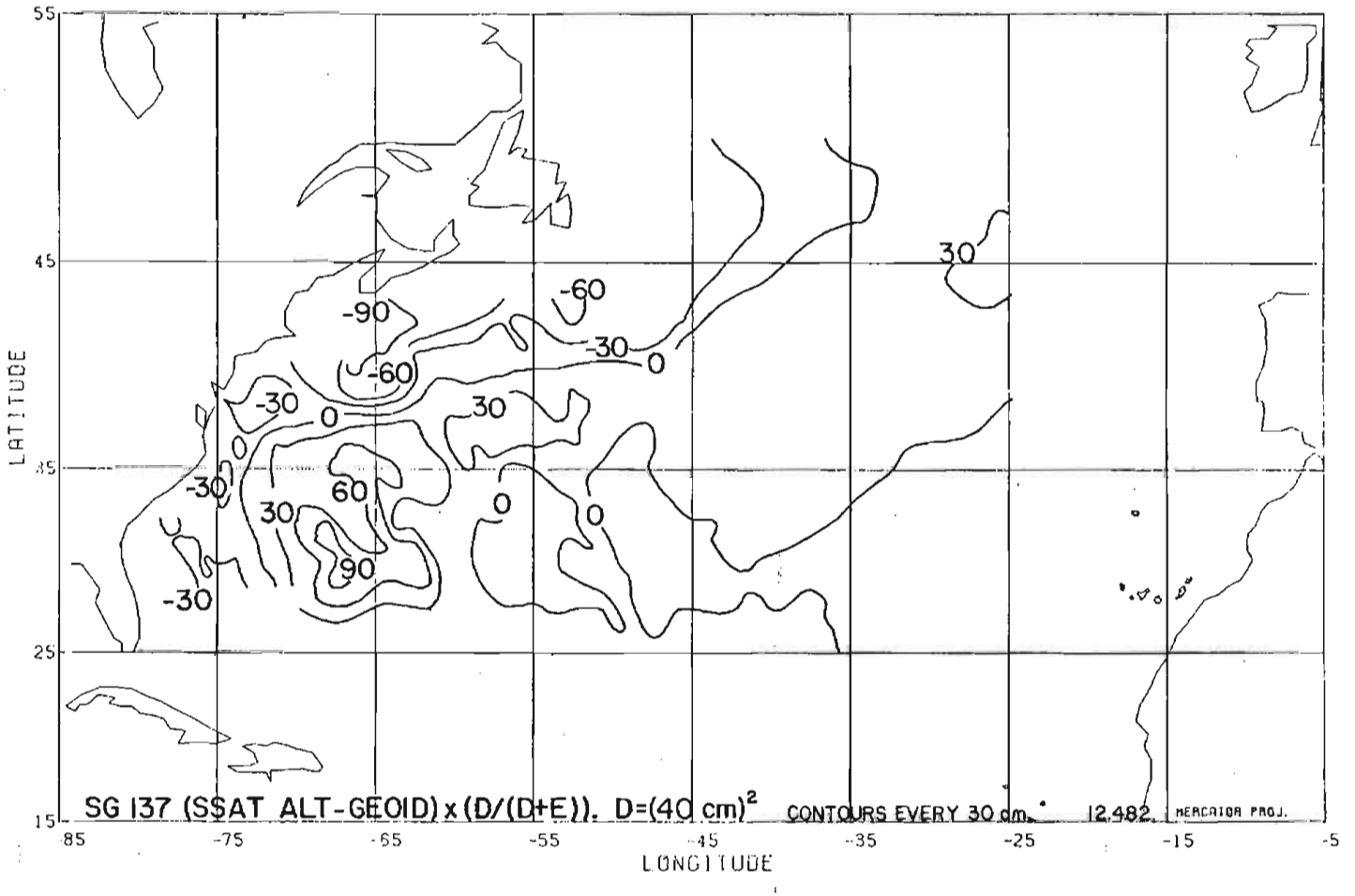

Figure 5-1: Difference (s-γ) weighted by D/(D+E) (equation 5-1), where D=(40 cm)^2 and E are the squared geoid errors of Figure 4-15. This figure is a better estimate of ζ than the unweighted difference (s-γ). See text for discussion. Contours every 30 cm.

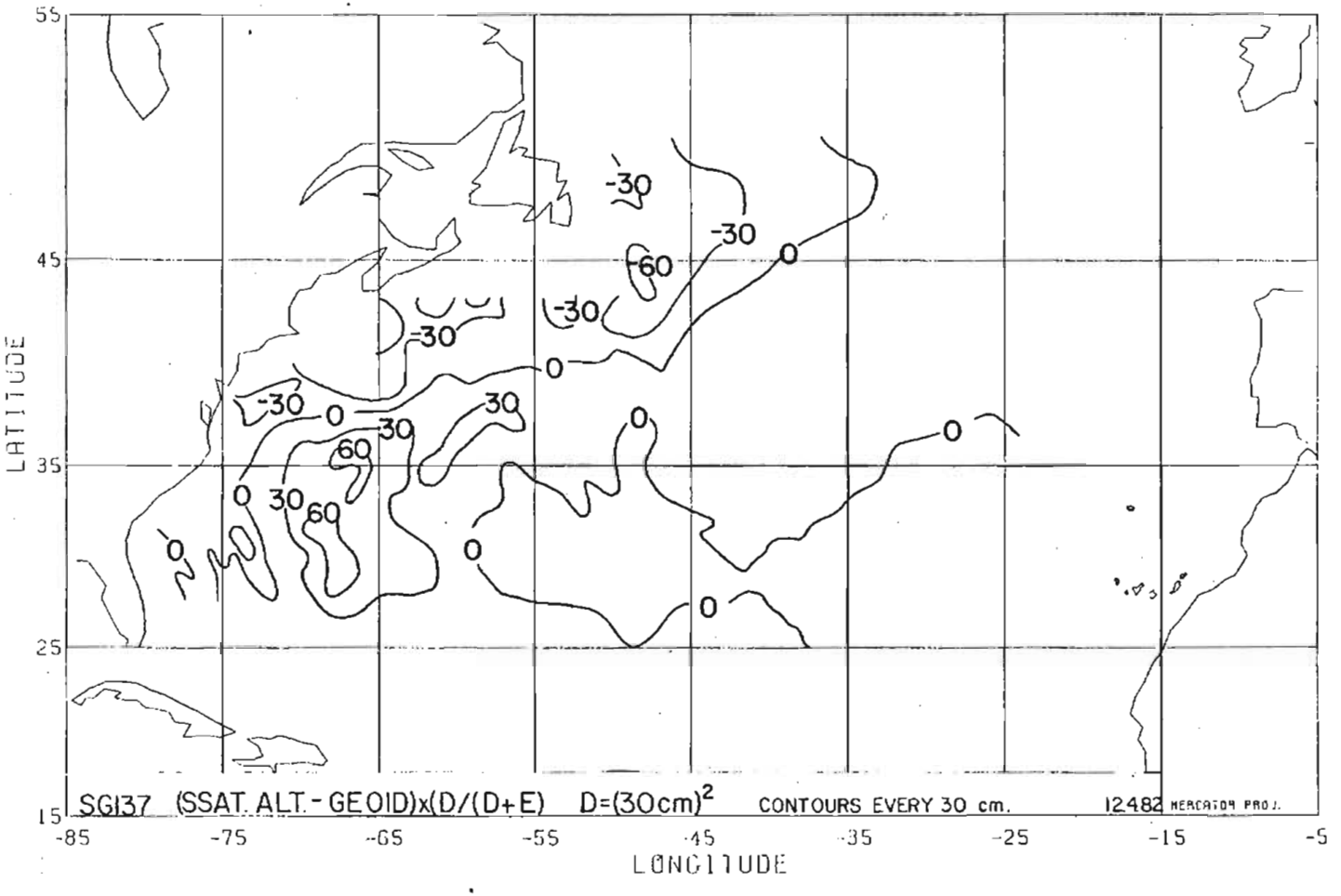

Figure 5-2: Difference (s-γ) weighted by D/(D+E) (equation 5-1), where D=(30 cm)^2 and E are the squared geoid errors of Figure 4-15. This figure is a better estimate of ζ than the unweighted difference (s-γ). See text for discussion. Contours every 30 cm.

