DATA ASSIMILATION IN METEOROLOGY AND OCEANOGRAPHY

MICHAEL GHIL

Climate Dynamics Center
Department of Atmospheric Sciences and
Institute of Geophysics and Planetary Physics
University of California, Los Angeles, California 90024

PAOLA MALANOTTE-RIZZOLI

Center for Meteorology and Physical Oceanography
Department of Earth, Atmospheric, and Planetary Sciences
Massachusetts Institute of Technology
Cambridge, Massachusetts 02139

1. INTRODUCTION AND MOTIVATION

The atmosphere and oceans are in motion on time scales comparable to those of human experience. By contrast, the solid earth—lithosphere, mantle, and core—moves on the average much more slowly. Geophysical inference for the solid earth has concentrated, therefore, on methods that interpret the available data in terms of a stationary structure. Atmospheric data, on the other hand, are routinely interpreted in the daily process of numerical weather prediction by assimilation into nonstationary, dynamical models. In oceanography, inverse modeling of the solid-earth type, as well as data assimilation of the meteorological type, are trying to make the best of the relatively limited, but rapidly-increasing data sets.

The oceanographic data revolution knocking at the door will bring the daily practice of physical oceanography closer to that of dynamic meteorology. Two major differences however must be kept in mind when discussing data assimilation for the two geofluids. First, the oceanographic data set to become available in the mid-1990s is expected to be considerably smaller than that currently available in meteorology and will fall far short of complete, uniform, and accurate coverage of the mass and velocity fields throughout the world ocean's width and depth.

The second major difference is in the motivation. Meteorological assimilation has been driven largely, but not exclusively, by the crucial need to forecast. This need is not as pressing at the present time for most oceanographers, with two exceptions: In the tropical ocean the capability to predict on seasonal and interannual time scales is being documented, and experimental forecasts are leading to quasi-operational ones. In mid-latitudes and on the global scale, there is substantial interest in nowcasting and short-range forecasting by the world's navies, fisheries, and off-shore drilling concerns. This interest is being met by a small, but very active segment of the oceanographic community, especially on the regional and subbasin scale.

The most pressing motivation for a much larger number of physical oceanographers is, however, the optimization of the use of the much-expanded, but still insufficient data sets expected in the near future, for the purposes of deepening and broadening our understanding of ocean circulation on regional, basin, and global scales. This will require the blending of actual current observations with the theoretical knowledge from past observations, as incorporated into numerical models, prognostic or diagnostic. Data sets from field programs will be archived and thus will be available for imaginative use with different numerical models and data assimilation or inverse methods, hence, the need to intensify the exploration and intercomparison of data assimilation methods in oceanography.

Numerical models can be used to assimilate meteorological and oceanographic data, creating a dynamically consistent, complete and accurate "movie" of the two geofluids, atmosphere and ocean, in motion. One key problem for oceanographic applications is how to determine variables not directly observed, such as the velocity components, from the observed variables, such as surface height or wind stress. The other key problem is how to use information in one part of the ocean, at the surface or in a western boundary current, in order to infer the state of the other parts, at depth and throughout a subtropical gyre. The answers to these two problems lie in the dynamical coupling between variables for the one, and the propagation of information with the flow for the other. This is the central role that dynamics plays in estimating the state of the ocean, as well as that of the atmosphere, from incomplete data. Numerical models, however, are not and never will be perfectly accurate representations of the atmosphere and ocean's large-scale motions. Both models and data have errors; hence the need to balance dynamical and observational information properly.

Meteorological data usage can thus provide some guidance to oceanographers. On the other hand, differences in the intrinsic properties of the two fluids and in the nature of the available data sets imply that oceanographers must proceed cautiously in building upon the experience accumulated by the meteorological community. The existence of complex continental borders and of narrow, intense currents along western boundaries and the equator, the difficulty in defining unambiguously a mean quasi-steady circulation, the importance of deep convective processes that are confined to very limited ocean areas are all major differences from the global atmosphere. These differences and many others will oblige oceanographers to reinterpret, adapt, and modify data assimilation techniques suggested by meteorology and other disciplines, such as geophysics and control theory. The purpose of this chapter is therefore twofold: (1) to provide a review of current operational practice and of advanced data assimilation techniques in meteorology, and (2) to illustrate the difference between the oceanic and atmospheric cases, showing how the theoretical framework developed in meteorology can best be adapted and modified for oceanography.

The chapter is organized as follows. In Section 2, the history of data usage in meteorology is outlined, and a number of methods for combining data with models are mentioned. In Section 3, we focus upon the differences and similarities between the two geofluids, the atmosphere and the ocean, comparing their physics and dynamics as well as the current and expected data sets for each medium and the available numerical models. In Section 4, the mathematical framework of estimation theory is presented, emphasizing sequential estimation and its connections with variational methods. Computational considerations are raised and the implications of model nonlinearity for data assimilation are discussed.

Section 5 is devoted to meteorological applications of data assimilation, starting with the currently most widespread operational technique, so called optimal interpolation (OI). The initialization of primitive-equation (PE) models is presented, and methods are outlined to eliminate the undesirable fast gravity waves allowed by PE dynamics. Among the advanced techniques currently under development, both variational and sequential methods are reviewed. In Section 6, we discuss the rapidly growing field of oceanographic data assimilation. We assess the important questions to be solved and their dependence upon the forthcoming data sets, as well as dependence of data assimilation methods on the model used. The existing literature is reviewed, and specific examples of relevance are given. Section 7 concludes this review with a critical discussion of results achieved so far and an outline of important paths for future research.

New data sets are becoming available in many other fields of geophysical fluid dynamics, such as planetary atmospheres or the earth's stratosphere (Panel, 1991). For the uninitiated readers, who need a quick overview of available methods for assimilating new data in their field of study, we recommend the following path of reading on the first go around: Section 2, Sections 4.1 and 4.2, Section 5.1, Sections 5.4.2 and 5.4.3, Section 6.3.2, and Section 7. This might turn you off foreover from data assimilation, or it might motivate you to read other sections of the chapter. Alternatively, you might continue the learning process by perusing additional references mentioned in these key sections or by working out simple examples with the methods outlined.

2. Evolution of Data Assimilation in Meteorology

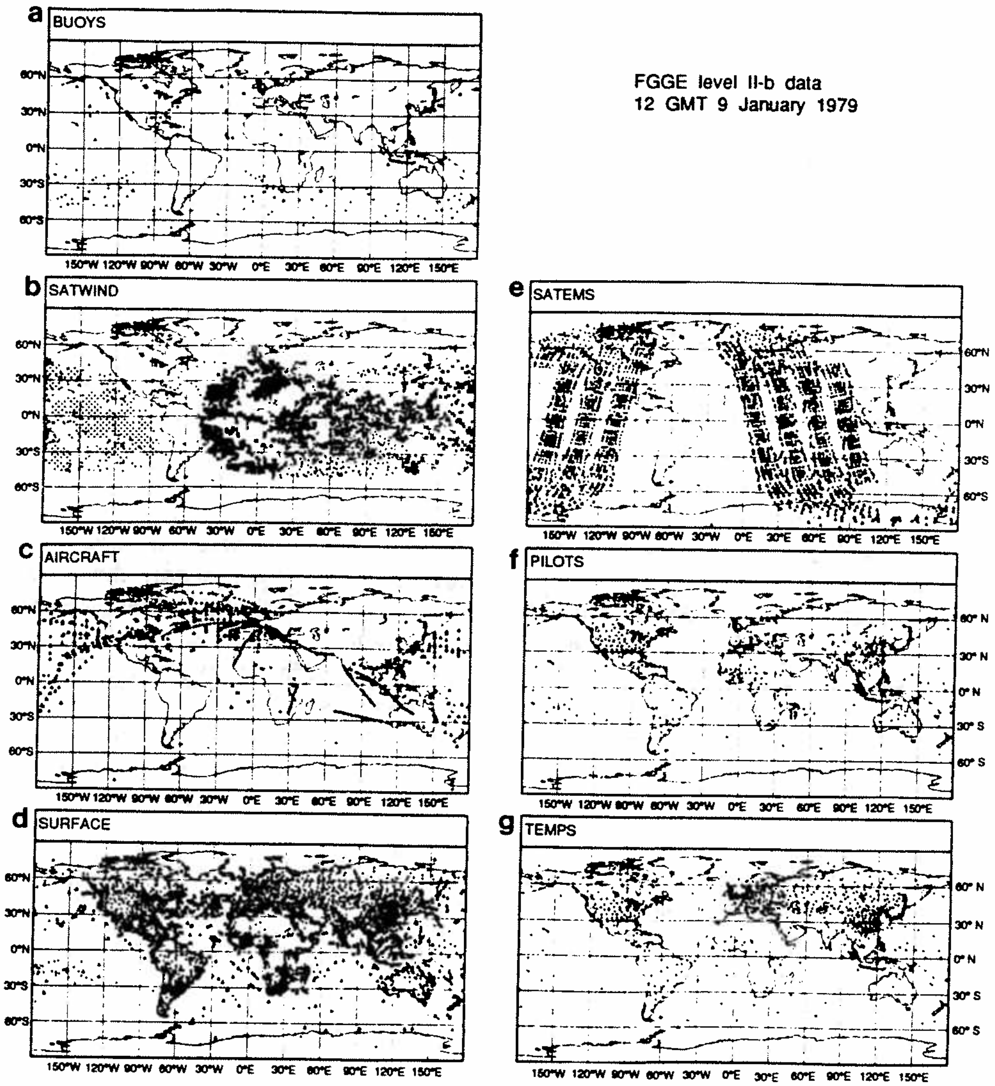
We not only want to know and understand the climatological or current state of either geofluid (the atmosphere or the ocean), we also want to predict their future state. Beyond the qualitative understanding of either geofluid, a quantitative estimate of its state in the past and present as well as quantitative prediction of future states is required. The estimate of the present state is a prerequisite for future prediction, and the accuracy of past prediction is essential for an accurate estimate of the present.

How does the estimation of the present proceed in meteorology? The first step along the road of quantitative numerical estimation in meteorology was *objective analysis*, which replaced manual graphic interpolation of observations by automated mathematical methods, such as two-dimensional (2-D) polynomial interpolation (Panofsky, 1949). Not surprisingly, this step was largely motivated by the use of rapidly improving knowledge of atmospheric dynamics to produce numerical weather forecasts (Charney *et al.*, 1950). The main ideas underlying objective analysis were statistical (Eliassen, 1954; Gandin, 1963; Phillips, 1976). Observations are considered to sample a random field with a given spatial covariance structure, which is predetermined and stationary in time.

This generalizes, in fact, Wiener's (1949, 1956) ideas on statistical estimation and prediction (cf. Ghil, 1989) from a finite-dimensional system, governed by ordinary differential equations (ODEs), to an infinite-dimensional system, governed by the partial differential equations (PDEs) of geophysical fluid dynamics. In practice, these statistical ideas appeared too complicated and computationally expensive at the time to be adopted as they stood into the fledgling numerical weather prediction (NWP) process. Instead, various shortcuts, such as the successive-correction method were implemented in the operational routine of weather bureaus (Cressman, 1959).

Two related developments led to the next step, in which the connection between statistical interpolation on the one hand and dynamics on the other became apparent and started to be used systematically. One development was the increasingly accurate nature of numerical weather forecasts; the other was the advent of time-continuous, space-borne observing systems. Together, they produced the concept of four-dimensional (4-D) space–time continuous data assimilation in which a model forecast of atmospheric fields is sequentially updated with incoming observations (Charney *et al.*, 1969; Smagorinsky *et al.*, 1970; Rutherford, 1972). Here the model carries forward in time the knowledge of a finite number of past observations, subject to the appropriate dynamics, to be blended with the latest observations.

Combining the 4-D assimilation of the new satellite, aircraft, and drifting buoy data with the usual objective analysis of the earlier conventional data



FGGE level II-b data
12 GMT 9 January 1979

FIG. 1. Meteorological observations available during one 12-hour period centered at 1200 GMT 9 January 1979. Each panel gives one type of observation with data type at the top left; numbers in parentheses here are typical of measurements available for a 12-hour period: (a) Drifting buoys, surface pressure p_s (270); (b) cloud-drift wind vectors **V** (two velocity components) from geostationary satellites, at one of two levels (2250 vectors); (c) **V** (two scalars) from aircraft and constant-level balloons (1100); (d) surface temperature T_s, wind $\mathbf{V}_s$ and pressure p_s (four) from land stations and ships (3450); (e) temperature T from polar-orbiting satellites (2050 × 5 levels); (f) **V** (two) from pilot ballons (660 × 10); (g) T, **V** and humidity q (four scalars) from radio- and dropsondes (750 × 10) (from Bengtsson *et al.*, 1981).

from radiosondes, ships, and land stations (see Fig. 1) led to an interesting realization. In fact, NWP operations had, of necessity, combined dynamics with observations all along in determining the state of the atmosphere at all times and in particular at those times from which forecasts had to be issued. Any weather bureau carries out two processes in parallel: one is the numerical forecast from a particular moment in time, or *epoch*, which we shall call initial time; the other is the 4-D assimilation of incoming data in order to estimate as well as possible the state of the atmosphere at the next epoch from which a forecast has to be issued.

Figure 2 shows the process of intermittent updating, in which all data within a certain interval, or window, are used together at the same epoch to update the state of the system as forecast by the NWP model (Bengtsson, 1975). Forecasts are typically started at so-called synoptic times, 0000 GMT and 1200 GMT, in which case a 12 hr assimilation cycle with ±6 hr windows is used. The subsynoptic times 0600 GMT and 1800 GMT also intervene when using a 6 hr cycle with ±3 hr windows. At analysis or update times, the numerical forecast is first verified against the new data and then combined or blended with them, i.e., the data are assimilated into the model. Finally, a new forecast is issued from the newly estimated state of the atmosphere.

The intermittent updating process described above was entirely appropriate as long as most data were taken, by international agreement, at the same time in order to provide a "synopsis" of the global weather; hence the word synoptic times and synoptic maps. With the advent of satellite data, time-continuous data assimilation, i.e., in practice every model time step (Ghil *et al.*, 1979), became possible. Thus, considerable interest developed throughout the 1970s in objective analysis and data assimilation methods, in preparation for

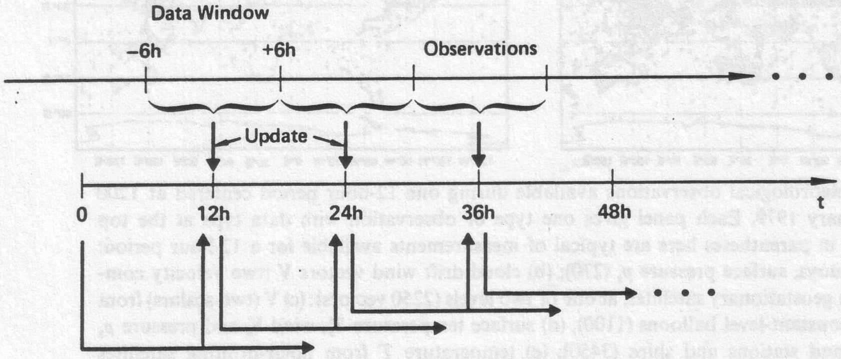


FIG. 2. Operational cycle of a weather service which combines the forecasting and data assimilation process (from Ghil, 1989).

the First GARP global experiment (FGGE), later relabeled the global weather experiment (GWE). The different methods in use by 10 advanced weather services at the end of that decade are reviewed by Gustafsson (1981) and by Ghil (1989). They are discussed further in Section 5.

At this point, we note merely that noisy, inaccurate data should not be fitted by exact interpolation, but rather by a procedure designed to achieve two goals simultaneously: (1) to extract the valuable information contained in the data, and (2) to filter out the spurious information, i.e., the noise. Thus, the analyzed field should be close to the data, but not too close. The *statistical* approach to this problem is linear regression. The *variational* approach is minimizing the distance, (e.g., in a quadratic norm) between the analyzed field and the data, subject to constraints that yield a smoother result. The connection between these two approaches in a stationary, ergodic context is intuitively obvious and is reflected in the fact that root–mean–square (rms) minimization is used in popular parlance for both approaches.