(which is the appropriate simplification of equation 2-17).

The first desirable property of any computation is that the result not be overly sensitive to uncertain parameters. Figures 5-1 and 5-2 show the estimate of ζ when its rms value Z is assumed to be 40 or 30 cm respectively (the rms of the hydrographic estimate of ζ depicted in figure 5-4 is 33 cm). Both figures show the same features; as expected, assuming more power in ζ (figure 5-1) shows more detail -most likely noise.

The most optimistic interpretation of a figure such as 5-2 is that it represents a statistically optimum estimate of the difference between measured sea-surface topography and computed geoid. But both the structure of ζ and the stucture of data noise were strongly simplified, hence figure 5-2 is perhaps better described as the result of a scaling scheme that automatically takes into account the variation in geoid errors -by an order of magnitude- over the North Atlantic.

The scaling factors w_p used to compute figure 5-2 are shown in figure 5-3 in the form $1-w_p$. The 'optimistic' interpretation of this figure is that it represents the actual expected errors of ζ, but the discussion in the previous paragraph also applies here.

Figure 5-2 shows a believable gyre and a believable time-averaged Gulf Stream, but only north of Cape Hatteras (35^{o}N). The position and width of the 'Gulf Stream' in figure 5-2 cannot be directly compared to a quasi-instantaneous picture of the circulation, such as computed by Wunsch (1981) and reproduced here in figure 5-4, because of the shifting

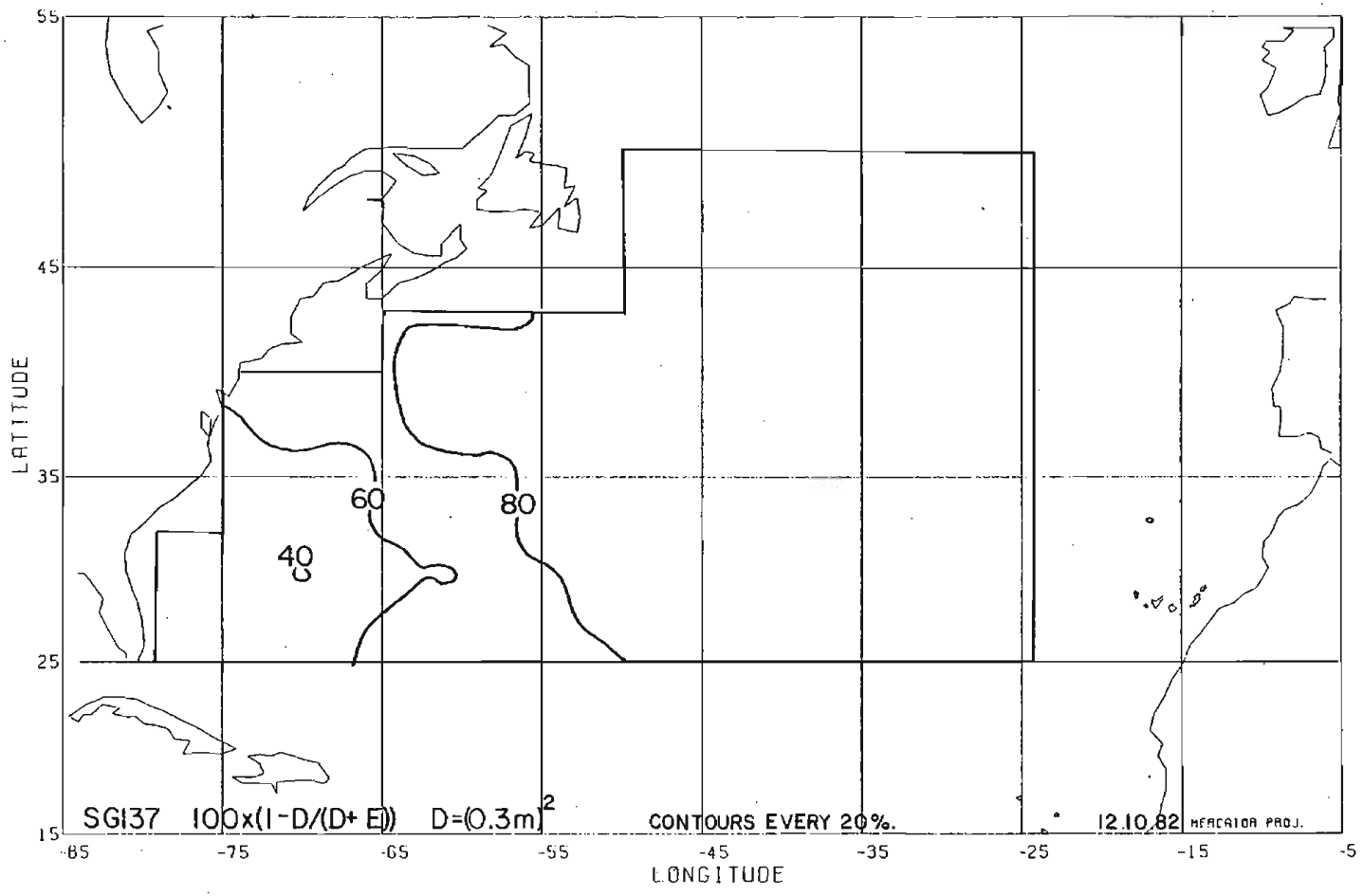

Figure 5-3: 100x(1-D/(D+E)). Expected errors of the estimate of ζ given in Figure 5-2, as percentage of the variance of ζ (equation 5-2). Contours every 20%. Notice that the complement to 100 of these numbers indicates the down-weighting applied to (s-Y) in Figure 5-2.