The analysis method in widest operational use today in NWP is a particular form of statistical interpolation, commonly referred to as optimal interpolation (OI) (Lorenc, 1981; McPherson *et al.*, 1979). Optimal interpolation is described within the broader context of estimation theory in Section 5.1. A particular implementation of variational methods, using the equations of motion as a strong constraint, is also being considered at present by some weather services (Courtier and Talagrand, 1987). This implementation by the adjoint method is discussed in Section 5.4 via the well-known duality between stochastic estimation and deterministic control (Gelb, 1974, Section 9.5). Oceanographic examples of the adjoint method are mentioned in Section 6. General reviews of meteorological analysis and assimilation methods are given by Bengtsson (1975), Bengtsson *et al.* (1981), Daley (1991), Thiébaux and Pedder (1987), and Williamson (1982). Brief unifying treatments are given by Ghil (1989), Lorenc (1986), Phillips (1982), and Wahba (1982). While providing its own unifying point of view, that of estimation and control theory, this chapter also addresses specific issues of data assimilation not covered by the preceding references, with a special emphasis on current and future applications to physical oceanography.

3. Atmosphere and Ocean: Dynamics, Data Sets, and Models

As discussed in Section 1, an important difference between oceanographic and meteorological data assimilation is in the motivation. Meteorological assimilation was driven at first by the need to forecast. At the present time and for the near future, oceanographic assimilation is and will be driven more by the need to understand better ocean dynamics through the blending of actually observed data with model-evaluated values of the same dynamical variables. This second approach emphasizes model parameter estimation, formal testing of the models against the data, and the need to calculate solution errors arising from the errors inherent in the model, in the data, and in their optimal blending. A discussion of these important aspects of oceanographic data assimilation will be given in Section 6 in the context of specific applications. Here we review the general similarities and differences between the two geofluids in physics and dynamics, in current and expected data sets, and in the numerical models used for each medium.

3.1. Dynamics and Thermodynamics

The similarities between the two geofluids are well known, and a unified theory is given by geophysical fluid dynamics (GFD). A number of books address GFD from this broad point of view (Ghil and Childress, 1987; Gill, 1982; Pedlosky, 1987). Nevertheless, crucial differences exist between the two fluid media. Hence, oceanographers cannot simply borrow the data assimilation techniques developed in meteorology; they must reinterpret the techniques and make them more suitable for oceanographic data sets.

Major similarities and differences are often paired. First, both atmosphere and ocean are forced, dissipative systems (Lorenz, 1963; Ghil and Childress, 1987, Section 5.4), but the atmosphere is forced only thermally, by equator-to-pole and land–sea temperature contrasts. Furthermore, this large-scale thermal forcing changes slowly on the time scale of purely deterministic prediction, 1 to 2 weeks.

By contrast, the major component of the ocean circulation on short time scales is the wind-driven circulation. Hence, in order to model, understand, and predict successfully oceanic currents, it is necessary to possess information on the ocean's internal dynamic variables, as well as on the surface forcing functions that drive these variables. Scatterometry will provide the wind-stress field at the sea surface with a space–time resolution adequate for global ocean circulation modeling and with reasonable accuracy.

The oceans' thermohaline circulation exhibits most of its variability on the much longer time scales of decades to millenia (Gill, 1982; Ghil *et al.*, 1987). But its short-term variability is also significant (Gill, 1982; Levitus, 1989), and it does interact strongly with the wind-driven circulation. Unfortunately, direct measurements of the heat and water fluxes, which drive the thermohaline component of the circulation, will not have adequate resolution and accuracy for global ocean modeling. The global distribution of the incoming solar radiation is relatively well known at the top of the atmosphere. At the sea surface, however, the thermodynamic fluxes have been modified by the dynamical and physical effects of the intervening fluid, the atmosphere, and are modified further by the reflecting and absorbing properties of the sea surface itself. These fluxes include: (1) incoming, short-wave solar radiation; (2) long-wave radiation reemitted at the sea surface; (3) sensible heat; (4) latent heat; (5) evaporation of water vapor; and (6) precipitation of liquid water. Their direct measurement, both remotely and *in situ*, is very difficult.

The heat and water fluxes are currently evaluated on a global basis through bulk formulae which depend upon a number of empirical coefficients, such as surface drag for wind stress. These bulk formulae provide order-of-magnitude estimates at best. Even in climatological studies, their sensitivity upon the specific values used for the best-fit coefficients may be so great as to reverse the sign of the total heat budget for a given basin (see, for instance, Bunker *et al.*, 1982). Hence, the oceanographic community will have to rely heavily upon numerical modeling and data assimilation to infer the thermohaline circulation, especially in the deep ocean layers where thermohaline processes are dominant.

In both ocean and atmosphere, dynamics and thermodynamics of the "pure" fluid, dry air or fresh water, interact through and are modified by a minor constituent: water in the atmosphere and salt in the ocean. Conservation equations for water in its three phases (in the atmosphere) and for salt (in the oceans) must be added to the equations of conservation of mass, momentum, and energy of the pure medium. The hydrologic cycle is, in fact, the most poorly observed component of atmospheric motions. Still, the atmospheric equation of state, even in the presence of diabatic processes, is relatively simple.

The ocean, on the other hand, is not a pure chemical solution with only one solute, which would still imply the existence of a unique functional relationship among three arbitrarily chosen thermodynamic variables, defining the equation of state of sea water. In fact, there are about 30 substances dissolved in the ocean, in quasi-steady relative proportions. These solutes are responsible for the "saltiness" of seawater quantified through the empirical concept of salinity, defined as the quantity of dissolved material in grams present in 1 kg of seawater. Because of the number and diversity of solutes, the equation of state for sea water is highly nonlinear and it is only available in the form of an empirical best fit, with consequent difficulties for data analysis and numerical modeling.

Another fundamental difference between the ocean and the atmosphere is that the ocean is essentially opaque to electromagnetic radiation. The atmospheric observational system is largely dependent upon radio and light waves for probing the atmosphere in the vertical and sending information back to the earth's surface and for looking into the atmosphere's interior from satellites. This is not possible for the ocean: one cannot see into its interior or communicate through it by electromagnetic means, only by acoustic ones. This major difference in the physics of the two fluids has had obvious and profound consequences for the capability of collecting synoptic data sets with global coverage, a capability not existing for the oceans and limited, even in the future, to the ocean surface only (Munk and Wunsch, 1982).

Further complications in oceanographic modeling are due to the presence of continents, which break the world ocean into major basins with complex geometries. This has two effects. First, the break in the longitudinally periodic configuration of the fluid makes it impossible to define a zonal-mean climatological component of the circulation analogous to the atmospheric subtropical jet. Many models and results for the atmosphere rely upon the expansion and linearization of the equations of motion around this dynamical mean state, which constitutes a considerable simplification. This powerful simplifying approach is impossible for the ocean.

Second, oceanic horizontal and vertical boundary conditions are much more complex. In the atmosphere, horizontal boundary conditions are periodic, which makes relatively simple spectral models extremely useful and efficient. At the upper boundary, a simple radiation condition suffices. In the ocean, continents introduce great flow distortions and multiple, model-dependent choices for the horizontal boundary conditions. At the surface, accurate knowledge of two major surface forcing functions, heat and momentum, is required, as previously discussed. Only the bottom boundary condition may be simpler, not because the ocean's bottom topography is less complex than the earth's surface topography, but because deep ocean motions are very weak, and the bottom boundary conditions can often be linearized.

All the differences between the two geofluids just reviewed briefly make the ocean system less easily tractable than the atmospheric one, with respect to realistic numerical modeling or the capability for data assimilation. In one respect, however, the ocean is simpler than the atmosphere, and this may simplify the development and adaptation of assimilation techniques. The ocean is a very stably stratified system with a time-constant, permanent pycnocline. Mixing occurs mostly along isopycnal surfaces, rather than across them. This stable stratification strongly inhibits vigorous vertical motions, and vertical velocities are usually of the order of 10^{-5} cm/sec compared to 1 cm/sec for the horizontal components. Unlike the atmosphere, where vertical convection plays a crucial role in the dynamics, deep convection cells in the ocean are very limited in horizontal extent; they are mostly confined to the North and South polar regions of the Atlantic, where the major water masses are formed.

The ocean's strong stratification also helps determine the most energetic scales and processes for the global ocean circulation. The counterpart of synoptic-scale cyclones in the atmosphere is mesoscale eddies in the mid-latitude ocean. Oceanic energy spectra (Wunsch, 1981) show a dominant mesoscale peak for the time and space scales of the general circulation, which has no clear counterpart in a well-defined synoptic-scale atmospheric peak (Boer and Shepherd, 1983). Energetic mesoscale eddies near the surface have been shown to be, in all likelihood, the major internal driving force of the deep ocean circulation (Holland and Rhines, 1980). The characteristic scale of oceanic mesoscale eddies is two to three times the internal Rossby radius of deformation, which is determined by the local stratification profile and in midlatitudes equals typically 50 km. Hence, in the ocean, the Burger number B satisfies $B < 1$ or $B \ll 1$ (Gill, 1982), while the most energetic, synoptic motions in the atmosphere have length scales comparable to or less than the Rossby deformation radius, $B \geq O(1)$ (Ghil and Childress, 1987, Section 4.3). Thus, we can expect, independently of the type of forcing (mechanical or thermal), that the ratio of gravity-wave energy to Rossby-wave energy in the ocean is smaller, and that the dissipation mechanisms of a stable numerical model are much more effective in damping the gravity waves.

The dominance of the mesoscale energy peak allows a major simplification in the oceanographic equations of motion, leading to the quasi-geostrophic (QG) approximation, which automatically filters out gravity-wave noise. Even when considering PE dynamics, the rigid-lid approximation is typically used in ocean models, thus filtering out the most troublesome waves, the surface gravity waves. This is done not only for reasons of computational economy, but also because surface tides are basically a linear phenomenon, well understood and highly predictable, and because the tides do not interact significantly with the dynamical processes and energy exchanges of the oceanic general circulation. We shall return to these simplifications of the dynamics in the discussion of available oceanographic models, Section 3.3.

We suspect, for all the heuristic reasons given previously, that the initialization "shocks" observed in some atmospheric PE models in the absence of appropriate initialization procedures will not constitute as important a problem in oceanic PE models, as long as surface gravity waves are not present thanks to the rigid-lid approximation. This inference is supported in fact by the results of Thompson (1986) and Malanotte-Rizzoli *et al.* (1989). In Section 6.2, we show that strongly unbalanced initial data are in fact required to produce strong internal gravity-wave noise in an oceanic PE model endowed with realistic stratification.

3.2. Data Sets in the Atmosphere and Ocean

To measure the distance with respect to availability of data between physical oceanography and dynamical meteorology at present, let us compare Fig. 1 with Figs. 3 and 4. Figure 1 represents the data typically available at present over one synoptic period, i.e., over 12 h for the global atmosphere.

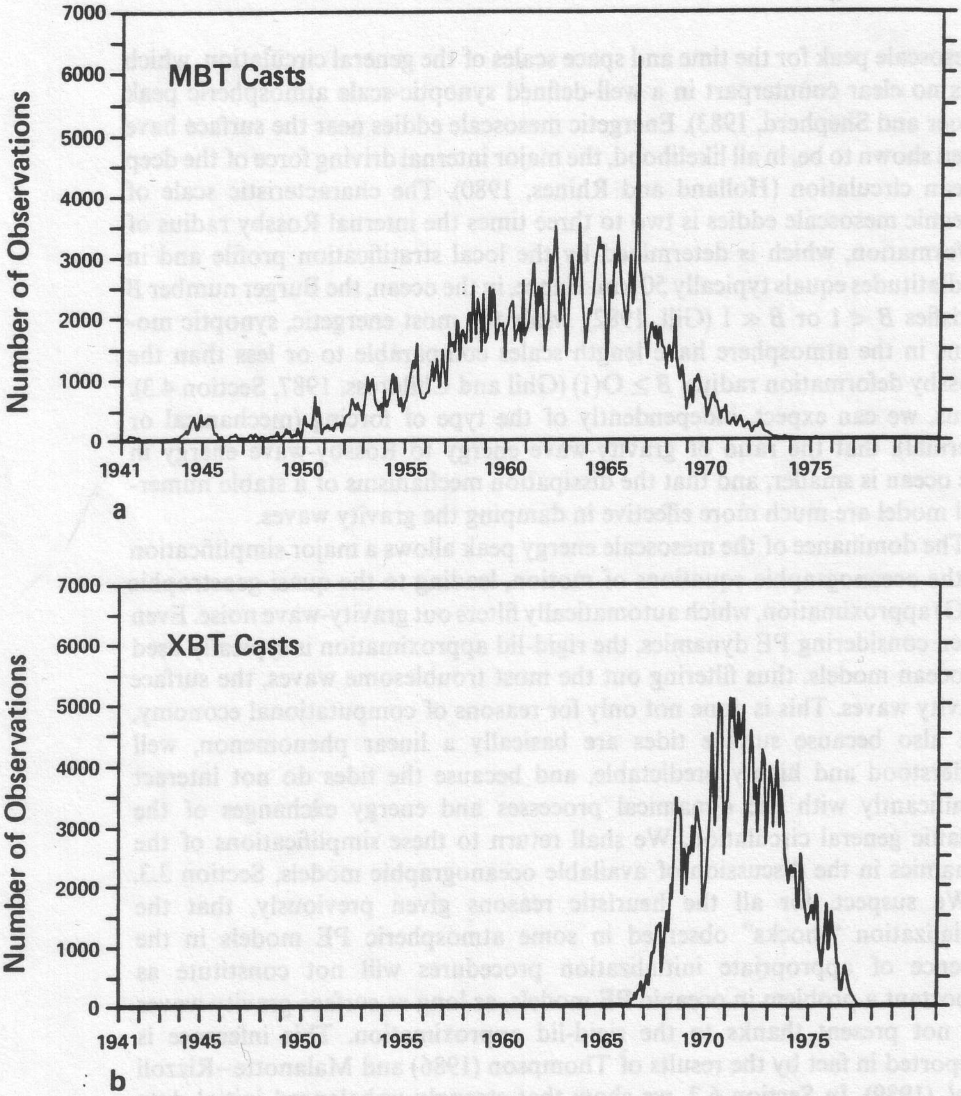


FIG. 3. Number of bathythermograph casts per month, 1941–1977. (a) MBTs; (b) XBTs (after Levitus, 1982).

Figures 3 and 4 represent the distribution in space and time of all oceanographic data up to 1978, archived by the National Oceanographic Data Center (NODC), Washington, D.C.