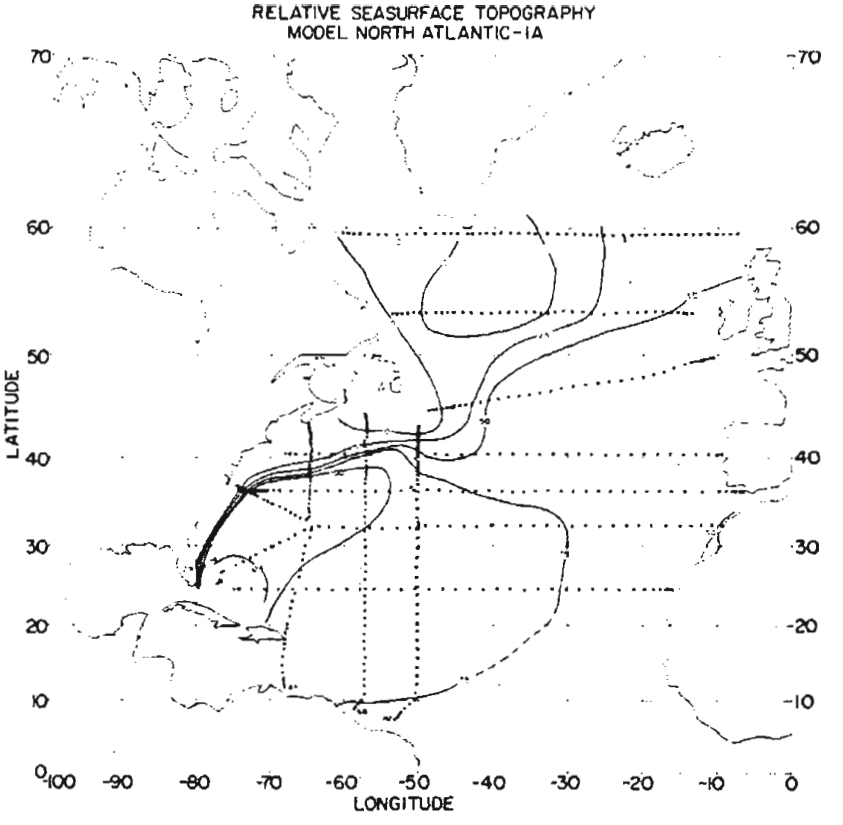

FIGURE 5-4. Surface elevation $\tilde{\zeta}$, in cm , computed by Wunsch (1981) solely from hydrographic data. Because each cruise provides an almost instantaneous profile of the circulation, but different cruises measured at different times of the year, this picture is strongly time-aliased. Wunsch (1981) described it as 'a sea surface as it might appear at one particular instant, but as it may never have actually been.'. The geostrophic relations only define the slopes, hence Bermuda was arbitrarily set at 100 cm..

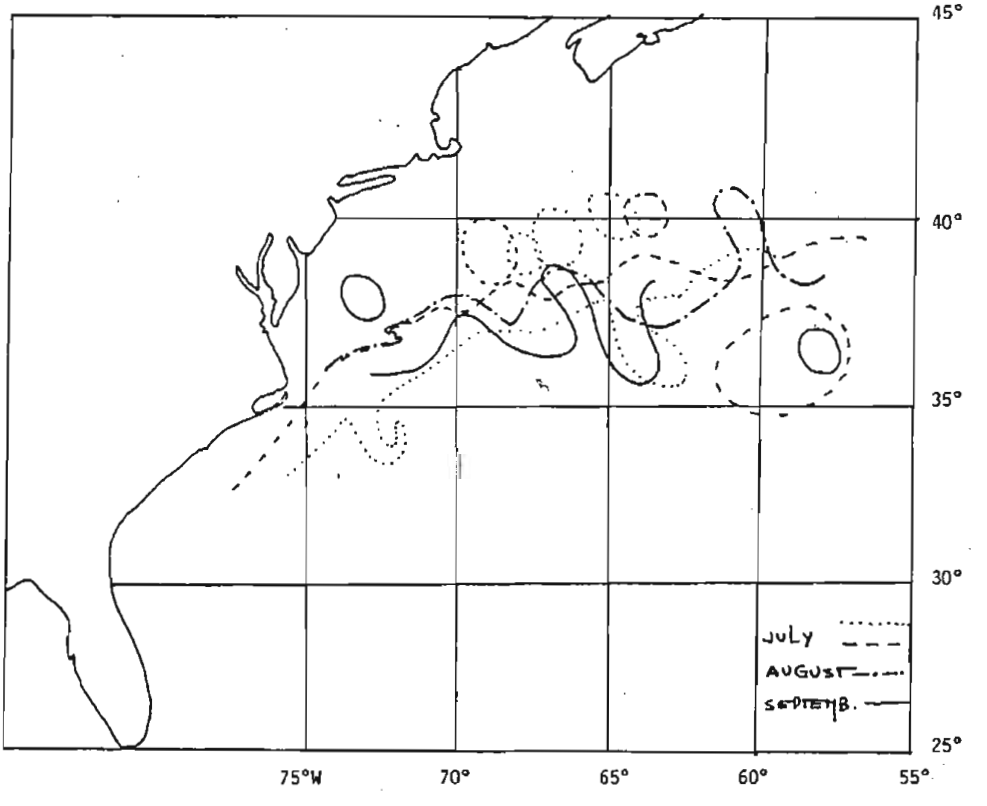

FIGURE 5-5. POsitions of the Gulf Stream and larger eddies during the SEASAT mission, as measured by infrared radiometry. From: GULFSTREAM, vol. IV, numbers 7, 8,9, july-september 1978. NOAA-NWS.

position of the current, and the passage through the area of large mesoscale eddies during the 3 months over which we are averaging (this point was discussed at length in section 1.3, chapter 1). These motions can be inferred from remote sensing of the temperature structure in the upper ocean, using satellite infrarred radiometry. Figure 5-5 shows the position of the boundaries of the Gulf Stream as determined by the position of "prominent sea-surface temperature gradients, or by the 15° Ç isotherm at 200 m", as published in the Gulf Stream bulletin (june-september, 1978); the agreement with the corresponding features of figure 5-2 is very good. There is also a hint of a northeastward flow towards Iceland (see Stommel et al., 1978; Wunsch 1981) in figure 5-2, but it is defined as the boundary of two obviously erroneous features in figure 4-14. The formal error estimates of figure 5-3 confirm that this feature cannot be taken too seriously from this data set.

The most disappointing feature of figures 5-2 and 5-3 is the absence of the powerful signal associated with the Florida current (the component of the circulation off of Florida in figure 5-4) and the failure of the error estimates to predict the large discrepancy in this region. The error estimates are rms worldwide averages, and as such can be expected to fail over a small fraction of the Earth's surface, but in this case the geophysical cause is apparent. Islands such as the Bahamas, Cuba and Puerto Rico, have strong positive signals in an otherwise negative gravity background. Unfortunately data over the islands are usually very sparse -and old- and

this is the case with the Bahamas, at the tip of Florida. Because of undersampling over the islands, the averages are biased towards negative values; because these are regions of steep gravity gradients, a worldwide average covariance function (power spectrum) tends to underestimate the energy being aliased (an idea of the difference in power between trenches and 'average' ocean floor can be found in Brammer and Sailor, 1983). In addition, the Wagner and Colombo spectrum underestimates power at short wavelength.

5.3 USING AN INITIAL ESTIMATE OF ζ

We now replace our simplistic description of ζ as a spatially uncorrelated process. The new description states that ζ should resemble the hydrographic estimate (figure 5-4) within certain expected errors. We assume that these errors are spatially uncorrelated. This assumption about the errors yields again an optimum estimation equation at each point, uncorrelated from data at other points.

$$\hat{\zeta} = \alpha\,(\hat{s} - \hat{m}) + \beta\,\tilde{\zeta} \qquad (5\text{-}3)$$

$$\alpha = \sigma_2^2/(\sigma_1^2+\sigma_2^2), \quad \beta = \sigma_1^2/(\sigma_1^2+\sigma_2^2)$$

σ_1^2 = error variance of $\hat{s}-\hat{m}$

σ_2^2 = error variance of the hydrographic estimate $\tilde{\zeta}$.