In Fig. 1, the total number of scalar measurements of the atmospheric mass and velocity fields over 12 hr is of the order of 10^5 (Ghil, 1986, 1989). This

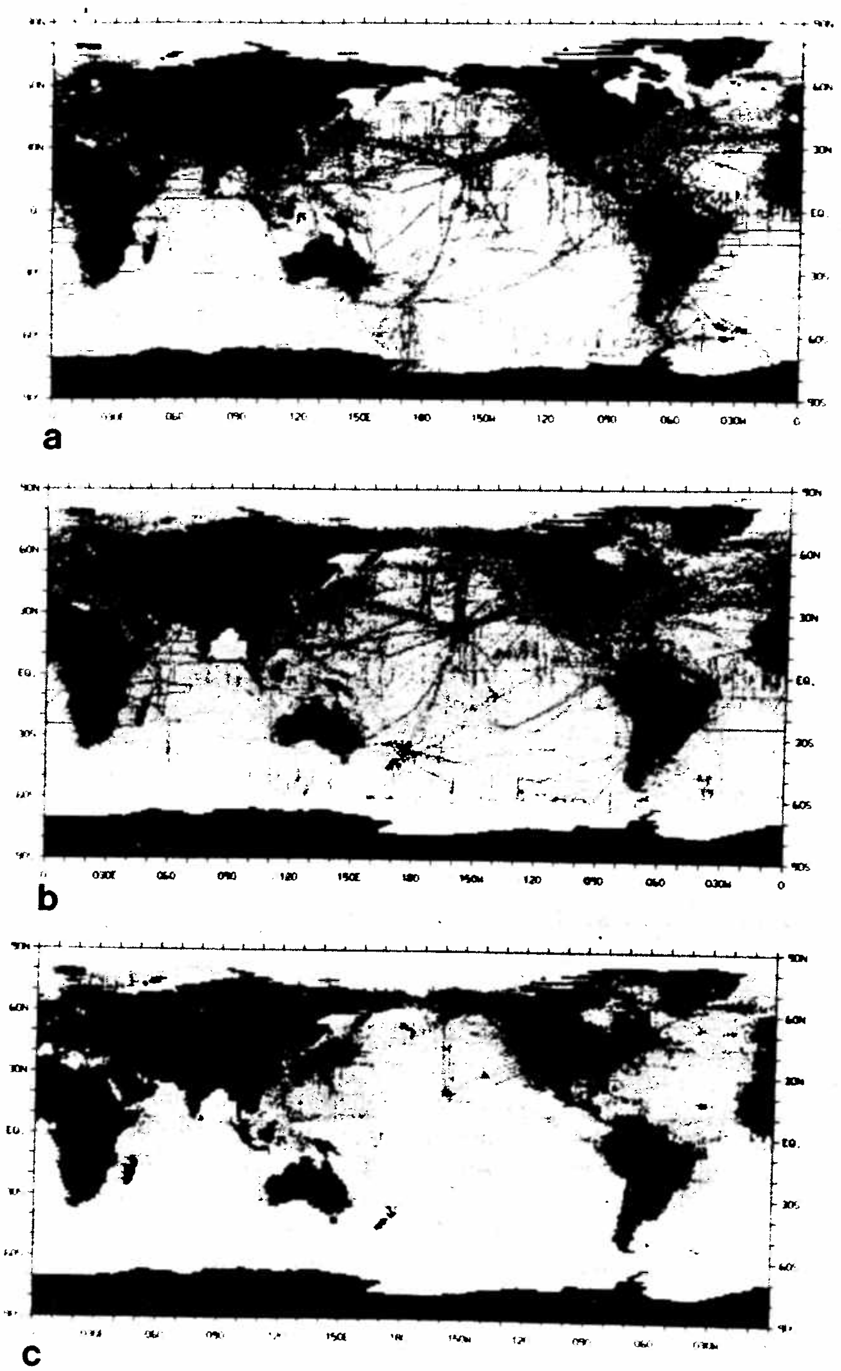


FIG. 4. Horizontal distribution of oceanographic observations at the sea surface: one-degree squares containing 1–4 observations (small dot) or 5 or more observations (large dot) from the merged NODC data set (SD + MBT + XBT). (a) Temperature for Northern Hemisphere (NH) winter (Feb., March, April); (b) temperature for NH summer (Aug., Sept., Oct.); (c) salinity for NH summer (from Levitus, 1982).

number is essentially adequate for a description of large-scale atmospheric fields, by using the methods of data assimilation into weather prediction models, which are currently operational in major weather bureaus. The test of adequacy here is relatively accurate prediction for a few days or a few synoptic periods.

The total number of archived oceanographic mass-field measurements over a period of 80 years or so is of the order of 10^7: (a) temperature T and salinity S from Nansen casts at about 500,000 hydrographic stations: (b) T from about 785,000 mechanical bathythermograph (MBT) and about 300,000 expendable bathythermograph (XBT) soundings, each with its own vertical distribution of individual measurements (Levitus, 1982). The situation for the ocean's velocity field is rather worse than for the mass field.

On the face of it, taking the number of atmospheric observations as the yardstick, there are 10^2 times more oceanic observations for a period of 10^5 times longer, i.e., 10^3 times fewer observations. This first estimate has to be corrected by allowing for the different time and space scales of the basic phenomena to be observed and predicted in the atmosphere and in the ocean. The Rossby radius of deformation, which is the characteristic length scale in both geofluids, is $O(10^2)$ km in the ocean vs. $O(10^3)$ km in the atmosphere, thus requiring an observational density 10^2 times higher. This is compensated only partially by the longer characteristic time in the oceans, requiring a frequency of observation 10 times lower than the atmosphere. The corrected estimate is therefore of 10^4 times fewer observations in the ocean.

Not only have oceanographers been accustomed to very few observations, but these are even more unevenly distributed in space and time. Figure 3 shows the distribution in time of MBT and XBT casts. The XBT is a more accurate and convenient instrument than the MBT, which it has essentially replaced. Unfortunately, the number of XBT casts has actually decreased, and there is also a lag in the entry of some XBT measurements into the NODC files. The distribution of observations in space, horizontally (Fig. 4) and with depth (Fig. 5), is also very uneven. Most data are in the Northern Hemisphere (NH, dotted line in Fig. 5), and there is further concentration of data in western boundary currents and along shipping lanes. The amount of data below the permanent thermocline is a tiny fraction of the total, and decrease of information with depth is quite rapid in the upper ocean as well.

In contrast to this situation, valid until just a few years ago, there are already about 40,000 satellite sea-surface temperature measurements daily. In addition, in the early 1990s, about 50,000 sea-surface height measurements and 180,000 surface wind vectors will be available daily (Halpern, 1987). Thus, the daily number of measurements in oceanography will become comparable to that currently available in meteorology. Even so, two problems remain; first, this is still a factor of 10 smaller, due to the difference in characteristic

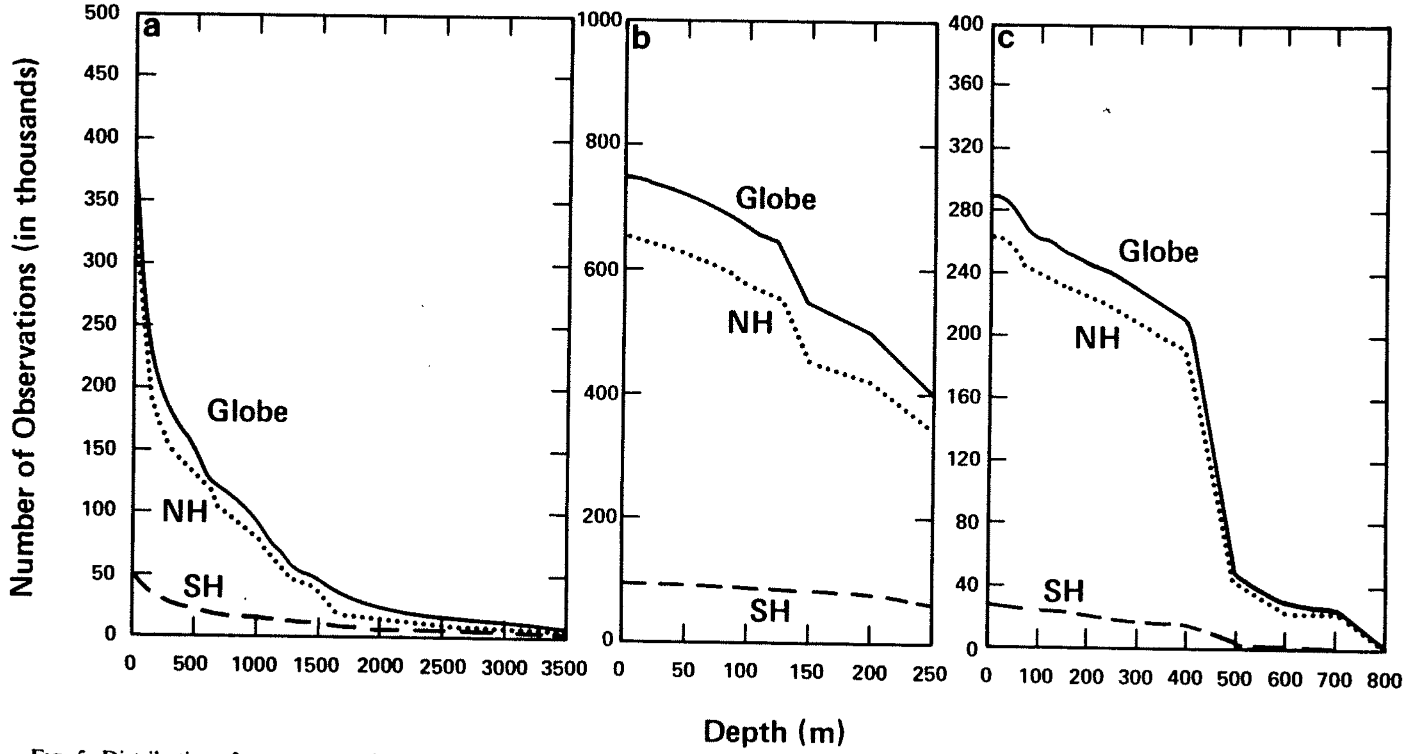


FIG. 5. Distribution of temperature observations at standard levels in NODC archives. (a) Station data (SD); (b) MBTs; (c) XBTs. Each panel shows the distribution with depth over the globe (solid), Northern Hemisphere (NH) (dotted), and Southern Hemisphere (SH) (dashed). Notice that both the ordinate and the abscissa in each panel have different scales (after Levitus, 1982).

scales; and second, the additional data mentioned are all surface data. It is hoped that the number of vertical soundings will increase somewhat, due to acoustic-tomography arrays and other advanced systems. But it is unlikely that this increase will be anywhere as dramatic as that for the surface. Furthermore, the data for the interior water mass, and especially the deep layers, will still be very unevenly distributed.

3.3. Oceanographic Models

Due to the smaller number and more uneven distribution of oceanographic data, a proportionately heavier burden will rely upon numerical models and data assimilation techniques to provide the dynamical interpolation of the circulation to data-poor water volumes and from observed to unobserved variables. Oceanographic models have followed their meteorological counterpart with a lag of about 10–20 years. As a result, most of them are still at the stage of craft rather than technology: each has been designed with a specific set of questions in mind or for limited domains; hence they are not portable in general and cannot be used for general-purpose global assimilation and prediction. Most numerical models in oceanography can be divided into four major categories:

(1) Quasi-geostrophic models, a prototype of which is discussed by Holland (1978).

(2) Layer models, based upon PE dynamics that use the adiabatic approximation. Examples are the subtropical gyre models of Holland and Lin (1975), models by Hurlburt (1986) and Thompson (1986) for the Gulf of Mexico, and, more recently, for the Gulf Stream system (Thompson and Schmitz, 1989); the equatorial models developed by Busalacchi and Picaut (1983), Luther and O'Brien (1985), Cane and Paden (1984) for different tropical oceans; and the Bleck and Boudra model (1986) in isopycnal coordinates.

(3) Primitive Equation models endowed with active thermodynamics, the most complete prototype of which, known as the Geophysical Fluid Dynamics Laboratory (GFDL) model, has been developed by Bryan (1969) and Cox (1984). It is the only ocean general circulation model (GCM) made available to scientists all over the world and is presently used for a vast variety of modeling and data assimilation studies. A Semi-Spectral PE (SPEM) model has been constructed by Haidvogel *et al.* (1991) and is being used in a variety of applications.

(4) Intermediate Models (IM) based upon different forms of the balance equation have been proposed by Gent and McWilliams (1984). They are not yet in widespread and standard use, but rather are in the development stage. Portable QG and PE versions of a limited-area, open-boundary regional

model have been developed by the Harvard group (Robinson and Walstad, 1987) and made available to other users.

Oceanographic QG models have been quite successful in simulating the mesoscale eddy field, the basin-wide, wind-driven general circulation, and the interactions between the two scales. This success reinforces the point made earlier about the QG approximation being even more appropriate in the ocean than in the atmosphere. The oceanic inertial-gravity wave and mesoscale bands of the spectrum appear to be sufficiently well separated so that little energy leaks from the one to the other. The mesoscale peak is moreover two orders of magnitude greater. Outside of relatively small regions characterized by systems endowed with strong curvature, like jet meanders and ring structures, the oceanic Rossby number is small enough for QG dynamics to model adequately mesoscale and planetary scale processes and reproduce their energetics accurately.

Still, PE models are needed to represent the thermohaline circulation and its interaction with the wind-driven circulation. Even in the more advanced PE models, however, including the complete GFDL model with active thermodynamics, the rigid-lid approximation is commonly used. This is justified by surface gravity waves, e.g., barotropic tides, being unimportant in the energy budgets and fluxes of the large-scale circulation. The rigid-lid approximation filters out the fastest component of the gravity-wave noise, and the model dissipation mechanisms suffice to damp out the much slower internal waves. Thus, ocean dynamics allow major simplifications in the governing equations without impairing the realistic representation of oceanic motions and of their energetics.

4. Estimation Theory and Data Assimilation

In this section, we give a brief review of the theoretical framework of estimation and control theory that provides the foundation of data assimilation techniques. Meteorological applications of the theory will be discussed in Section 5; oceanographic applications in Section 6. Statistical methods are closer to the estimation aspects of the theory, variational methods to the control aspects. The connections between the two aspects should become clear. The basic mathematical concepts of estimation and minimization can be given quite simply.

Take two measurements y and z of an observable x. The simplest estimate of the variable x, $\hat{x}$, is sought as a linear combination of y and z:

$$\hat{x} = \alpha_1 y + \alpha_2 z \qquad (4.1a)$$

We assume that the two measurements are unbiased,

$$Ey = Ez = Ex \tag{4.1b}$$

E is the expectation operator, the mean or average of a theoretically infinite number of measurements, and Ex is not known *a priori.* Requiring the estimate itself to be unbiased, $E\hat{x} = Ex$ immediately implies that $\alpha_1 + \alpha_2 = 1$ and hence Eq. (4.1a) can be rewritten as

$$\hat{x} = y + \alpha_2(z - y) \tag{4.2a}$$

We also assume that the measurement errors are uncorrelated,

$$E(y - Ex)(z - Ex) = 0 \tag{4.2b}$$

and that their variances σ_1^2 and σ_2^2 are known

$$\sigma_1^2 \equiv E(y - Ex)^2, \qquad \sigma_2^2 \equiv E(z - Ex)^2 \tag{4.2c,d}$$

The *optimal* linear unbiased estimate of x is given by choosing α_2 and $\alpha_1 = 1 - \alpha_2$ so as to *minimize the variance*

$$\sigma^2 = E(\hat{x} - x)^2 \tag{4.3}$$

of the estimation error. The required minimum is achieved by choosing

$$\alpha_1 = \hat{\sigma}^2/\sigma_1^2, \qquad \alpha_2 = \hat{\sigma}^2/\sigma_2^2 \tag{4.4a,b}$$

here $\hat{\sigma}^2$ is just the variance of the optimal estimate given by

$$\hat{\sigma}^{-2} = \sigma_1^{-2} + \sigma_2^{-2} \tag{4.4c}$$

The weights α_1 and α_2 thus reflect the relative uncertainties in y and z, respectively, and $\hat{\sigma}^2$ is smaller than both σ_1^2 and σ_2^2. In fact, it is convenient to define as *accuracy* $\mathrm{A} \equiv \sigma^{-2}$ the inverse of the variance of a random error. With this terminology, Eq. (4.4c) states that the accuracy of a linear unbiased optimal estimate equals the sum of the accuracies of unbiased mutually uncorrelated measurements.

Formally, the variational approach would have required here to minimize $J = J(x; \beta_1, \beta_2)$,

$$J \equiv \beta_1(x - y)^2 + \beta_2(x - z)^2 \tag{4.5}$$

with respect to x for arbitrary weights β_1 and β_2. The result will be the same, Eq. (4.1) with (4.4), provided

$$\beta_1 = \alpha_1, \qquad \beta_2 = \alpha_2 \tag{4.6a,b}$$

or, eliminating $\hat{\sigma}^2$,

$$\beta_1 = \sigma_1^{-2}, \qquad \beta_2 = \sigma_2^{-2} \tag{4.6c,d}$$

Equation (4.5) appears to be simpler, since no explicit statistical assumptions (4.2a–d) need to be made. But Eq. (4.6) shows that implicit access to information similar to that given in Eq. (4.2c,d) is required.