The key quantitities are now the σ_2 -the errors in $\tilde{\zeta}$-. Wunsch (1981) and Roemmich and Wunsch (1981) argued that the expected error of a hydrographic estimate of ζ -obtained using

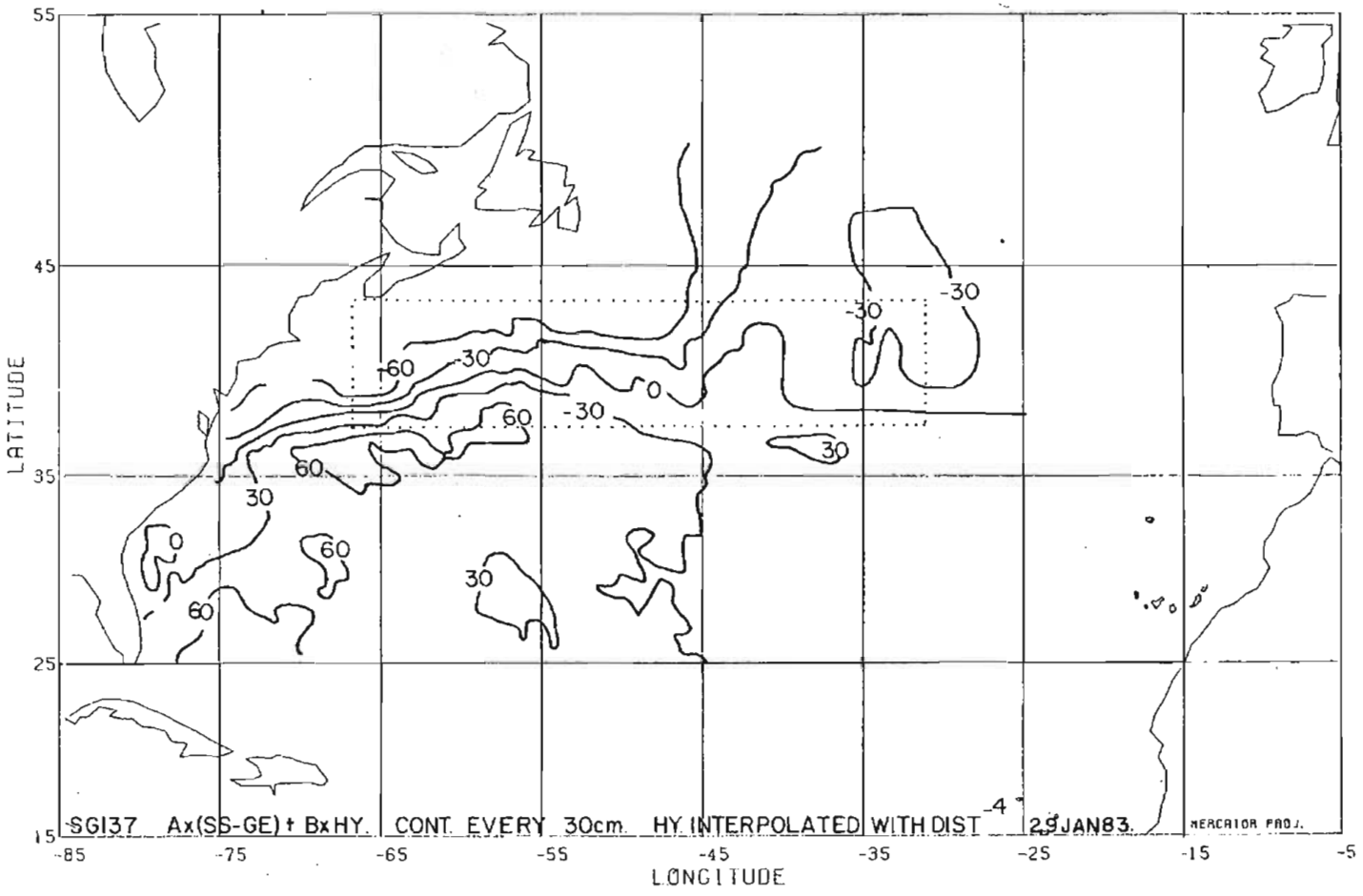

FIGURE 5-6: Estimate of ζ given by equation 5-3. It represents those parts of $s-\gamma$ and ζ that are common to boith surfaces, within their respected errors. The hydrographic estimate, ζ, was obtained by Wunsch (1981). Its mean (70 cm) was removed prior to computations. Its errors were set at 10 cm outside the dotted box (37°-43°N; 75°-40°W), and 30 cm inside. Geoid errors are those of figure 4-15; errors in s were neglected. The standard deviation of this surface is 38 cm; that of the hydrographic estimate is 33 cm.

inverse methods-is about 10 cm away from western boundary currents. This estimate is based on the changes in the computed $\tilde{\zeta}$ when different initial reference levels are used, and on the envelope of null space solutions that can be added up to an expected variance for the current velocity (this description of the resolution error is entirely equivalent to equation 2-17). Within 100 km of a boundary current, the expected error should be about 50 cm -to allow for known time changes in the current axis-. Between 100 and 200 km from the axis, a 25 cm error is likely, again according to Wunsch (1981). For the computations that led to figure 5-6, σ_2 was set at 10 cm everywhere outside a box that contains the Gulf Stream north of Cape Hatteras; in this box, σ_2=30 cm. Off of Florida, σ_2 was left at the 10 cm level to offset -partially- the large underestimate of geoid error in this area.

The hydrographic estimate of Wunsch (1981) has a mean of 70 cm, and a standard deviation of 33 cm. The mean is a consequence of arbitrarily setting Bermuda at 100 cm -the geostrophic relation does not define this quantity-hence it was removed prior to combining the hydrographic surface with the other two. Figure 5-6 is significantly better than the crude estimate in fig. 5-3 only in the neighbourhood of Florida; of course, 'significantly better' still means closer to the hydrographic estimate. Part of the geoid error in the neighbourhood of the Grand Banks is apparent in figure 5-6.

Removing the surface of figure 5-6 from $\hat{s}$ yields the best estimate of the geoid γ that can be obtained with this data set

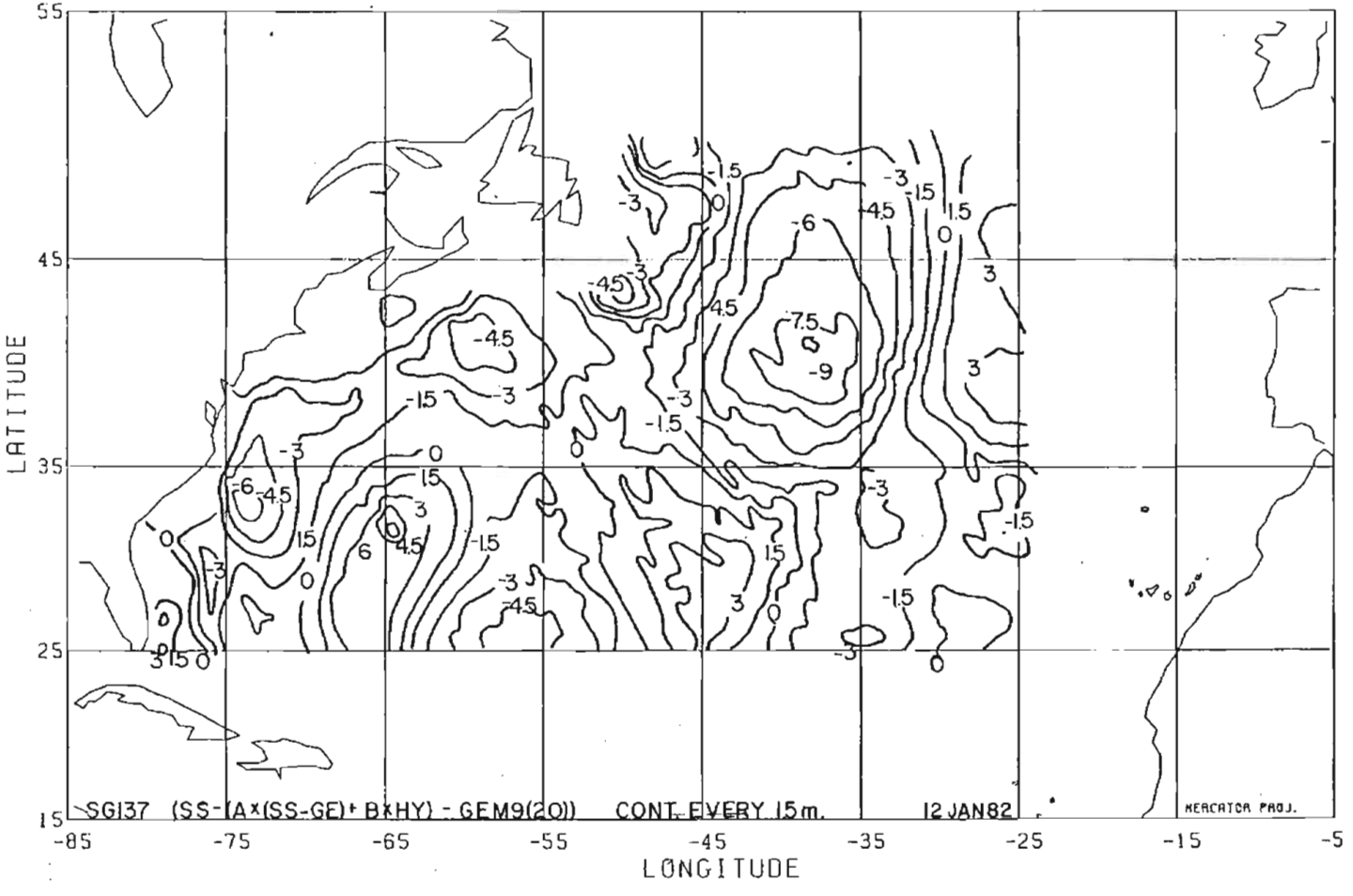

FIGURE 5-7: estimate of the geoid based on altimetry, gravity and hydrography, relative to GEM-9 (up to degree 20). It is the difference s-z-GEM9, where z is that of figure 5-6. Contours every 1.5 m.

and the simplifications introduced. Figure 5-7 shows such an estimate of γ based on altimetry, hydrography and gravity. It is dominated by the altimetric data -compare to figure 4-14- but the differences due to figure 5-6 can be detected, particularly near Florida, and around 38°N.

5.4 SUMMARY

The unscaled difference $\hat{s}-\hat{\gamma}$ does not measure the surface expression of the general circulation, ζ, because the errors in γ both dominate the difference surface, and vary by an order of magnitude over the North Atlantic. Only when the expected errors in $\hat{s}-\hat{\gamma}$ are combined with an estimate of ζ, even a very rough one, do the known features of the circulation begin to appear above the noise background.

Even the simplest of assumptions -that ζ is spatially uncorrelated over distances greater than 100 km, with constant $(30\ cm)^2$ variance- produces a believable, but blurred, picture of the main gyre in the North Atlantic. The failure to define the Florida current is intrinsic to the gravity data set, and no amount of unprejudiced optimization (i.e., short of using the hydrographic estimate itself) can recover it. Even using expected geoid errors that underestimate actual errors, as these do, is better than no scaling at all. The reason for this partial success is that the spatial variability of expected geoid errors follows closely the actual discrepancies

whose main source is undersampling of short wavelengths in the gravity field..

No new feature of ζ can be assured from these results, and disturbingly enough, some well known and distinctive features -such as the Florida current- fail to appear at all. Only additional information about the gravity field can recover these features. However, precisely because such features are large and fairly well constrained from hydrographic data, a significant correction to an estimate of γ from s can be applied.

CHAPTER 6

SUMMARY, DISCUSSION AND CONCLUSIONS

The review of least squares inverse methods used in geophysics (chapter 2) showed that the diverse criteria of optimality that use the L_2 norm produce only one form of optimum inverse. The difference in practice lies in the choice of 1) the weighting function used to describe the smoothness of the unknown; 2) the signal-to-noise parameter μ; 3) a desire to obtain an unbiased inverse. It was argued that the choices of the Gauss-Markov theorem are best suited when there is an overabundance of data -and they cannot be inverted simultaneously or stepwise-. This approach was then used in chapters 4 and 5.

Our analysis of the optimum construction of geoids from gravity data (chapter 3) emphasized resolution functions rather than rms errors, and explored a variety of possible solutions, rather than a single one. The main finding of that chapter was that accurate bandpassed versions of the geoid could be constructed from fairly limited data sets, but only if the data themselves were accurate averages over small areas.

Altimetric measurements of sea-surface topography ($\hat{s}$), such as those obtained with Seasat, have expected errors ranging between 10 and 30 cm after crossover adjustments,

but the wavenumber spectrum of this error, particularly its accuracy at long wavelengths is still unknown. The altimetric accuracy is fairly uniform throughout the North Atlantic. In constrast, estimates of the North Atlantic geoid ($\hat{\gamma}$) in the wavelength band 2000 to 100 km, have errors ranging between 30 and 260 cm, when computed from surface gravity acceleration data and satellite orbit perturbations. The second major component of the time-averaged s, the surface expression ζ of the general circulation, has an rms value around 30 cm. Obvious consequences follow from these values: 1) $\hat{s}$ gives more information about γ than gravity acceleration data do, even before removing $\hat{\zeta}$; 2) time-varying oceanographic components are easier to recover from altimetry, because they do not require an independent estimate of γ; 3) given $\hat{s}$ and $\hat{\gamma}$, one can only recover ζ by considering the spatial variation in expected geoid errors, and its relation to expected ζ. Points 1 and 2 were well known before this thesis, but point 3 was never implemented (except for the long wavelength component, by Tai (1982)) because the best marine geoids lack any useful accuracy statements.

Publicly available marine gravity data cannot be converted into geoidal heights with the same or higher accuracy than that of the altimeter because of aliasing -gravity power at wavelengths between 5 km and a few hundred km acts as noise when the sampling distance is longer than its wavelength, and research vessels -slow and expensive-have not

been able to cover the ocean with a pattern dense enough to define these wavelengths. Not surprisingly, one can only compute an accurate gravimetric geoid in the immediate vicinity of the continental U.S., from where a majority of ship tracks leave, and where they cross in a dense pattern. The geoid computation of chapter 4 was aimed both at obtaining a good geoid from the available data, and also at providing realistic expected errors for this geoid, later needed in order to combine $\hat{\gamma}$ with $\hat{s}$.

The nonuniform geoid accuracy precludes recovering ζ by directly substracting $\hat{\gamma}$ from $\hat{s}$. A much better result is obtained through a scaling scheme, equivalent to the optimum estimation of geostrophic heights ζ when both ζ and the geoid errors are assumed to be spatially uncorrelated. The computational procedure that follows from such a simple description of ζ, applied in chapter 5, yielded a believable estimate of the main gyre in the time-averaged North Atlantic circulation.