As discussed by Lorenc (1986), Wahba (1978, 1982), and their references, this information can also be retrieved variationally, given a distribution of observations in space or time which by ergodicity is equivalent to a distribution in probability space. The variational problem has to be reformulated for these purposes as minimizing

$$J' \equiv (x - y)^2 + (x - z)^2 + \lambda x^2 \tag{4.7}$$

and the regularization or smoothing parameter λ has to be determined from the data. This can be done by a resampling scheme (Efron, 1982), such as the (i) bootstrap, (ii) (generalized) cross-validation, or (iii) the jackknife. The results should still be the same to within sampling error.

Thus, it is clear that preference for the statistical approach [Eqs. (4.1, 4.4)] or the variational approach [Eq. (4.7)] hinges on computational considerations. The relative efficiency of numerical algorithms derived from either approach cannot be determined from the previous pedagogical example, but only within the context of specific applications (Sections 5 and 6).

4.1. Sequential Estimation and Optimal Data Assimilation

Estimation theory deals with the solutions of randomly perturbed systems of differential equations, ODEs or PDEs, as determined from noisy data distributed arbitrarily in space and time. For our purposes, it is sufficient to consider the ODE, or lumped parameter case, in discrete time since any numerical model of the atmosphere or ocean has to be presented in such a finite form to modern computational devices (Ghil *et al.*, 1981).

With the insight gained from the previous example, a linear unbiased data assimilation scheme for the geofluid can be written as:

$$\mathbf{w}_k^{\mathrm{f}} = \Psi_{k-1}\mathbf{w}_{k-1}^{\mathrm{a}} \tag{4.8a}$$

$$\mathbf{w}_k^{\mathrm{a}} = \mathbf{w}_k^{\mathrm{f}} + K_k(\mathbf{w}_k^{\mathrm{o}} - H_k\mathbf{w}_k^{\mathrm{f}}) \tag{4.8b}$$

The state vector $\mathbf{w}$ represents all model variables, such as temperatures and velocity components, at a set of grid points or in the form of spectral or finite-element coefficients. The forecast model [Eq. (4.8a)] is advanced in discrete time steps Δt, $\mathbf{w}_k = \mathbf{w}(t_k)$, $t_k = k\Delta t$. Superscript f stands for the forecast, o for observations and a for the analysis. Ψ is the system matrix describing its dynamics, which are linear at first, and H_k is the observation matrix. Extension to nonlinear models is addressed in the following discussion.

In the applications we are interested in, the system matrix Ψ represents the discretized version of a partial differential operator. In principle, the discretization can be made by finite differences, spectral transform, or finite elements. In the case of finite-element or of implicit finite-difference methods, Eq. (4.8a) becomes

$$\Psi_k^{(1)}\mathbf{w}_k^{\mathrm{f}} = \Psi_{k-1}^{(2)}\mathbf{w}_{k-1}^{\mathrm{a}} \tag{4.8a$'$}$$

The matrix $\Psi_k^{(1)}$ is in either case band limited and invertible, so that (4.8a′) can be reduced to (4.8a) by writing

$$\Psi_{k-1} = [\Psi_k^{(1)}]^{-1}\Psi_{k-1}^{(2)} \tag{4.8a$''$}$$

In the examples given in Section 5 and in Ghil (1989), we use for simplicity explicit finite-difference methods. It is obviously desirable in data assimilation, as well as in numerical prediction and simulation to use stable discretization methods (but see Miller, 1986).

The observation vector $\mathbf{w}_k^{\mathrm{o}}$ has dimension $p_k \ll N$, where N is the dimension of $\mathbf{w}_k^{\mathrm{f}}$ and $\mathbf{w}_k^{\mathrm{a}}$. The matrix H_k represents the fact that only certain variables or combinations thereof are observed at a set of points much smaller than the total number of grid points (Figs. 1, 3, 4, and 5). For instance, remote soundings of radiance by polar-orbiting satellites combine atmospheric temperatures, or tomographic soundings of acoustic travel times combine oceanic densities. H_k also represents the interpolation of grid values to data location for a grid-point model and (inverse) spectral transforms to physical space for a spectral model. The vector $\boldsymbol{\eta}_k \equiv \mathbf{w}_k^{\mathrm{o}} - H_k\mathbf{w}_k^{\mathrm{f}}$ contains the new information provided by the data. It is called *innovation vector* in the engineering literature and *observational residual* in the meteorological literature.

Equation (4.8b) has the form of Eq. (4.2a) with $y = \mathbf{w}_k^{\mathrm{f}}$, $z = \mathbf{w}_k^{\mathrm{o}}$, and $\alpha_2 = K_k$. The conceptual difference between Eqs. (4.8) and (4.1) is that $\mathbf{w}_k^{\mathrm{f}}$ represents past observations, and the practical difference is that $p_k \neq N$, i.e., H_k is not square, and it may have a different size at each time step. In fact, all operational data assimilation schemes have the form of Eq. (4.8b), whether the model in Eq. (4.8a) is linear or nonlinear. Existing assimilation schemes differ from each other by the weight matrix K_k and we wish to find the optimal K_k in a precise sense to be defined forthwith; in the engineering literature, K_k is often called the gain matrix.

Optimality is defined in the context of the following assumptions. First the true evolution of the geofluid, $\mathbf{w}_k^{\mathrm{t}}$, is governed by

$$\mathbf{w}_k^{\mathrm{t}} = \Psi_{k-1}\mathbf{w}_{k-1}^{\mathrm{t}} + \mathbf{b}_{k-1}^{\mathrm{t}} \tag{4.9a}$$

where $\mathbf{b}_k^{\mathrm{t}}$ is a (Gaussian) white-noise sequence, i.e.,

$$E\mathbf{b}_k^{\mathrm{t}} = 0, \qquad E\mathbf{b}_k^{\mathrm{t}}(\mathbf{b}_l^{\mathrm{t}})^{\mathrm{T}} = Q_k\delta_{kl} \tag{4.9b,c}$$

δ_{kl} being the Kronecker delta, and superscript T indicates the transpose (of a column vector, in this case); $\mathbf{b}_k^t$ is called system noise by engineers and oceanographers and model error by some meteorologists. No difficulty arises by adding a deterministic forcing $\mathbf{b}_k$ to the governing equation (4.8a), in which case Eq. (4.9b) would become

$$E\mathbf{b}_k^t = \mathbf{b}_k \neq 0 \tag{4.9b$'$}$$

As discussed in Section 3, forcing is more important for oceanic than for atmospheric flows. But in a linear problem, one can always separate the particular forced solution from the homogeneous one. It is the lack of complete initial data for the latter which we wish to compensate for by observations distributed in time. Deterministic forcing is not considered further here, but it will be introduced when considering the duality between deterministic control and stochastic estimation in Sections 5 and 6.

Current NWP models are in fact close to perfect in the sense that their error is almost white. Balgovind *et al.* (1983) have shown that the potential vorticity error of the second-order accurate NWP model then in use at the NASA Goddard Laboratory for Atmospheres (GLA), verified at 24 hr and 36 hr against a special set of satellite and conventional data comparable to that in Fig. 1, is essentially random, stationary in time, and nearly white in space. Numerical models of the ocean are not at all at this stage, but systematic errors can be eliminated by physical insight, numerical trial and error, and by applying systematically estimation and control theory.

The second assumption used in optimizing the weight matrix K_k concerns the error model for the observations

$$\mathbf{w}_k^o = H_k\mathbf{w}_k^t + \mathbf{b}_k^o \tag{4.10a}$$

where $\mathbf{b}_k^o$ is the observational noise or measurement error. One assumes that $\mathbf{b}_k^o$ is also a (Gaussian) white-noise sequence,

$$E\mathbf{b}_k^o = 0, \qquad E\mathbf{b}_k^o(\mathbf{b}_l^o)^{\mathrm{T}} = R_k\delta_{kl} \tag{4.10b,c}$$

for convenience and without any real loss of generality, it is assumed that system noise and observational noise are uncorrelated with each other,

$$E\mathbf{b}_k^t(\mathbf{b}_k^o)^{\mathrm{T}} = 0 \tag{4.11}$$

The assumptions given by Eqs. (4.9)–(4.11) permit us to derive the evolution in time of the error covariance matrices

$$P_k^{\mathrm{f,a}} \equiv E(\mathbf{w}_k^{\mathrm{f,a}} - \mathbf{w}_k^t)(\mathbf{w}_k^{\mathrm{f,a}} - \mathbf{w}_k^t)^{\mathrm{T}} \tag{4.12}$$

of the forecast $\mathbf{w}_k^{\mathrm{f}}$ and the analysis $\mathbf{w}_k^{\mathrm{a}}$, respectively. This evolution follows from Eqs. (4.8), (4.9a), and (4.10a), using (4.9b,c), (4.10b,c), and (4.11), and it is governed by

$$P_k^f = \Psi_{k-1} P_{k-1}^a \Psi_{k-1}^T + Q_{k-1} \tag{4.13a}$$

$$P_k^a = (I - K_k H_k) P_k^f (I - K_k H_k)^T + K_k R_k K_k^T \tag{4.13b}$$

Hence, by advancing $P_k^{f,a}$ along with $\mathbf{w}_k^{f,a}$, one can know how well the true state $\mathbf{w}_k^t$ is estimated for any weight matrix K_k. This in turn permits one to determine the optimal K_k [cf. Eq. (4.15)].

There are two problems which arise at this point. First and foremost, one must consider the computational complexity of advancing in time the error covariance matrices. While Eqs. (4.8a,b) represent $0(N)$ computations per time step, Eqs. (4.13a,b) represent at face value $O(N^2)$ computations. This is quite tolerable for typical engineering applications with $N \leq 1000$, but prohibitively expensive for atmospheric and oceanic prediction or simulation models with $N \geq 10^5$. However, by exploiting special features of the dynamics matrix Ψ and the covariance matrix P, which arise in the latter applications, the operation count can be reduced to $O(N)$, i.e., it can be made comparable to that for currently operational, less sophisticated data assimilation methods (Parrish and Cohn, 1985; Todling and Ghil, 1990; and Section 5.3).

Second, the noise covariance matrices Q_k and R_k are assumed to be known in the subsequent derivation of the optimal K_k. This is not so in practice, and finding the actual magnitude of model errors and observational errors is an important function of the data assimilation process. An adaptive filter to achieve this in GFD was formulated by Dee *et al.* (1985). It was tested only for the linear, one-dimensional (1-D) shallow-water equations, and substantial future work on this problem is necessary.

The optimal weight matrix K_k at each time step is obtained by minimizing the expected mean-square (m-s) estimation error

$$J \equiv \operatorname{tr} P_k^a \equiv E(\mathbf{w}_k^a - \mathbf{w}_k^t)^T(\mathbf{w}_k^a - \mathbf{w}_k^t) \tag{4.14}$$

This is done by using Eq. (4.13b) for the matrix P_k^a and setting the derivative of J with respect to each element of K_k equal to zero. A unique, absolute minimum is attained for

$$K_k = K_k^* \equiv P_k^f H_k^T (H_k P_k^f H_k^T + R_k)^{-1} \tag{4.15}$$

The linear unbiased data assimilation schemes Eq. (4.8a,b) with the optimal gain matrix K_k^* in Eq. (4.15) is called the Kalman filter (Kalman, 1960). Its continuous-time counterpart [see Table IV(A) in Section 5.4] is often called the Kalman–Bucy filter (Kalman and Bucy, 1961).

To complete the analogy between the Kalman filter (K-filter) and the two-measurement example given at the beginning, it is useful to rewrite Eqs. (4.13b) and (4.15) as

$$(P_k^{\mathrm{a}})^{-1} = (P_k^{\mathrm{f}})^{-1} + H_k^{\mathrm{T}} R_k^{-1} H_k \tag{4.16a}$$

$$K_k^* = P_k^{\mathrm{a}} H_k^{\mathrm{T}} R_k^{-1} \tag{4.16b}$$

It then becomes clear that Eq. (4.16a,b) is the counterpart of Eq. (4.4a–c), i.e., the weight given to the current observations is inversely proportional to their variance, and the accuracy of the analysis is the sum of the accuracies of the forecast, based on the past observations, and of the current observations. Note that $H_k = 0$ and hence $K_k^* = 0$ when no observations are available at time k.

The formula for P_k^{a} [Eq. (4.13b)] can be simplified when $K_k = K_k^*$, and the entire filter with this simplification is rewritten here for easy reference:

$$\mathbf{w}_k^{\mathrm{f}} = \Psi_{k-1} \mathbf{w}_{k-1}^{\mathrm{a}} \tag{4.17a}$$

$$P_k^{\mathrm{f}} = \Psi_{k-1} P_{k-1}^{\mathrm{a}} \Psi_{k-1}^{\mathrm{T}} + Q_{k-1} \tag{4.17b}$$

$$K_k^* = P_k^{\mathrm{f}} H_k^{\mathrm{T}} (H_k P_k^{\mathrm{f}} H_k^{\mathrm{T}} + R_k)^{-1} \tag{4.17c}$$

$$P_k^{\mathrm{a}} = (I - K_k^* H_k) P_k^{\mathrm{f}} \tag{4.17d}$$

$$\mathbf{w}_k^{\mathrm{a}} = \mathbf{w}_k^{\mathrm{f}} + K_k^* (\mathbf{w}_k^{\mathrm{o}} - H_k \mathbf{w}_k^{\mathrm{f}}) \tag{4.17e}$$

It must be noted that the K-filter minimizes the estimation error variance not only at every time step, but over the entire interval over which data are provided. This fact, and connections to deterministic variational methods via control theory, are also discussed in the next subsection. Moreover, the filter [Eq. (4.17)] is *sequential* or recursive, i.e., current observations are discarded as soon as they are processed or assimilated. This is due simply to the filter's extracting all useful information from the innovation vector at each time step, by an application of Bayesian ideas in a dynamical context (Kalman, 1960; Lorenc, 1986; Ghil, 1989). The sequential nature of the K-filter [Eq. (4.17)] makes it conceptually easy to grasp, and it has great practical advantages as we shall see in the following discussion. It is probably the major reason for the astounding success of the K-filter, and of its various computational modifications (e.g., Bierman, 1977; Budgell, 1986a), in engineering applications. For the application of the K-filter to a 1-D linearized barotropic model of geophysical relevance, see Ghil *et al.* (1981), Miller (1986), Ghil (1989), and Section 5.2.1 here; a 2-D application is given in Section 5.3.1 here.

Various extensions of the optimal filter [Eq. (4.17)] to nonlinear models are possible (Jazwinski, 1970). A promising approach for GFD appears to be the extended Kalman filter (EKF) (Ghil *et al.*, 1982; Budgell, 1986b). It proceeds by successive linearizations in time of the nonlinear dynamics

$$\mathbf{w}_k^{\mathrm{f}} = \mathrm{N}_{k-1}(\mathbf{w}_{k-1}^{\mathrm{f}}) \tag{4.18}$$

Here Eq. (4.18) replaces the linear evolution of Eq. (4.8a), while Eq. (4.8b) is still used at analysis time to update the model with incoming data. The explicit, nonautonomous dependence of $\mathbf{N}$ on time t is indicated by the subscript k.

The EKF is defined by using Eq. (4.17) with Ψ_k being given by the Jacobian matrix

$$(\Psi_k)_{ij} = \left.\frac{\partial N_k^i(\mathbf{w})}{\partial \mathbf{w}^j}\right|_{\mathbf{w}=\mathbf{w}_k^{\mathrm{f}}} \tag{4.19}$$

With this Ψ_k, the error evolution in Eq. (4.13a) is still correct to first order in $\mathbf{w}_k^{\mathrm{t}} - \mathbf{w}_k^{\mathrm{f}}$. Operational experience with suboptimal filters in NWP, OI in particular, suggests that the linearization of Eq. (4.19) need in fact not be recomputed at every time step. Ψ_k can be kept fixed over time intervals over which the flow does not change dramatically. The problem of filter divergence, common in engineering applications, is likely to be encountered for planetary flows only in the presence of strong instability combined with strong nonlinearity of the flow (cf. Budgell, 1986b, for nonlinearity, and Miller, 1986, for instability). This is because nonlinearity in GFD is essentially quadratic and, while of paramount importance in long-term behavior (Ghil and Childress, 1987; Pedlosky, 1987), is typically well behaved over the short time spans involved in data assimilation. Nonlinear estimation and the EKF are discussed further in Section 5.3.2.