A hydrographic estimate of ζ can be combined with $\hat{\gamma}$ and $\hat{s}$, but only if the modelling errors in $\hat{\zeta}$ are reasonably well described. The main source of modelling error is the neglect of time variations in the geostrophic equations. Such a computation was performed in chapter 5; the improvements over the estimate described in the previous paragraph are most obvious off of Florida, where geoid errors completely mask the well known Florida current. This feature of

the circulation is an extreme example of how knowledge of ζ can be used to correct an altimetric estimate of γ.

It is disturbing to find that the success of the computations required to recover ζ from $\hat{s}$ depends mostly on the estimates of expected errors in $\hat{\gamma}$ and $\hat{\zeta}$. Disturbing, but not surprising because the size of ζ is at or below the noise level in $\hat{\gamma}$.

Of the many simplifications introduced in the estimates of chapters 4 and 5, disregarding the spatial correlation in geoid errors -and not knowing those of the altimetric surface- is probably the most critical: just as it is possible to distinguish the sea surface signal of a seamount from that produced by a mesoscale eddy on the basis of their behaviour in time, it is possible to distinguish geoid errors from geostrophic features on the basis of their different wavenumber behaviour. Furthermore, a new generation of computers is now available, and many of the lengthy computations avoided in this work can now be carried out, but only small improvements in the signal-to-noise can be expected.

The fact that the largest source of error in the geoidal estimates is due not to measurement noise but to missing data, suggests an alternative approach to the problem of recovering oceanographic information from altimetry. Suppose we convert $\hat{s}$ into gravity accelerations,

a computation that is much less affected by missing data because of the excellent coverage of the satellite. We would now substract estimated gravity from measured gravity only at those locations where gravity accelerations were actually measured. At this stage we would be left with profiles of the gravity acceleration equivalent of ζ. A final transformation into slopes of sea-surface (deflections of the vertical) would yield the desired quantities. The advantage of this approach lies in its ability to minimize the aliasing error introduced every time data are gridded or an integral transform is applied.

"A lack of information cannot be remedied by any mathematical trickery" (Lanczos, 1961, chapter 3). The only way to improve significantly an estimate of the time-averaged circulation from sea surface measurements, is to include independent information about the gravity field. Among already existing data, bathymetry is probably the only source that can be used. It has been observed for many years that gravity and the topography of the solid earth are strongly correlated, but the degree of correlation changes both as a function of wavelength and with tectonic setting. During the last ten years, physical models describing loading over elastic plates of different ages have successfully explained the main features of this correlation over the oceans. The available bathymetric data set, when properly used, becomes a source of gravity information at

those wavelengths where the correlation is strong. But this correlation must be used with care, because a rotating fluid like the ocean reacts strongly to bottom topography, a simple consequence of the conservation of potential vorticity; the circulation itself tends to follow contour lines away from the equator, and seamounts can be expected to generate swirls in the flow above them.

Other than bathymetry, satellite-to-satellite tracking and satellite gradiometry are the most promising among future sources of data; surface gravity measurements can be expected to contribute only short wavelength information because of their high resolution and poor coverage. Until such new information about the Earth's gravity field is available -and is not derived from altimetry- recovery of time-averaged features of the circulation will remain restricted to either small portions of the ocean, or to the longest wavelength components.

ACKNOWLEDGEMENTS

Most theses reflect, to some extent, the ideas of the thesis advisors, and this work is not an exception. But my gratitude to Carl Wunsch and Barry Parsons has to do more with example than with specific suggestions. I would be very happy if, given a chance to supervise, I were to combine guidance and freedom, willingness to discuss any technical obstacle at length, physical intuition, or the will to correct a foreign student's prose, as Carl Wunsch has. Or pay as much attention to small details and physical meaning, and pursue scientific subjects as stubbornly as Barry Parsons has.

John Sclater's "spectacular" and contagious enthusiasm for marine geophysics guided me in my first two years in the Joint Program. I hope I will never lose it. I am also happy that Charlie Hollister demanded "familiarity with the planet Earth", or else ..., with its intended effect.

When I got into computer trouble, Barbara Grant was always there to pull me out. Whenever I found an administrative obstacle at MIT, Debby Gillett Roecker showed me how to dodge. When admitted to the Joint Program, Jake Peirson accompanied his official acceptance note with 'buen viaje', in my own language. Bud Brown taught me how to handle a propane torch and various power tools without losing fingers; he also drew many of the figures in chapter 4, and never charged for

his services as consulting car mechanic. When the time to type this thesis came, Dorothy Frank did most, and taught me how to do the rest.

David Johnson not only tried hard to inculcate the details of marine sediments into an avowed geophysicist, but he and Diann acquainted this foreigner (international student in doublespeak) with the American way of life -Cape Cod version.

Professor Sclater warned me early that I would learn more from my peers than from his. Almost. My fellow student Tom Herring introduced me to collocation and to Molodenskii's work; John Crowe spent days and nights teaching me how to measure heat flow at sea; Ken Green's advice - based on a decade at MIT - helped me in my first years; Bruce Cornuelle's understanding of -and faith in-statistically optimum estimation contributed much to Chapter 2 of this work. To them, and to all G & G students in the Joint Program I am grateful.

Carl Bowin and Jim Cochran kindly provided me with gravity data from cruises I could not get through NGSDC.

Nestor Lanfredi and Daniel Vara introduced me to marine research (at the Naval Hydrographic Service, Argentina). Antonio Saralegui, Angel Cerrato, Oscar Mingo and Abel Burna introduced me to Surveying, Geodesy and Geophysics at the Universidad de Buenos Aires. I thank them for their example in the midst of obstacles, and for their encouragement.

NASA's research Grant NAG6-9 not only funded this work- it also paid for a major part of my graduate education. For both reasons I am very grateful.

My English prose has many 'rough edges'. Diana Granat contributed to this thesis by rounding them off. But I owe her much, much more.

This thesis is dedicated to Celina Fingerman de Zlotnicki and Salomon Zlotnicki, my parents. What I learned from them cannot be summarized in any number of lines.

Aki, K., and P. G. Richards: Quantitative Seismology, Theory and Methods, W. H. Freeman, San Francisco, 1980.

Apel, J.R.: Satellite sensing of ocean surface dynamics. Ann. Rev. Earth Planet. Sci., 8, 303-342. 1980.

Backus, G. E., and F. Gilbert: The resolving power of gross Earth data. Geophys. J. R. Astron. Soc., 16, 169-205, 1968.

Backus, G. E., and F. Gilbert: Uniqueness in the inversion of inaccurate gross Earth data, Phil. Trans. R. Soc. London Ser. A, 266, 123-192, 1970.

Barrick, D.E. and C.T. Swift: The Seasat microwave instruments in historical perspective. IEEE Jour. Oceanic Meas., OE-S, 74-79, 1980.

Bendat, J.S. and A.G. Piersol: Random Data: analysis and measurement procedures. 407 pp., Wiley-Interscience. 1971.

Bjerhammar, A.: Theory of errors and generalized matrix inverses, Elsevier, New York, 1973.

Bomford, G.: Geodesy (4^{th} edition). Clarendon Press, Oxford. 855 pp. 1980.

Bowin, C.: Caribbbean gravity and tectonics. Special paper 169. Geolog. Soc. of America, 1976.

Bowin, C., W. Warsi, and J. Milligan: Free air gravity anomaly Atlas of the World. Geol. Soc. Amer., Map and Chart Series No. MC-46. 1982.

Bracewell, R.: The Fourier transform and its applications. McGraw Hill, 1965.