4.2. Variational Methods: Fundamentals and Variants

The point of view taken in the previous section is that of optimizing assimilation techniques [Eq. (4.8)] developed since the mid 1970s in NWP for the purposes of assimilating satellite data (Ghil *et al.*, 1979; Lorenc, 1981; McPherson *et al.*, 1979). A point of view which at first appears to be more general is to minimize the distance between a given trajectory $\mathscr{C} \equiv \{\mathbf{w}(t): 0 \le t \le t^*\}$ and a set of data $\mathbf{z}(t) = H(t)\mathbf{w}^{\mathrm{t}}(t) + \mathbf{b}^{\mathrm{o}}(t)$ over the time interval $0 \le t \le t^*$, subject to a dynamical or smoothness constraint $\mathbf{S} \equiv \mathbf{S}(\mathbf{w}, t) = 0$, i.e., minimize the functional J_{var},

$$J_{\mathrm{var}}[\mathbf{w}] \equiv \int_0^{t^*} \{\boldsymbol{\eta}^{\mathrm{T}}(t)A(t)\boldsymbol{\eta}(t) + \mathbf{S}^{\mathrm{T}}\Gamma(t)\mathbf{S}\}\, dt \tag{4.20}$$

Here we use for simplicity (Bennett and Budgell, 1987) continuous-time notation, but still let vectors stand for spatial dependence to retain some similarity with the notation in the previous section. The vector $\boldsymbol{\eta}(t) \equiv \mathbf{z} - H\mathbf{w}^{\mathrm{f}}$ is the observational residual mentioned already in the previous subsection, and $A(t)$ is typically zero, except at discrete times between 0 and t^* when observations are available. The variational formalism was introduced by Sasaki (1958) into dynamic meteorology and by Provost and Salmon (1986) into physical oceanography. $\mathrm{S}(\mathbf{w}; t)$ can represent the equations of motion with or without time dependence (Sasaki, 1970) or a smoothness constraint, such as the second derivative in one or more space dimensions or powers of the Laplacian in the geometry of interest (Wahba, 1982).

The matrices $A(t)$ and $\Gamma(t)$ are $p(t) \times p(t)$ and $N \times N$, respectively; they should be symmetric and positive definite. Both can be prescribed *a priori* or be computed adaptively from the data. If S is linear in the model variables and represents the dynamics, then the minimizing trajectory $\mathscr{C}$ is given by the Kalman–Bucy (1961) filter, which is the continuous-time extension of Eqs. (4.17) [e.g., Bucy and Joseph, 1987; Gelb, 1974; Table IV(A) here]. In this case, subject to a suitable generalization of the assumptions in Eqs. (4.9–4.11), the matrices $A(t)$ and $\Gamma(t)$ are proportional to the accuracies of the data and of the model, i.e., proportional to $R^{-1}(t)$ and $Q^{-1}(t)$, with an obvious extension of the discrete-time notation of Eqs. (4.9c) and (4.10c). The analogy with Eqs. (4.5, 4.6) in the simple example at the beginning of this section should be obvious.

In the oceanic case of data distributed very sparsely in time and space (Figs. 3–5), a steady-state form of Eq. (4.20) can be regarded as the generalized inversion of the rectangular matrix $H^* = \sum_{H \neq 0} H(t_k)$ summing over all observations at distinct times (and noncoincident locations). Various regularization constraints S have been used by Wunsch (1978), Fiadeiro and Veronis (1984), and Bennett and McIntosh (1982).

In the absence of statistical information like Eqs. (4.9–4.11), there are two standard formulations of the minimization problem in Eq. (4.20). Using the terminology of Sasaki (1970), the case in which $\mathrm{S} = 0$ is required to be satisfied exactly is that of a strong constraint. Requiring $\mathrm{S}^{\mathrm{T}}\Gamma\mathrm{S}$ to be minimized, not necessarily to zero, but only along with $\eta^{\mathrm{T}}A\eta$ is called imposing a weak constraint. For a weak constraint, $A(t)$ and $\Gamma(t)$ have to be prescribed independently of solving for $\mathbf{w}(t)$. In this case, the matrices A and Γ are typically chosen to be constant in time and have a very simple structure (e.g., $A = \alpha I_p$ and $\Gamma = \gamma I_N$) with I_p and I_N identity matrices of suitable dimensions; α and γ are scalars that have to reflect, however crudely, the relative confidence in model and data, respectively. For a strong constraint, $\Gamma \rightarrow \infty$, which is conceptually equivalent to letting the model error vanish, $Q \rightarrow 0$; $A(t) = \alpha I_p$ as before, in most applications.

There are two general approaches to the minimization problem of Eq. (4.20). One is to compute the first variation of J with respect to $\mathbf{w}(t)$, δJ, and obtain the extrema of J by setting $\delta J = 0$. To be a little more explicit, let us switch from vector notation to spatial dependence in a simplified, but typical case:

$$\begin{aligned} J[\phi, V] \equiv \int_0^{t^*}\int_\Sigma & \{\alpha(\mathbf{x}, t)h_\phi(\mathbf{x}, t)[\phi(\mathbf{x}, t) - \phi^o(\mathbf{x}, t)]^2 \\ & + \beta(\mathbf{x}, t)h_v(\mathbf{x}, t)\{\mathbf{V}(\mathbf{x}, t) - \mathbf{V}^o(\mathbf{x}, t)]^2 \\ & + \gamma(\mathbf{x}, t)s^\nu(\phi, \mathbf{V}; \mathbf{x}, t)\}\, d\Sigma\, dt \end{aligned} \tag{4.21}$$

Here ϕ is geopotential and $\mathbf{V}$ horizontal velocity, $\mathbf{x}$ is the position vector of horizontal and possibly vertical coordinates within the area (or volume) Σ, $\alpha(\mathbf{x}, t)$ and $\beta(\mathbf{x}, t)$ take the place of the matrix $A(t)$, with $\gamma(\mathbf{x}, t)$ playing the role of $\Gamma(t)$ and $s(\phi, V; \mathbf{x}, t)$ that of $\mathbf{S}(\mathbf{w}; t)$. The functions h_ϕ and h_v are sums of Dirac delta functions at the discrete points $(\mathbf{x}_i, t_j)$ at which observations of ϕ and V, respectively, exist. Their presence highlights the need for the constraints s, in order to obtain smooth fields $\phi(\mathbf{x}, t)$ and $V(\mathbf{x}, t)$ from discrete observations $\phi^o(\mathbf{x}_i, t_j)$ and $\mathbf{V}^o(\mathbf{x}_{i'}, t_{j'})$. The exponent ν equals 1 for a strong constraint and 2 for a weak constraint.

The classical variational approach of $\delta J = 0$ leads to the Euler–Lagrange equations for Eq. (4.21). This approach is particularly suited for the strong-constraint formulation, since $\gamma(\mathbf{x}, t)$ becomes a Lagrange multiplier, while α and β are prescribed *a priori*. A system of Euler–Lagrange PDEs is then obtained for the unknown functions ϕ, V, and γ. The form of this PDE system depends on the functional form of the constraint $s(\phi, \mathbf{V})$ and on the geometry; in general the system will be neither elliptic nor hyperbolic, but of mixed type (Stephens, 1970). Certain PDE systems of mixed type lead to well-posed initial boundary value problems and can be solved numerically with reasonable computational cost (Ghil *et al.*, 1977, and references therein). In most cases of real interest, however, this approach has not proved particularly useful or promising.

In the case of weak constraints, with α, β, and γ prescribed, $\delta J = 0$ can also lead to a set of PDEs for the minimizing solution (Bennett and McIntosh, 1982; Bennett and Budgell, 1989; Miller, 1987, pp. 18–23). Considerable numerical and regularity problems arise in all but the simplest problems, in this case as well.

The second approach of direct minimization by an iterative numerical algorithm is more in tune with modern computational tools. This approach circumvents the Euler–Lagrange equations and minimizes J in Eq. (4.21) directly with respect to the trajectory $\{\phi(\mathbf{x}, t), V(\mathbf{x}, t)\}$ yielding, if the constraint is strong, also $\gamma(\mathbf{x}, t)$. If α, β, and γ are given, positive constants (i.e., in the simplest case of a weak constraint) and s is linear, the functional J is quadratic in the solution and hence will have a unique minimum. This can be computed by discretizing Eq. (4.21) in space and time and minimizing the discretized $\tilde{J}$ with respect to all the components $\phi(\mathbf{x}_i, t_j)$ on a regular grid (Hoffmann, 1982).

A particularly efficient method for direct minimization of the distance between a trajectory $\mathscr{C}$ and data $\mathbf{z}$ over $0 \leq t \leq t^*$, subject to the strong constraint of flow equations, is the adjoint method. The functional to be minimized is

$$J_{\mathrm{Ad}}[\mathbf{w}] \equiv \int_0^{t^*} \mathscr{J}(\mathbf{w}(t); t)\, dt \tag{4.22}$$

where $\mathscr{J} = \eta^{\mathrm{T}} A \eta$, as in Eq. (4.20) for instance. The equations of motion are taken as the continuous-time form of Eq. (4.18), i.e.,

$$\dot{\mathbf{w}} = \mathbf{N}(\mathbf{w}) \tag{4.23}$$

with $(\dot{\ }) = d(\ )/dt$.

The method proceeds to find the initial vector $\mathbf{w}_0 = \mathbf{w}(0)$ for which the solution $\mathbf{w}(t)$ of Eq. (4.23) will minimize J_{Ad} in Eq. (4.22). To do so, one takes the first variation δJ with respect to $\delta \mathbf{w}_0$. An excellent derivation of the method is given by Talagrand and Courtier (1987). We summarize here only the steps of the algorithm.

(1) A first-guess trajectory $\mathbf{w}^{(1)}(t)$ is computed from some $\mathbf{w}_0^{(1)}$, and $J^{(1)} = J[\mathbf{w}^{(1)}]$ is calculated for it.

(2) The gradient of J with respect to $\mathbf{w}_0$, $\nabla_{\mathbf{w}_0} J$, is obtained in three steps:

(i) The flow Eqs. (4.23) are linearized about $\mathbf{w}(t) = \mathbf{w}^{(1)}(t)$,

$$\delta \dot{\mathbf{w}} = L_{\mathbf{w}}(t)\, \delta \mathbf{w} \tag{4.24a}$$

where $L_{\mathbf{w}}$ is the Jacobian matrix

$$L_{\mathbf{w}} = \left. \frac{\partial \mathbf{N}(\mathbf{u})}{\partial \mathbf{u}} \right|_{\mathbf{u} = \mathbf{w}(t)} \tag{4.24b}$$

(ii) One computes the adjoint $L_{\mathbf{w}}^*$ of $L_{\mathbf{w}}$ and the data-forcing function $\mathbf{M}_{\mathbf{w}}(t)$,

$$\mathbf{M}_{\mathbf{w}}(t) = \left. \frac{\partial \mathscr{J}(\mathbf{u}; t)}{\partial \mathbf{u}} \right|_{\mathbf{u} = \mathbf{w}(t)} \tag{4.25}$$

(iii) The inhomogeneous adjoint equation

$$-\delta' \dot{\mathbf{w}} = L_{\mathbf{w}}^* \delta' \mathbf{w} + \mathbf{M}_{\mathbf{w}}(t) \tag{4.26}$$

is integrated backward in time from final data $\delta' \mathbf{w}(t^*) = 0$. The solution of this problem, $\delta' \mathbf{w}(0)$ is the required gradient $\nabla_{\mathbf{w}_0} J^{(1)}$.

(3) The initial value $\mathbf{w}_0^{(1)}$ in Step (1) is corrected in successive descent steps, $n = 1, 2, \ldots$,

$$\mathbf{w}_0^{(n+1)} = \mathbf{w}_0^{(n)} - \rho_n \mathbf{D}_n \tag{4.27}$$

$\mathbf{D}_n$ is a descent direction determined from $\nabla_{\mathbf{w}_0} J^{(n)}$ and ρ_n is an appropriate scalar, the step length in the given direction.

If $\mathbf{D}_n$ and ρ_n are properly chosen, the sequence $\mathbf{w}_0^{(n)}$ will tend to a $\mathbf{w}_0^{(\infty)}$, which leads to the trajectory $\mathscr{C}^{(\infty)}$ minimizing $J_{\mathrm{Ad}}[\mathbf{w}]$, at least locally in phase space. Various descent algorithms, including in particular steepest descent, conjugate gradient, and quasi-Newton (QN) are described by Gill *et al.* (1982).

This method was introduced into GFD by Marchuk (1975). Its most general form for the related problem of sensitivity analysis was presented by Cacuci (1981) and applied to a general circulation model by Hall (1986). Assimilation results with a 2-D version of the inviscid barotropic vorticity equation and 24 hr of radiosonde observations over the Northern Hemisphere were obtained by Courtier and Talagrand (1987). The slightly different problem of recovering wind stress in the tropics from oceanographic data was treated by Thacker and Long (1988).

Data assimilation for fully nonlinear problems such as Eqs. (4.18) or (4.23), in GFD and elsewhere, is incompletely understood, and no algorithms satisfactory in all cases exist so far. Multiple minima of the cost functional [Eqs. (4.21) or (4.22)] and rapidly changing growth rates of instabilities along a trajectory approximating a minimum are only some of the difficulties encountered (see Sections 5.3.2 and 6.3.4). Linearization is necessary in both the sequential estimation approach (Sections 4.1 and 5.3.2) and the direct minimization approach (Section 5.4). The EKF [Eq. (4.19)] and its generalizations proceed by successive linearizations in time along a given trajectory, while the adjoint method and its variants proceed by successive linearizations in function space over the entire time interval in question. In neither case is the optimal solution of the problem guaranteed *a priori.*

The adjoint method has no direct access to statistical information, and it is not clear at this point how model errors can be taken into account, while the K-filter imposes a very large computational burden in order to provide the necessary error estimates. The relative advantages and disadvantages of statistical and variational methods are a matter largely of numerical and practical considerations (Lorenc, 1986). These are all changing rapidly in a climate of intense research and of swift improvement in the computing environment of GFD.

5. Current Status of Meterological Data Assimilation

This section is devoted to the meteorological applications of the estimation and control theory discussed in Section 4. Meteorological data assimilation is a mature subdiscipline, characterized by the following features:

(1) Well-developed numerical prediction models with forecasts validated on a daily basis against observations;

(2) A large number of observations, albeit distributed irregularly in space and time and possessed of considerable variety in their statistical properties;

(3) A relatively broad consensus on operational assimilation methods, along with an active program of exploration into more advanced algorithms.

The emergence of this situation over a period of 40 years was described in Section 2. The theoretical basis of data assimilation methods was outlined in Section 4. We start this section with a description of the statistical, sequential data-assimilation method in broadest operational use today, optimal interpolation (OI). This is followed by a description of the initialization problem and of some of its solutions. The section concludes with research implementations of two classes of advanced methods described in Sections 4.1 and 4.2, the Kalman filter of sequential estimation theory, and the adjoint method of control theory, respectively.