Brammer R. F.: Estimation of the ocean geoid near the Blake Escarpment using Geos-3 satellite altimetry data. J. Geophys. Res. 84, 3843-3852. 1979.

Brammer, R.F. and R.V. Sailor: Spectrum analysis of the ocean geoid using Seasat Altimeter Data and shipboard gravity survey data. J. Geophys. Res., in press, 1983.

Bretherton, F., R.E. Davis, and C.B. Fundry: A technique for objective analysis and design of experiments applied to MODE-73. Deep Sea Res., 23, 559-582. 1976.

Chapman, M.E.: Techniques for interpretation of geoid anomalies. J. Geophys. Res. 84, 3793-3802. 1979.

Chapman, M.E. and M. Talwani: Comparison of gravimetric geoids with Geos3 altimetric geoid. J. Geophys. Res. 84, 3803-3816. 1979.

Charney, J.G. and G.R. Flierl: Oceanic analogues of large-scale atmospheric moptions. In Evolution of Physical Oceanography, B.A. Warren and C. Wunsch (ed.). The MIT Press, Cambridge, Mass. 623pp. 1981.

Cheney, R.E. and J.G. Marsh: Oceanic eddy variability measured by GEOS-3 altimeter crossover differences. EOS, vol. 62, 743-752, 1981.

Chovitz, B.H.: Modern geodetic Earth reference models. EOS 62, 65-67 1981.

Christodoulidis, D.C.: Influence of the atmospheric masses on the gravitational field of the Earth. Bull. Geod., 53, 61-77, 1979.

Claerbout J.F. and F. Muir: Robust modelling with erratic data. Geophysics, 38, 826-844. 1973.

Colombo, O.L.: Optimal estimation from data regularly sampled on a sphere with applications to Geodesy. OSU-DGS† Report 291. 1979.

Davis, P.J.: Interpolation and approximation. 392 pp. Dover. 1975.

Davis, R.E.: Estimating velocity from hydrographic data. J. Geophys. Res. 83, 5507-5509. 1978.

Dongarra, J.J., C.B. Moler, J.R. Bunch, and G.W. Stewart: LINPACK User's guide. Soc. Ind. and Applied Math. 1979.

Freeden, W.: On approximation by harmonic splines. Manuscripta geodaetica, 6, 193-244. 1981.

Gihman I. and A. V. Skorohod: The theory of stochastic processes, vol. I. Springer Verlag, 570 ,pp. 1974.

Gradshteyn, I. S., and I. M. Ryzhik: Table of Integrals, Series and Products. Academic Press, New York, 1965.

Hagiwara, Y.: A new formula for evaluating the truncation error coefficients. Bull. Geodesique, 50, 131-135. 1976.

Hancock, D.W., R.G. Forsythe and J. Lorell: Seasat altimeter sensor file algorithms. IEEE Jour. Oceanic Meas., OE-S, 93-99, 1980.

Heiskanen, W. A., and H. Moritz: Physical Geodesy. W. H. Freeman, San Francisco, 1967.

Henderschott, M.C.: Long waves and ocean Tides. In Evolution of Physical Oceanography, B.A. Warren and C. Wunsch (ed.). The MIT Press, Cambridge, Mass. 623 pp. 1981.

Jackson, J. D.: Classical Electrodynamics, John Wiley, New York, 1975.

Jekeli, C.: Global accuracy estimates of point and mean undulation differences obtained from gravity disturbances, gravity anomalies and potential coefficients, †OSU-DGS Report 288, 1979.

Jekeli, C.: Modifying Stokes' function to reduce the error of geoid unudulation computations, J. Geophys. Res., 86, 6985-6990, 1981.

Kaula, W. M.: Theory of Satellite Geodesy. Blaisdell, Waltham, Mass., 1966.

Kennett, B., and G. Nolet: Resolution Analysis for discrete systems, Geophys. J. R. Astron. Soc., 53, 413-425, 1978.

Kinsman, B.: Wind waves. Their generation and propagation on the ocean surface. Prentice Hall Inc., 1965.

Lame, D.B. and G.H. Born: Seasat measurement system evaluation: achievements and limitations. J. Geophys. Res. 87, 3173, 1982. (this issue of JGR is wholly devoted to Seasat).

Lanczos, C.: Linear Differential Operators. Van Nostrand, New York, 1961.

Lawson, C.L., and R.J. Hanson: Solving least squares problems. 340 pp. Prentice Hall, Inc. 1974.

Lerch, F.T., S.M. Klosko, R.E. Laubscher, C.A. Wagner: Gravity model improvement using GEOS-3. (GEM 9 and GEM 10) GSFC Document X-921-77-246, 1977. (also: J. Geophys. Res. 84, 3897-3916. 1979).

Lerch, F.J., S.M. Klosko and G.B. Patel: A refined gravity model from Lageos (GEM-L2). Geophys. Res. Letters 9, 1263, 1982.

Liebelt, P. B.: An Introduction to Optimal Estimation. Addison Wesley, Reading, Mass., 1967.

Marsh, J. G., and E.S. Chang: 5' detailed gravimetric geoid in the northwestern Atlantic ocean. Mar. Geod., 1, 253-261, 1978.

Marsh, J.G., B.D. Marsh, R.G. Williamson and W.T. Wells: The gravity field in the central Pacific ocean from satellite to satellite tracking. J. Geophys. Res. 86, 3979-3997, 1981.

Marsh, J.G., R.E. Cheney, T.V. Martin, J.J. McCarthy: Computation of a precise mean sea surface in the Eastern North Pacific using SEASAT altimetry EOS, Vol. 63, 178-179, 1982.

Marquardt, D.W.: Generalized inverses, ridge regression, biased linear estimation, and nonlinear estimation. Technometrics, 12, 591-612, 1970.

Mather, R.S.: The Earth's gravity field and ocean dynamics. NASA TM 79540, 1980.

McAdoo, D.C.: Geoid anomalies in the vicinity of subduction zones. J. Geophys. Res. 86, 6073-6090, 1981.

Molodenskii, M. S., V. F. Eremeev, and M. I. Yurkina: Methods for the Study of the External Gravitational Field and Figure of the Earth. Israel Program for Scientific Translations, Jerusalem, 1962.

Moritz, H.: Advanced Physical Geodesy. 500 pp. Herbert Wichmann Verlag, Karlsruhe (Germany), 1980.

Moritz, H.: Integral formulas and collocation. †OSU-DGS Report 234, 1975.

Moritz, H.: Covariance functions in Least Squares Collocation, †OSU-DGS Report 240, 1978.

Moritz, H.: Least Squares Collocation, Rev. Geophys. Space Phys., 16, 421-430, 1978.

Morner, N.-A.: Geoid Changes with time. (Abstract) EOS, Trans. Am. Geophys. Union, 63, 1280, 1982.

Neumann, G. and W.J. Pierson: Principles of Physical Oceanography. Prentice-Hall, 545 pp., 1966.

Olea, R.A.: Optimal contour mapping using universal kriging. J. Geophys. Res., 79, 695-702, 1974.

Papoulis, A.: Probability, random variables and stochastic processes. McGraw Hill, 583 pp., 1965.

Parker, R.L.: Linear inference and underparameterized models. Rev. Geophys. Sp. Phys., 15, 446-456, 1977b.

Parker, R.L.: Understanding Inverse Theory, Ann. Rev. Earth Planet. Sci., 5, 32-64, 1977.

Pedlosky J. Geophysical Fluid Dynamics. Sprionger Verlag. 624 pp. 1979.