5.1. Optimal Interpolation

Within the general framework of sequential estimation theory (Section 4.1), OI (Rutherford, 1972; Schlatter *et al.*, 1976; McPherson *et al.*, 1979; Lorenc, 1981) is a particular suboptimal filter in which intermittent updating is used and the true forecast error covariance matrix P_k^{f}, defined in Eq. (4.12), is replaced by an approximation, S_k^{f}. This approximation is computed (Cohn *et al.*, 1981; Ghil *et al.*, 1982; Marshall, 1985a) as the product of a time-independent correlation matrix C and a diagonal variance matrix D_k^{f},

$$S_k^{\mathrm{f}} = (D_k^{\mathrm{f}})^{1/2} C (D_k^{\mathrm{f}})^{1/2} \tag{5.1}$$

The matrix C is based on assuming that the mass field errors are homogeneous, isotropic, and that their correlations have a Gaussian dependence on distance in a horizontal tangent plane

$$C_{ij}^{\phi\phi} = \exp\{-\|\mathbf{x}_i - \mathbf{x}_j\|^2/s_0^2\} \tag{5.2a}$$

or on the surface of a sphere of given radius, i.e., at a given height. Here $\|\mathbf{x}_i - \mathbf{x}_j\|$ is the usual (linear or spherical) distance between the two points $\mathbf{x}_i$ and $\mathbf{x}_j$ and s_0 is a decorrelation distance, chosen typically as 500 km $\leq s_0 \leq$ 1000 km (Ghil *et al.*, 1979; McPherson *et al.*, 1979; Lorenc, 1981; Hollingsworth *et al.*, 1985), i.e., comparable to the Rossby radius of deformation and hence to the characteristic size of synoptic systems in the medium. The additional assumption used to determine C is that the geopotential and velocity forecast errors are geostrophically related, and hence the cross-correlations between the velocity and mass fields and between the velocity components themselves are derived from Eq. (5.2a) by partial differentiation with respect to the eastward and northward horizontal coordinates,

x^1 and x^2,

$$C_{ij}^{\phi u_\alpha} = \sqrt{2}\{(x_i^\beta - x_j^\beta)/s_0\} C_{ij}^{\phi\phi} = -C_{ij}^{u_\alpha \phi} \tag{5.2b,c}$$

$$C_{ij}^{u_\alpha u_\alpha} = \{1 - 2(x_i^\beta - x_j^\beta)^2/s_0^2\} C_{ij}^{\phi\phi} \tag{5.2d}$$

$$C_{ij}^{u_\alpha u_\beta} = 2\{(x_i^\alpha - x_j^\alpha)(x_i^\beta - x_j^\beta)/s_0^2\} C_{ij}^{\phi\phi} \tag{5.2e}$$

Here the Greek subscripts α and β correspond to eastward (1) or northward (2) components, and $\alpha \neq \beta$ (i.e., if $\alpha = 1$ then $\beta = 2$ and vice versa).

The geostrophic assumption for the forecast errors breaks down, obviously, near the equator, and relations (5.2b–e) have to be modified there. Vertical correlations are treated at present differently from the horizontal correlations (Lönnberg and Hollinsworth, 1986; Baker *et al.*, 1987), since the analysis is done separately on pressure or sigma-coordinate surfaces. A unified, truly 3-D treatment of forecast error correlations would be an important step in the direction of true optimality; indeed, baroclinic instability affects strongly short-range forecasts and has a fully 3-D structure.

The evolution of the forecast error variance matrix D_k^{f} between update times $k' = (J - 1)r$ and $k'' = Jr$ is prescribed,

$$D_{k''}^{\mathrm{f}} = D_{k'}^{\mathrm{a}} + D \tag{5.3a}$$

where D is an empirically determined approximation of mean forecast error growth over r model time steps (6 h or 12 h, cf. Section 2). At update time, the new D_k^{a} is obtained by using Eqs. (5.1, 5.2) and

$$S_k^{\mathrm{a}} = (I - K_k H_k) S_k^{\mathrm{f}} (I - K_k H_k)^{\mathrm{T}} + K_k R_k K_k^{\mathrm{T}} \tag{5.3b}$$

$$D_k^{\mathrm{a}} = \mathrm{diag}(S_k^{\mathrm{a}}) \tag{5.3c}$$

with $k = k''$. Thus Eqs. (5.1–5.3) are the OI counterparts of Eqs. (4.13a,b) of sequential estimation, and (4.15) is still used, with P_k^{f} replaced by S_k^{f},

$$K_k^{\mathrm{OI}} = S_k^{\mathrm{f}} H_k^{\mathrm{T}} (H_k S_k^{\mathrm{f}} H_k^{\mathrm{T}} + R_k)^{-1} \tag{5.4}$$

The OI procedure uses exactly the same forecast and update equations as the general linear unbiased data assimilation scheme [Eqs. (4.18, 4.8b)],

$$\mathbf{w}_k^{\mathrm{f}} = N_{k-1}(\mathbf{w}_{k-1}^{\mathrm{a}}) \tag{5.5a}$$

$$\mathbf{w}_k^{\mathrm{a}} = w_k^{\mathrm{f}} + K_k^{\mathrm{OI}}(\mathbf{w}_k^{\mathrm{o}} - H_k \mathbf{w}_k^{\mathrm{f}}) \tag{5.5b}$$

Equations (5.1–5.5) describe completely, in compact vector-matrix notation, the OI assimilation procedure.

In a practical assimilation cycle, beside the operations implied by Eqs. (5.1, 5.3a–c, and 5.5b), considerable work is expended on the related problems of quality control and data selection (Gustafsson, 1981; Gandin, 1988;

Lorenc and Hammon, 1988; Hollingsworth, 1989). The selection and quality control of data in OI aim at reducing the computational burden, severely limiting the number of measurements used to update any given grid point value, and aim at eliminating altogether outliers, i.e., data with very large apparent errors (Ghil, 1989, Section 6). The total computational expense of a typical 24 hr assimilation cycle exceeds considerably the expense of a pure forecast for the same period of time; it is comparable to that of a 3–5 day forecast issued every day (L. Bengtsson, personal communication, 1989).

Still, OI is obviously much less expensive than a straightforward implementation of the full K-filter, with its $0(N^2)$ operations. It is in turn much more expensive than other data assimilation methods, such as direct insertion or the successive-correction method (Bergthorsson and Döös, 1955; Cressman, 1959). The variety of methods in operational use at the end of the 1970s is illustrated in Table I.

A comparison of three data assimilation methods and variations thereof was carried out by Ghil *et al.* (1979) in the context of maximizing the impact of satellite-derived temperature observations on the accuracy of numerical weather forecasts. The three basic methods were direct insertion, successive corrections, and OI. All three methods were applied in a time-continuous rather than intermittent mode to the temperature retrievals, while conventional data were all assimilated at synoptic times by the same method, successive corrections. It should be noted that temperature observations from polar-orbiting satellites exceed at present in number any other class of observations, and are comparable to all others combined (see Fig. 1).

The results of this comparison are shown in Table II. The data are designated as NoSat, i.e., conventional data only, or Sat, i.e., all data available during the Data System Test (DST-6) conducted by the U.S. National Aeronautics and Space Administration (NASA), January–March 1976. The methods are designated as DIM for direct-insertion method, SCM for successive-correction method, and SAM for statistical assimilation method; the latter is essentially a time-continuous version [see especially Fig. 1 and Eq. (16) of Ghil *et al.*, 1979] of OI.

The impact of the satellite data was also determined by careful consideration of the changes in initial states and by the subjective evaluation of the quality of the forecasts. These results (not shown) are consistent overall with the numerical measures of impact given in the table. Listed are the improvements in S_1-skill score, which is a nondimensional measure of accuracy for gradients in the height field, and in the root–mean–square (rms) difference from a validating analysis, both given for the 500 mb height field. Statistical significance is measured by the average difference divided by the standard error for the set of forecasts; values of 0.5, 1.0, and 2.0 correspond to confidence levels of 69%, 84%, and 98%, respectively.

TABLE I. CHARACTERISTICS OF DATA ASSIMILATION SCHEMES IN OPERATIONAL USE AT THE END OF THE 1970s[a]

Organization or country	Operational analysis methods	Analysis area	Analysis/forecast
Australia	Successive correction method (SCM)	SH[d]	12 hr
	Variational blending techniques	Regional	6 hr
Canada	Multivariate 3-D statistical interpolation	NH[d] Regional	6 hr (3 hr for the surface)
France	SCM; wind-field and mass-field balance through first guess	NH	6 hr
	Multivariate 3-D statistical interpolation	Regional	
F.R. Germany	SCM. Upper-air analyses were built up, level by level, from the surface	NH	12 hr (6 hr for the surface)
	Variational height/wind adjustment		Climatology only as preliminary fields
Japan	SCM	NH	12 hr
	Height-field analyses were corrected by wind analyses	Regional	
Sweden	Univariate 3-D statistical interpolation	NH	12 hr
	Variational height/wind adjustment	Regional	3 hr
United Kingdom	Hemispheric orthogonal polynomial method		
	Univariate statistical interpolation (repeated insertion of data)	Global	6 hr
U.S.A.	Spectral 3-D analysis	Global	
	Multivariate 3-D statistical interpolation	Global	6 hr
U.S.S.R.	2-D[c] statistical interpolation	NH	12 hr
ECMWF[b]	Multivariate 3-D statistical interpolation	Global	6 hr

[a] After Gustafsson (1981).

[b] European Centre for Medium Range Weather Forecasts.

[c] 2-D is in a horizontal plane.

[d] Southern Hemisphere and Northern Hemisphere, respectively.

TABLE II. IMPROVEMENT IN THE MEAN ACCURACY OF 11 48-hr FORECASTS AS A FUNCTION OF ASSIMILATION METHOD[a]

			Percent impact		Statistical significance	
Experiment	Data	Method	S_1	rms	S_1	rms
NO	No Sat	(SCM)	0	0	–	–
DN	Sat	DIM	0.21	2.43	0.13	0.94
C2i	Sat	SCM, intermittent	2.79	3.28	1.75	1.33
C2t	Sat	SCM, time-continuously	5.01	9.31	2.10	2.34
S2μ	Sat	SAM	4.09	12.09	1.99	4.08

[a] Due to the use of satellite data (after Ghil *et al.*, 1979).

It is clear from the table that use of the satellite data does provide an improvement in the short-range forecasts. While negative impacts were present on certain days (Tables 2 and 4 of Ghil *et al.*, 1979), and on other days the improvements were not synoptically significant (Section 4c of Ghil *et al.*, 1979), the average impact is positive and it is dominated by a number of large, synoptically significant cases of improvement in the initial states. These results, while hotly debated at the time (Tracton *et al.*, 1981), are generally accepted today, and satellite data are in broad operational use.

But it is also apparent from Table II that a relatively unsophisticated method such as DIM does little to extract the information content of the data. For the same method, SCM, time-continuous assimilation (C2t) is much more efficient than the intermittent version (C2i) at extracting information from the continuous datastream. A number of operational and research centers are considering at present 3-hr updating intervals, as recommended by Ghil *et al.* (1979), instead of the 12-hr cycles still largely in use at the end of the 1970s (Table I) or the 6-hr cycles in prevalent use now.

Finally, the time-continuous OI method designated as SAM in Table II provides a further substantial improvement over time-continuous SCM. This is, however, less dramatic than that of SCM over DIM and, in fact, Bratseth (1986) has shown that SCM can be formulated iteratively so that, in the limit, it converges to an OI method (see also Seaman, 1988).

What is the computational cost of these improvements? For the numerical model and the computer used by Ghil *et al.* (1979), a 24-hr forecast took 40 min of CPU, a 24-hr NoSat assimilation ran in 48 min, a 24-hr time-continuous SCM assimilation ran in 59 min, and a 24-hr time-continuous OI ran in 96 min. As sufficiently powerful computers became available at the turn of the decade, most weather bureaus who could afford them implemented one or another version of OI.

As much more powerful computers are likely to become available during the coming decade, should we be content with OI and with increasing simply the numerical resolution of our forecast and assimilation models? A strong case can certainly be made for the positive impact of increased resolution on both assimilation and forecasting accuracy (Atlas *et al.*, 1982; Hollingsworth *et al.*, 1985). But the rather crude approximation of forecast error covariance evolution in OI has certain deleterious effects on assimilation results. This might require us to use some of the increase in computing power to improve and modify OI in the direction of a better approximation to the K-filter or to replace it by a variational method.

Figure 6 shows the estimated analysis errors [cf. Eqs. (5.3b,c)] of a 6-hr assimilation cycle at the U.S. National Meteorological Center (NMC) (McPherson *et al.*, 1979). The errors in the mass field (upper panel) as well as in the wind field (lower panel) have large inhomogeneities, with local maxima as large as 6°C in temperature and 30 m/sec in zonal velocity. Some, but not all of these maxima occur in regions of data sparseness, and most exhibit strong

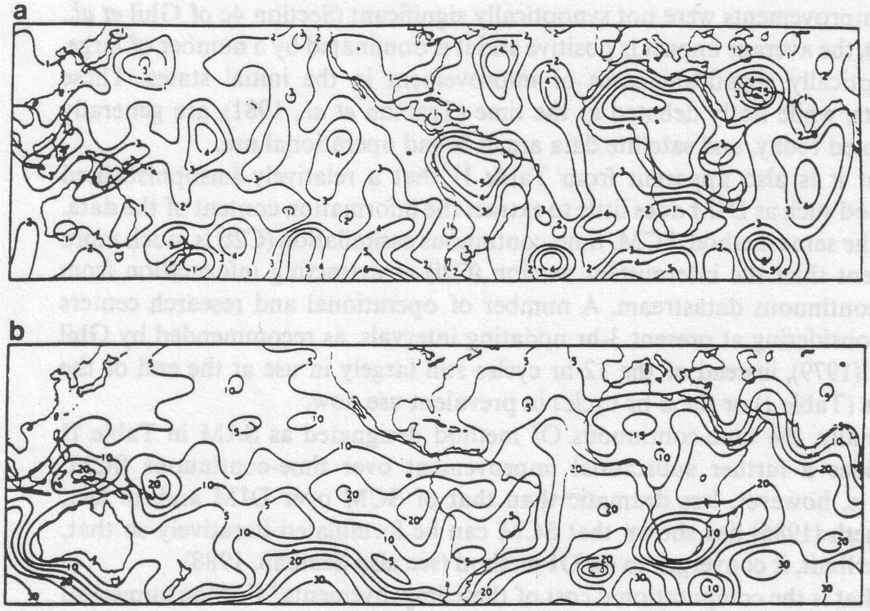


FIG. 6. Estimated analysis error at 250 mb for 0000 GMT 14 December 1977. (a) Temperature error; contour interval is 1°C. (b) Eastward wind component error; contours are 5 ms^{-1} apart (from McPherson *et al.*, 1979).

gradients (see for instance the maxima over South America, North Africa and the Arabian Sea in Fig. 6a and those over South Africa and the Maritime Continent in Fig. 6b). These inhomogeneities and strong gradients in analysis errors contradict the basic assumption of homogeneity in mass field errors made in OI and the derivation of cross correlations involving velocity components without paying attention to these gradients (Cohn and Morone, 1984). In OI, there is no way to estimate reliably and separately forecast errors per se, but a comparison of Eqs. (4.17) with Eqs. (5.1)–(5.5) strongly suggests the existence of inhomogeneities in forecast errors induced by those in analysis errors, and vice versa.

These inconsistencies in OI do not greatly affect its results in regions where data are plentiful and relatively accurate, such as the continents of the Northern Hemisphere. But they do lead to severe problems in the neighborhood of isolated data (see bull's eyes in Figs. 6a,b over islands in the South Atlantic) and at the boundaries between data-dense and data-sparse regions (Cohn *et al.*, 1981; Dee, 1991). Thus, the pursuit of more advanced statistical and variational methods seems certainly justified, and we shall consider these in greater detail in Sections 5.3 and 5.4.

5.2. Initialization Problem

5.2.1. Fast Waves, Initialization, and Projection

Many aspects of synoptic-scale atmospheric and (mesoscale) oceanic motion are well approximated by relatively slow Rossby waves. These are the only type of waves described by the (linearized) QG equations. In NWP, however, PE models have replaced QG models at all major operational centers. (Linearized) PE models also describe relatively fast inertia-gravity waves, which carry a much smaller, but nonvanishing amount of the total energy of the flow.

In terms of describing and predicting the slow, meteorologically and oceanographically significant flow features, such as midlatitude storms or meanders and eddies and rings, the faster waves would seem at first to be more of a nuisance than a help; hence, the inspired use of the QG approximation in early NWP (Charney *et al.*, 1950) and its continued use in theoretical studies of long-term behavior (Pedlosky, 1987; Ghil and Childress, 1987), and also the attempt to justify rigorously the QG approximation by the existence of a slow manifold in the PE system (Leith, 1980; Lorenz, 1980).