Rapp, R.H.: A global 1° × 1° anomaly field combining satellite, GEOS-3 altimeter and terrestrial anomaly data. †OSU-DGS Report 278, 1978.

Rapp, R.H.: Potential coefficient and anomaly degree variance modelling revisited. †OSU-DGS Report 293, 1979.

Rapp, R.H.: A comparison of altimeter and gravimetric geoids in the Tonga Trench and Indian Ocean areas. Bull. Geod., 54, 149-163, 1980.

Rapp, R. H.: Ellipsoidal corrections for geoid undulation computations. †OSU-DGS Report 308, 1981.

Rapp, R. H.: A global atlas of sea-surface heights based on adjusted Seasat altimeter data. †OSU-DGS Report 333, 1982a.

Rapp, R. H.: A summary of the results from the O.S.U. analysis of Seasat altimetry. †OSU-DGS Report 335, 1982b.

Rhines, P.B.: The dynamics of unsteady currents. In The Sea, vol. 6, John Wiley, 1977.

Richardson, P.L.: The Benjamin Franklin and Timothy Folger charts of the Gulf Stream. In Oceanography, the past. M. Sears and D. Merriman (ed.) Springer Verlag, 812 pp., 1980.

Riesz, F. and B. Sz.-Nagy: Functional analysis. New York, 1955.

Robinson, E.A.: Predictive decomposition of time series with application to seismic exploration. Ph.D. Thesis, MIT, 1954, (Also in Geophysics, 32, 418-484, 1967).

Roemmich, D. and C. Wunsch: On combining satellite altimetry with hydrographic data. J. Marine Res., 40, 605-619. 1982.

Rowlands D.: The adjustment of Seasat altimeter data on a global basis for geoid and sea surface height determinations. OSU-DGS 325. 1981.

Sabatier P. C.: On geophysical inverse problems and constraints. J. Geophys., 43, 115-137. 1977.

Tarantola A. and B. Valette: Inverse Problems = Quest for Information. J. Geophys., 50, 159-170. 1982a.

Tarantola A. and B. Valette: Generalized nonlinear inverse problems solved using the least suqares criterion. Rev. Geophys. and Sp. Physics, 20, 219-232. 1982b.

Schutz, B.E., B.D. Tapley and C. Schum: Evaluation of the Seasat altimeter time tag bias. J. Geophys. Res. 87, 3239, 1982.

Shure, L., R.L. Parker and G.E. Backus: Harmonic splines for geomagnetic modelling. Phys. Earth and Planet. Int., 28, 215-229, 1982.

Sjoberg, L.: The accuracy of geoid undulations by degree implied by mean gravity anomalies on a sphere. J. Geophys. Res., 84, 6226-6230, 1979.

Stanley, H. R.: The GEOS-3 project. J. Geophys. Res., 84, 3779, 1979, (this issue of JGR is wholly devoted to GEOS-3). Stokes, G. G.: On the variation of gravity on the surface of the Earth. Trans. Cambridge Phil. Soc., 8, 672-695, 1849.

Stommel, H. and F. Scott: The beta spiral and the determination of the absolute velocity field from hydrographic station data. Deep Sea Res., 24, 325-329, 1977.

Stommel H., P. Niiler and D. Anati: Dynamic topography and recirculation of the North Atlantic. J. Marine Res., 36, 449-468. 1978.

Tai, C.-K.: On determining the large-scale ocean circulation from satellite altimetry. Submitted to J. Geophys. Res., 1983.

Talwani, M.: Gravity. In The Sea, ideas and observations in progress in the study of the seas. Vol. 4, part 1. A.E. Maxwell (ed.) 251-298, J. Wiley, 1971.

TOPEX Science Working Group: Satellite altimetric measurements of the oceans. Jet Propulsion Laboratory publication. Pasadena, Cal., 1981.

Tscherning, C.C.: A note on the choice of norm when using collocation for the computation of approximations to the anomalous potential. Bull. Geodesique. 51, 137-147, 1977.

Townsend, W.F.: An initial assessment of the performance achieved by the SEASAT-1 radar altimeter. IEEE Jour. Oceanic Meas., OE-S, 80-92, 1980.

Van Trees, H.: Detection, estimation and modulation theory, part 1. 697 pp., J. Wiley, 1968.

Vonbun, F.O., J.G. Marsh and F.J. Lerch: Computed and observed ocean topography: a comparison. Boundary Layer Meteorology, 13, 253, 1978.

Wagner, C. A., and O. L. Colombo: Gravitational spectra from direct measurements. NASA Tech. Memo. 79603, 1978. (Also: J. Geophys. Res., 84, 4699-4712. 1979)

Wagner, C. A.: The geoid spectrum from altimetry. J. Geophys. Res., 84, 3861, 1979.

Wahba, G.: Vector splines on the sphere, with application to the estimation of vorticity and divergence from discrete, noisy data. In Multivariate Approx. Theory, Vol. 2; W. Scherupp, K. Zeller (eds.). Birkhauser Verlag, Basel, 1982 (in press).

Warren, B.A., and C. Wunsch (ed.): Evolution of physical oceanography. Scientific Surveys in honor of Henry Stommel. 623 pp. The MIT Press, Cambridge. 1981.

Wong E.: Stochastic Processes in Information and Dynamical Systems. McGraw-Hill, 308 pp. 1971.

Wunsch, C.: Bermuda sea level in relation to tides, weather and baro-clinic fluctuations. Rev. Geophys. and Sp. Phys. 10, 1-49, 1972.

Wunsch, C.: Determining the general circulation of the oceans: a preliminary discussion. Science 96, 871-875, 1977.

Wunsch, C.: The North Atlantic general circulation west of 50°W determined by inverse methods. Rev. Geophys. and Space Phys., 16, 583-620, 1978.

Wunsch, C. and E.M. Gaposhkin: On using satellite altimetry to deter-mine the general circulation of the oceans, with application to geoid improvement. Rev. Geophys. and Space Phys., 18, 725-745, 1980.

Wunsch, C.: An interim relative sea-surface for the North Atlantic. Marine Geodesy, 5, 103-119, 1981.

Wunsch, C.: Low-frequency variability of the sea. In Evolution of Physical Oceanography, B.A. Warren and C. Wunsch (ed.). The MIT Press, Cambridge, Mass., 623 pp., 1981.

Wyrtki, K.: Sea level variation: monitoring the breath of the Pacific. EOS, Trans. Am. Geophys. Union 60, 25-27, 1979.

Zlotnicki, V., B. Parsons, and C. Wunsch: The inverse problem of constructing a gravimetric geoid. J. Geophys. Res., 87, 1835-1848, 1982.

†OSU-DGS Report: Report of the Department of Geodetic Science, The Ohio State University, Columbus, Ohio 43210.

BIOGRAPHICAL NOTE

The author was born -on april 26, 1952- and raised in Buenos Aires, Argentina. He graduated from the Colegio Nacional de Buenos Aires in 1970. He then entered the Facultad de Ingenieria, Universidad de Buenos Aires, to pursue a career where he could merge his taste for physics and mathematics with his desire to work in open spaces. He graduated as Surveyor (Agrimensor) in 1974, and completed the coursework required for the degree of Engineer in Geodesy and Geophysics in 1975. On that year he started work as a research assistant at the Naval Hydrographic Service, Argentine Navy. He travelled abroad for the first time in june 1977 -to come to the U.S. and to the Joint Program in Oceanography.