In practice, PE models were adopted in NWP because of the need to extend the model domain to the tropics and the entire globe in order to extend the range of validity of the numerical forecasts. This tropical extension required the use of more elaborate nonlinear balance equations in the QG system, leading to loss of ellipticity of the Monge–Ampère equation in question (e.g., Miyakoda, 1956). While a generalized Monge–Ampère equation can be solved efficiently in the mixed-type case (Ghil *et al.*, 1977), the PE system proved much easier to use in an operational setting.

On the theoretical side, it turns out that the slow manifold does not exist in a rigorous mathematical sense (Vautard and Legras, 1986), and that inertia-gravity waves are an inseparable part of the total behavior of the synoptic scales (Errico, 1982; Lacarra and Talagrand, 1988). In oceanography, global or basin-wide PE models are necessary to account correctly for the interaction between the thermohaline and the wind-driven circulation (Bryan and Sarmiento, 1985), and they are the only models available for the description and prediction of tropical phenomena (Gill, 1982).

In the process of data assimilation, NWP experience has shown that the discrepancy between current data, with their random errors, and model first guess, with its errors, can excite a spuriously large amount of inertia-gravity waves in a PE model. These fast waves are damped out over 12–24 hr and have been shown not to affect 24–48 hr forecasts substantially (e.g., Balgovind *et al.*, 1983). However, in an assimilation scheme without proper built-in error estimation, they can lead to a rejection of data at the next subsynoptic update time, being too different from the first guess (see also Daley, 1981, for additional undesirable features of the fast waves).

Therefore, a long-standing approach in NWP has been to eliminate entirely or reduce as much as possible the amount of inertia-gravity waves at initial forecast time. The minimization of the fast-wave energy at initial time goes by the name initialization in NWP. In other disciplines, including sometimes physical oceanography, initialization often means just the assignment of initial values, whatever their properties otherwise, to a forecast field (e.g., Robinson *et al.*, 1987, 1988, 1989). The word is used in its narrow technical NWP meaning throughout this chapter.

The optimal compromise between statistical minimization of the errors in the initial state, on the one hand, and dynamical minimization of the fast components in this state, on the other, is a topic of considerable current interest in NWP, as witnessed by an entire volume of contributions dedicated to it (Williamson, 1982; see also Ghil, 1980). The relevance to oceanographic data assimilation is discussed in Section 6.2.

A reasonable recipe for this compromise can be given in a simple linear shallow-water model (Ghil *et al.*, 1981; Cohn, 1982). In this model, the Rossby waves form a linear subspace, denoted by $\mathscr{R}$ in Fig. 7, and the inertia-gravity waves form a complementary subspace, denoted by $\mathscr{G}$ in the figure.

In the standard formulation of slow manifold theory (Daley, 1980, 1981; Leith, 1980), the two linear subspaces $\mathscr{R}$ and $\mathscr{G}$ are presented as orthogonal to each other. This is only the case if the linear system under study is normal, e.g.,

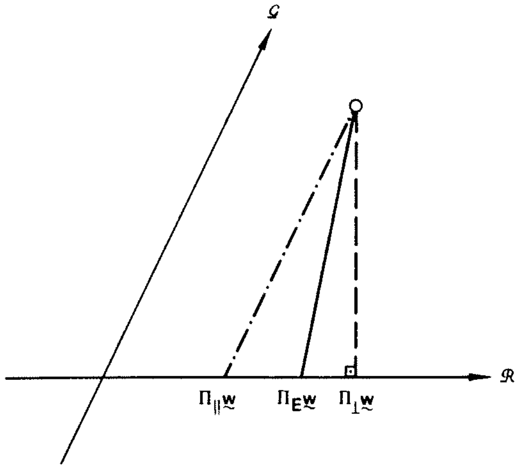


FIG. 7. Schematic representation of the slow subspace $\mathcal{R}$ of Rossby waves and the fast subspace $\mathcal{G}$ of inertia-gravity waves. Three projections onto $\mathcal{R}$ are shown: the parallel projection $\Pi_{\parallel}$ (dash-dotted), the perpendicular projection $\Pi_{\perp}$ (dashed), and the E-perpendicular projection Π_E (solid: from Ghil, 1989).

skew-symmetric, in particular if the full governing equations are linearized about a state of rest. In practice, GFD flows have large shear and linearization about a particular solid-body rotation is not a good approximation for the purposes of data assimilation. Linearization about nonzero mean flow, cf. Eq. (4.19), yields an associated linear operator which is not normal, i.e., does not commute with its adjoint.

As a consequence, projection onto the slow subspace $\mathcal{R}$ of the state $\mathbf{w}^f$ or $\mathbf{w}^a$ can be carried out in more than one way. The parallel projection $\Pi_{\parallel}$ eliminates the fast modes of $\mathbf{w}$ without changing the slow ones. The perpendicular projection operator $\Pi_{\perp}$ minimizes the distance between $\mathbf{w}$ and its projection onto $\mathcal{R}$, $\Pi_{\perp}\mathbf{w}$, in the usual Euclidean metric of the phase space. The oblique or A-perpendicular projection Π_A minimizes this distance in a modified metric, with nonnegative semi-definite weight matrix $\mathbf{A} \geq 0$,

$$J_A \equiv E(\mathbf{w}_k^a - \mathbf{w}_k^t)^T\mathbf{A}(\mathbf{w}_k^a - \mathbf{w}_k^t) \tag{5.6}$$

The simple model to which these distinct projections have been applied is governed by a linearized, spatially 1-D version of the shallow-water equations

$$u_t + Uu_x + \phi_x - fv = 0 \tag{5.7a}$$

$$v_t + Uv_x + fu = 0 \tag{5.7b}$$

$$\phi_t + U\phi_x + \Phi u_x - fUv = 0 \tag{5.7c}$$

The features that make this system worthy of interest, in spite of its great simplicity, are the presence of advection, of the Coriolis acceleration and β-effect, and of two physically distinct types of waves, slow Rossby waves and fast inertia-gravity waves. Non-stationary Rossby waves arise in this constant-f model from the equivalent β-effect due to the $-fUv$ term in the continuity equation (5.7c). The equivalent β is given by $\beta_* \cong f^2U/\Phi$ (Phillips, 1971).

As usual, the coordinate x points eastward, u and v are perturbation velocities eastward and northward, while ϕ is the perturbation geopotential. The parameters are chosen with meteorological midlatitude applications in mind. Thus, the mean zonal velocity is taken to be $U = 20\ \mathrm{m\ sec}^{-1}$, the mean geopotential is $\Phi = 3 \times 10^4\ \mathrm{m}^2\ \mathrm{sec}^{-2}$, and the Coriolis parameter is $f = 10^{-4}\ \mathrm{sec}^{-1}$. The resulting equivalent β_* is $6.7 \times 10^{-12}\ \mathrm{m}^{-1}\ \mathrm{sec}^{-1}$, so that $\beta_* \cong \beta/2$ with β having the usual value at 45° latitude.

The components of the state vector $\mathbf{w}_k$ are the values of (u, v, ϕ) on a space-time grid $(j\Delta x, k\Delta t)$ over which Eqs. (5.7) are discretized by a finite-difference approximation (Ghil *et al.*, 1981). The approximation in question is the Richtmyer two-step version of the Lax–Wendroff scheme, which is second-order accurate in both space and time. The number of points used, $1 \leq j \leq M$, is $M = 16$, so that $N = 3M = 48$. A spatially 2-D version of system (5.7), with $N = 3 \times 60 \times 61 = 10{,}980$ is discussed in the next section.

The time step, chosen close to the Courant–Friedrichs–Lewy stability limit, is $\Delta t = 30$ min. In this simple case, the dynamics matrix Ψ_k is constant in time, $\Psi_k \equiv \Psi$. But the reason for using $U \neq 0$ and the equivalent β-term in the first place is the desire to build towards a satisfactory solution of the data assimilation problem for nonlinear models. The EKF and its adaption to GFD problems requires successive linearizations about realistic flows (Ghil *et al.*, 1981, 1982) [cf. Eqs. (4.18, 4.19) here and the accompanying discussion]. It was shown (Budgell, 1986b; Lacarra and Talagrand, 1988) that the estimation can still proceed quite successfully in this more general and realistic case (see also Sections 4.1 and 5.3.2).

Details about the linear subspaces $\mathscr{R}$ and $\mathscr{G}$ in the continuous system (5.7), as well as in the actual discrete system used in the numerical examples, can be found in Cohn (1982). The different projections are written down explicitly there as matrix operators for the discrete system. The projection used in the following numerical example has some physical justification, being the minimum-energy projection, or E-perpendicular projection, which minimizes the expected energy of the analysis error, $\Pi_{\mathbf{E}}$. In this special case, the weight matrix $\mathbf{A}$ will be denoted by $\mathbf{E}$; it is positive definite, diagonal, and the diagonal entries are, at each grid point, unity for the velocity components u and v and $1/\Phi$ for the geopotential ϕ.

With these dynamical facts in mind, we can address the issue of the compromise between minimum errors and minimum fast waves by modifying the standard K-filter $\mathbf{K}_k^*$. The modified filter minimizes the error functional in Eq. (5.6) subject to the constraint that

$$\mathbf{w}_k^{\mathrm{a}} \in \mathscr{R} \tag{5.8}$$

at all update times k. It is assumed that $\mathbf{w}_0^{\mathrm{a}} \in \mathscr{R}$, i.e., that initialization has been performed at the outset.

The solution of this constrained minimization problem (Cohn, 1982; Ghil *et al.*, 1982) is to take for the gain matrix

$$\mathbf{K}_k = \mathbf{K}_k^{\Pi} \equiv \Pi_{\mathbf{A}} \mathbf{K}_k^* \tag{5.9}$$

where $\Pi_{\mathbf{A}}$ is the A-orthogonal projection matrix onto $\mathscr{R}$, defined by

$$\operatorname{Range} \Pi = \mathscr{R} \tag{5.10a}$$

$$\Pi^2 = \Pi \tag{5.10b}$$

$$(\mathbf{A}\Pi)^{\mathrm{T}} = \mathbf{A}\Pi \tag{5.10c}$$

The (dynamically) modified K-filter, or ΠK-filter, is the data assimilation scheme based on the choice of gain matrix $\mathbf{K}_k^{\Pi}$.

For any given choice of A, the ΠK-filter also has the property that it minimizes the functional

$$\bar{J}_{\mathbf{A}} \equiv E(\mathbf{w}_k^{\mathrm{a}} - \bar{\mathbf{w}}_k^{\mathrm{a}})^{\mathrm{T}} \mathbf{A} (\mathbf{w}_k^{\mathrm{a}} - \bar{\mathbf{w}}_k^{\mathrm{a}}) \tag{5.11}$$

subject to the constraint (5.8), where $\bar{\mathbf{w}}_k^{\mathrm{a}}$ denotes the analyzed field that would be produced by using the standard K-filter at time k. In fact, we have

$$\mathbf{w}_k^{\mathrm{a}} = \Pi \bar{\mathbf{w}}_k^{\mathrm{a}} \tag{5.12}$$

Thus, the ΠK-filter combines the standard K-filter with variational normal mode initialization (Daley, 1981; Tribbia, 1982), i.e., with variational projection onto $\mathscr{R}$; $\bar{\mathbf{w}}_k^{\mathrm{a}}$ is an objective analysis, $\mathbf{w}_k^{\mathrm{a}}$ is the initialized field, and the elements of A are the variational weights. The ΠK-filter, though, minimizes not only the A-distance of Eq. (5.11) between the final initialized field $\mathbf{w}_k^{\mathrm{a}}$ and the analyzed field $\bar{\mathbf{w}}_k^{\mathrm{a}}$, but also the A-distance of Eq. (5.6) between $\mathbf{w}_k^{\mathrm{a}}$ and the true field $\mathbf{w}_k^{\mathrm{t}}$, which is a measure of the actual analysis error.

When using the standard K-filter, the estimated state at any grid point (Fig. 7 in Ghil, 1989, not shown here) is given by the superposition of small-amplitude, rapidly evolving inertia-gravity waves upon large-amplitude, slowly-evolving Rossby waves. The former are excited by the system noise, $Q \neq 0$, at every time step and by the discrepancy between estimated state and noisy observations at synoptic times.

When using the ΠK-filter instead, the evolution of the Rossby waves is the same as before, while the fast waves are completely eliminated (Fig. 8 of Ghil, 1989, not shown here). In particular, fast waves are no longer excited at update time, even when the analysis $\mathbf{w}_k^a$ differs markedly from the first guess $\mathbf{w}_k^f$. Changing the type of projection to $\Pi_\|$ or $\Pi_\perp$ does not seem to make too much of a difference in the estimate (Ghil, 1989).

At what cost to the estimation error are the fast waves eliminated? It is obvious that constrained optimization, Eqs. (5.6, 5.8), can only yield a minimum larger than or equal to the result of unconstrained optimization [Eq. (5.6)]. In Fig. 8, we see side by side the expected rms errors for the K-filter and ΠK-filter.

The excess estimation error of the ΠK-filter over the K-filter, for all the components of the energy as well as for the total, increases with time in the assimilation cycle, but is still quite small in the asymptotic regime at day 10. So the loss of accuracy in estimation is not too great. But what is the gain?

As pointed out earlier, inertia-gravity waves are an inseparable part of the geofluid's behavior. They are essential in tropical phenomena, and in fact their suppression in operational NWP practice by nonlinear normal-mode initialization (Daley, 1981) has led to serious estimation errors in tropical analyses (Kanamitsu, 1981). The correct amount of fast-wave energy could

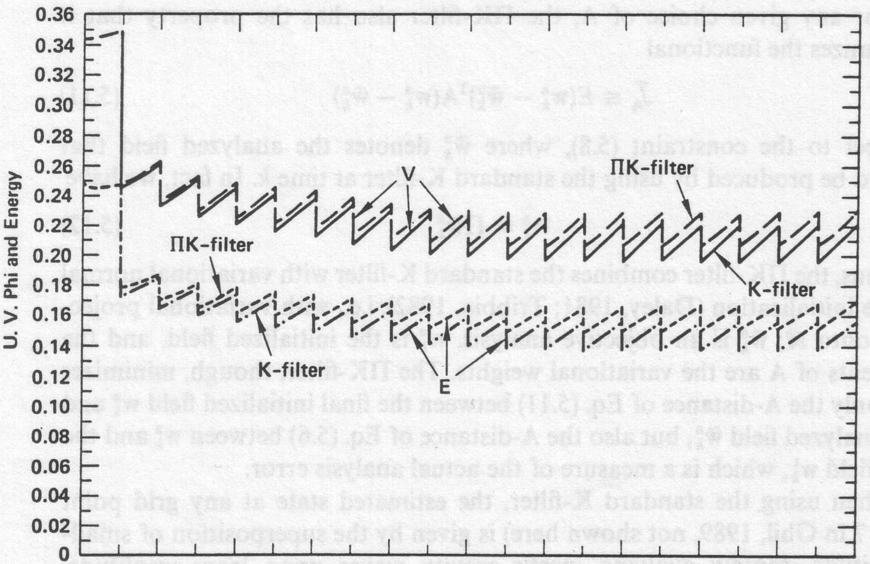


FIG. 8. Evolution of the estimation error for a conventional network with and without initialization of the K-filter. Only the expected rms error for v (solid for the standard K-filter, dashed for the ΠK-filter) and for the total energy E (dashed-dotted for **K**, short dashes for ΠK) is shown (after Ghil *et al.*, 1981).

be determined from the observations by using optimal or nearly optimal filters as suggested by the preliminary results of Dee *et al.* (1985). But large errors in the fast waves are very harmful to the correct estimation of the energetic slow waves in an assimilation scheme that is far from truly optimal, such as OI. Thus, initialization, albeit easy, is neither necessary nor particulary useful when a nearly optimal data assimilation scheme is implemented, but it is very helpful as an improvement to the highly suboptimal assimilation schemes in current NWP use (Cohn *et al.*, 1981). We turn therefore to a description of the initialization scheme in widest operational use at present, nonlinear normal-mode initialization. While of less interest for oceanographic data assimilation (see Sections 3.1 and 6.2), it turns out to provide considerable theoretical insight into the reasons why Rossby waves dominate in fluid systems governed by the full nonlinear primitive equations.

5.2.2. *Nonlinear Normal-Mode Initialization*

The development of nonlinear normal-mode initialization (NNMI) (Baer and Tribbia, 1977; Machenhauer, 1977) was strongly motivated by the practical desire to remove the spurious inertia-gravity waves generated by initialization shocks in PE models for NWP. Its actual application to the models of most leading NWP centers in the world today encountered substantial difficulties of three kinds: (i) the lack of time-scale separation between the relatively slow internal inertia-gravity waves with small equivalent depth, on the one hand, and Rossby waves, on the other; (ii) the close connection, throughout the atmosphere, between vertical velocities and hence precipitation, on the one hand, and horizontally divergent motions projecting significantly onto inertia-gravity modes; and (iii) the dominance of divergent motions in the tropics, with its ascending branch of the Hadley cell and massive latent-heat release. These three problems are clearly related to each other and have generated a substantial literature attempting to cope with the effect of cloud processes, diabatic heating, and divergent motions on the research and operational practice of NNMI (Donner, 1988; Kitade, 1983; Krishnamurti *et al.*, 1988; Rasch, 1985).

As seen earlier in this section, initialization might no longer be necessary in order to prevent data rejection once a more advanced form of data assimilation, i.e., one closer to true optimality than OI, has been implemented. The practical issues of initialization are also less critical in oceanography, cf. Section 6.2. On the other hand, NNMI has provided considerable theoretical insight into the structure of the nonlinear primitive equations, and helped researchers understand and justify at a deeper level the QG approximation. It is from this dynamical perspective that NNMI is reviewed here. The presentation follows Leith (1980) and Tribbia (1979).

Consider the discretization, spectrally or by finite differences, of a typical system of flow equations

$$\dot{\mathbf{w}} + iL\mathbf{w} = \varepsilon M(\mathbf{w}, \mathbf{w}) \tag{5.13a}$$

Here the generic nonlinear operator N of Eq. (4.23) has been decomposed explicitly into a linear hermitian part $L^* = L$ and a quadratic part M,

$$\mathrm{N}(\mathbf{w}) \equiv -iL\mathbf{w} + \varepsilon M(\mathbf{w}, \mathbf{w}) \tag{5.13b}$$

L is assumed to arise by linearization about a state of rest, hence the characteristic skew-hermitian character of iL, which is associated with classical tidal operators. M is quadratic and represents advective nonlinearities. Other nonlinearities arising from physical processes, such as convection and its interaction with radiation, have been mentioned before and are beyond the treatment of NNMI given here. The small parameter ε is typically a Rossby number; ε^{-1} measures the (nondimensional) time scale over which nonlinear effects are significant.

L has n real eigenvalues that are the frequencies of the system and are assumed to fall into two distinct ranges such that, without loss of generality,

$$0 < \sigma_1 \leq \cdots \leq \sigma_{k_0} \ll \sigma_{k_0+1} \leq \cdots \leq \sigma_n \tag{5.14}$$

$\sigma_k = 0(1)$ for $k_0 + 1 \leq k \leq n$ and $\sigma_k = 0(\varepsilon)$ for $1 \leq k \leq k_0$; typically $n \cong 3k_0$ in PE systems. The eigenvectors associated with the small frequencies span a slow subspace $\mathscr{R}$; the others span a fast subspace $\mathscr{G}$. Due to the linearization about a state of rest and the resulting skew symmetry of L, all eigenvectors are mutually orthogonal, and so are $\mathscr{G}$ and $\mathscr{R}$, in contradistinction from Fig. 6.

Changing by an orthogonal transformation from the arbitrary basis of Eqs. (4.23) or (5.13) to that of the eigenvectors of L yields a system

$$\dot{\mathbf{x}} + i\varepsilon\Lambda_x\mathbf{x} = \varepsilon N_x(\mathbf{z}, \mathbf{z}) \tag{5.15a}$$

$$\dot{\mathbf{y}} + i\Lambda_y\mathbf{y} = \varepsilon N_y(\mathbf{z}, \mathbf{z}) \tag{5.15b}$$

Here the matrices Λ_x and Λ_y are diagonal, having respectively the k_0 small frequencies (rescaled explicitly by ε) and $n - k_0$ large frequencies on the diagonal. The vector $\mathbf{z} = (\mathbf{x}^{\mathrm{T}}, \mathbf{y}^{\mathrm{T}})^{\mathrm{T}}$ is the vector $\mathbf{w}$ expressed in the new basis.

For $\varepsilon = 0$, i.e., no nonlinearity and complete scale separation, a solution starting from initial data $\mathbf{y}(0) = \mathbf{0}$, $\mathbf{x}(0) \equiv \mathbf{x}_0 \neq \mathbf{0}$ would stay forever in $\mathscr{R}$. For $0 < \varepsilon \ll 1$, this is not the case: nonlinear interactions between $\mathbf{x}$ and $\mathbf{y}$ will lead over time $0(\varepsilon^{-1})$ to significant fast components. Thus we wish to determine, for any given $\mathbf{x}_0$, a $\mathbf{y}(0) \equiv \mathbf{y}_0 \neq 0$ such that no high-frequency oscillations arise over time $0(\varepsilon^{-1})$.

Slightly different solutions to this problem were proposed by Baer (1977), who applied his procedure to a nonlinear version of the 1-D shallow-water model of Eqs. (5.7), and by Machenhauer (1977), who applied his to a nonlinear shallow-water model on the sphere. Both procedures are iterative and agree to first order in ε. Machenhauer's procedure is simpler and has therefore been applied more widely in operational practice. Baer's, as further clarified and elaborated by Baer and Tribbia (1977) and by Tribbia (1979), is more consistent and more systematic, involving an asymptotic expansion in ε.

The nonlinear balance condition of Machenhauer (1977) is that $\dot{\mathbf{y}} = 0$. This condition yields, from Eq. (5.15b), the implicit nonlinear equation for $\dot{\mathbf{y}}$:

$$\mathbf{y}_0 = -i\varepsilon\Lambda_y^{-1}N_y(\mathbf{z}_0, \mathbf{z}_0) \tag{5.16}$$

where $\mathbf{z}_0^{\mathrm{T}} = (\mathbf{x}_0^{\mathrm{T}}, \mathbf{y}_0^{\mathrm{T}})$. Eq. (5.16) can now be solved iteratively, starting with the natural first guess $\mathbf{y}_0^{(0)} = \mathbf{0}$.

Baer and Tribbia (1977) carried out a two-time scale expansion of Eq. (5.15a,b), with fast time $t^* \equiv t$ and slow time $\tau \equiv \varepsilon t$, and required that the solution be free of fast motion, i.e., of t^* derivatives. They showed that, to first order in ε, this will happen for $\mathbf{y}_0^{(1)}$ given by the first Machenhauer iteration of Eq. (5.16) with the natural first guess $\mathbf{y}_0^{(0)}$, $\mathbf{y}_0^{(1)} = -i\varepsilon\Lambda_y^{-1}N_v(\mathbf{z}_0^{(0)}, \mathbf{z}_0^{(0)})$.

Even assuming that the Machenhauer iteration converges, $\mathbf{y}_0^{(m)} \rightarrow \mathbf{y}_0^{(\infty)}$, to a solution of Eq. (5.16), this does not in fact guarantee that a solution of Eq. (5.15) with initial data $\mathbf{z}(0) = \mathbf{z}_0^{(\infty)}(\mathbf{x}_0)$ will stay free of fast motions for $t > 0$. To second order in ε, Baer and Tribbia (1977) showed that the initial data for the $\mathscr{G}$-component should satisfy

$$\mathbf{y}_0 = -i\varepsilon\Lambda_y^{-1}[N_y(\mathbf{z}_0^{(0)} + \mathbf{z}_0^{(1)}, \mathbf{z}_0^{(0)} + \mathbf{z}_0^{(1)}) + i\varepsilon\Lambda_y^{-1}N_y(\zeta, \zeta) + \mathbf{y}_0^{(1)}] \tag{5.17a}$$

Here $\mathbf{z}_0^{(0)}$ has $\mathscr{R}$-component $\mathbf{x}_0$, arbitrary and prescribed, and $\mathscr{G}$-component $\mathbf{y}_0^{(0)} = \mathbf{0}$, while $\mathbf{z}_0^{(1)}$ has components $\mathbf{0}$ and $\mathbf{y}_0^{(1)}$, respectively; furthermore, ζ has a null fast component and a slow component

$$\zeta = \mathbf{x}_0 + \varepsilon[-i\Lambda_x\mathbf{x}_0 + N_x(\mathbf{z}_0, \mathbf{z}_0)] \tag{5.17b}$$

Leith (1980) has shown that, to the first order to which the two procedures agree, it is possible to define, for relatively simple nonlinear PE models, a slow manifold $\mathscr{M}_s$ on which the dynamics is quasi-geostrophic in the following sense: (i) $\mathscr{M}_s$ is tangent at the origin to $\mathscr{R}$ and hence perpendicular there to $\mathscr{G}$; (ii) given a point in $\mathscr{R}$ away from the origin diagnostic equations can be found to determine the corresponding point on $\mathscr{M}_s$; and (iii) the motion of points within $\mathscr{M}_s$ is governed by QG equations. These results, however, can only be expected to hold in an ε neighborhood of the origin in phase-parameter space, i.e., for nearly-geostrophic motions deviating slightly from a state of rest. It is not clear whether a global manifold invariant under the PE equations for large-amplitude motions with substantial ageostrophic components exists and is totally free of fast motions (Lorenz, 1986; Vautard and Legras, 1986). For a review of the theory and applications of nonlinear normal mode initialization, see Errico (1989) and Daley (1991, chapters 6, 9, and 10).

5.3. Kalman Filtering Applications

As indicated in Section 4.1, two major issues in the practical application of the K-filter to GFD problems are its computational cost and its reliable extension to nonlinear systems. In this subsection, we show ways to address these two issues.

5.3.1. Efficient Implementation

The key feature of the K-filter [Eqs. (4.17a–e)] is its estimating optimally the state of a dynamical system as well as the error in this estimate due to observational and system errors. The computation of the forecast and analysis error covariance $P_k^{f,a}$ requires $O(N^2)$ operations for a state vector with N variables, in the absence of any simplifications. Even with the expected rapid progress in computing speed and memory devices, given $N = 0(10^5 - 10^6)$ for state-of-the-art meteorological and oceanographic models and $0(10^7)$ in the near future, the full implementation of such an algorithm would be out of the question for the 1990s.

The computational burden in the forecast step [Eq. (4.17a,b)] can be reduced in a number of ways, all of them involving simplifications in the algorithm along with some loss of optimality. The trade off is between computational cost and degree of optimality. To evaluate the best possible trade off, it is still necessary to have, at least for development purposes, a full implementation of the K-filter for a smaller size test problem in order to evaluate the performance of the proposed suboptimal algorithm (e.g., Cohn *et al.*, 1981; Ghil *et al.*, 1982; Section 5.1 here).

The approaches proposed or under study include various improvements to OI, such as taking into account the inhomogeneity of forecast errors (Cohn and Morone, 1984) or advecting mass-field error variances by OI-estimated winds (J. Pfaendtner, personal communication, 1990). Another approach would simply use a lower resolution model for the assimilation than for the forecasting, possibly tuning some of the coefficients of the lower resolution model to match as closely as possible its forecast fields to those of the higher resolution one. Still another approach is to assume a certain spectral distribution of model errors, e.g., restricted entirely to slow modes (cf. Phillips, 1986) and with the energy decreasing with wave number among the latter (cf. Balgovind *et al.*, 1983; Bennett and Budgell, 1987). Cohn and Parrish (1991), making such simple assumptions, obtained for a linear barotropic 2-D shallow-water model much better results with the K-filter than with OI, given a number of realistic as well as hypothetical data distributions. Yet another interesting simplification of the K-filter is that of Dee (1991), who chooses to advect the covariance matrix of height-field errors only, with the height-wind and wind-wind errors computed as in OI by Eqs. (5.2b–d). Clearly this is an active field of research and all these ideas need to be tested in more and more realistic settings with regard to model complexity and data distribution.

We present one approach here in further detail to give a better feel for the computational issues involved. In this approach, the computational complexity of the forecast step can be reduced from $0(N^2)$ *to* $0(N)$ by exploiting special features of the dynamics matrix Ψ_k and covariance matrix P_k, which arise in the GFD applications of interest here. The crucial feature is that both Ψ_k and P_k are sparse, rather than full matrices. In any finite-difference explicit form of the governing evolution equations, values of each variable at a given grid point and time step depend only on the variables at a few adjacent points and at the preceding time. Hence Ψ_k can be written as a banded matrix or as a block-banded matrix with banded blocks. The band width is given by the finite-difference stencil, i.e., by the relative position of adjacent grid points involved, whose number is typically 4–8 for 2-D problems and second-order accurate schemes.

The covariance matrix P_k is also banded or block banded to very good approximation, since forecast and analysis error correlations decay to zero over a distance comparable to the Rossby radius of deformation. For the large-scale and synoptic-scale problems to which discussion is restricted in the present review, a few grid points per Rossby radius are sufficient to resolve the motions of interest. Hence, in fact, the bandwidth of nonnegligible entries for P_k is comparable to that for Ψ_k.

The computational device for exploiting this bandedness is to actually write and store Ψ_k as a matrix, rather than as a finite-difference scheme, and to use an algorithm for multiplication by diagonals (Madsen *et al.*, 1976; Parrish and Cohn, 1985) on a vector processor with parallelism in calculating Eq. (4.17b), which is rewritten for this purpose as

$$P^{\mathrm{f}}_{k+1} = \Psi_k(\Psi_k P^{\mathrm{a}}_k)^{\mathrm{T}} + Q_k \qquad (5.18)$$

Equation (5.18) makes use of the symmetry of P_k and requires only programming the multiplication by Ψ_k from the left, which occurs twice.

Substantial computational savings in the analysis step [Eq. (4.17c,d,e)] can be obtained by avoiding the $p \times p$ matrix inversion of Eq. (4.17c) through the use of a device called sequential processing of observations (Ho, 1963; Gelb, 1974, pp. 304–305; Bierman, 1977; Rodgers, 1977; Parrish and Cohn, 1985). The idea is to view multiple observations, even when they arrive at the same update time, as a string of individual observations separated by zero-time intervals. Thus, Eqs. (4.17a,b) are not used to advance the state $\mathbf{w}_k$ and covariance $\mathbf{P}_k^{\mathrm{f}}$ from one pseudo-time step to the next, while Eqs. (4.17c,d,e) yield, by using the Woodbury formula [(cf. also Eq. (4.16a)],

$$(P_k^{\mathrm{a}})^{-1} = (P_k^{\mathrm{f}})^{-1} + H_k^{\mathrm{T}} R_k^{-1} H_k \tag{5.19}$$

The second term on the right-hand side of Eq. (5.19) can be decomposed further into smaller batches of observations or to individual observations.

In fact, this sequential processing could also be applied to reduce the computational cost of the OI analysis step [Eqs. (5.3b,c) and (5.4)] by replacing P_k with S_k in Eq. (5.19). Sequential processing is ideally suited conceptually to the unified treatment of time-continuous remote-sensing data, on the one hand, and of synchronous synoptic conventional data on the other. The only hitch is that, since OI does not provide a reliable estimate of the analysis error, separate empirical procedures for quality control, such as buddy checks within batches of data had to be developed in NWP operations [see discussion following Eq. (5.5), and references there].

The algorithmic simplifications to the forecast step and the analysis step outlined previously, i.e., multiplication by diagonals of Ψ and P and sequential processing of observations, have been applied to a two-layer shallow-water model in a 2-D domain:

$$\partial \mathbf{V}_k/\partial t = -(\mathbf{V}_k \cdot \nabla)\mathbf{V}_k - f\mathbf{k} \times \mathbf{V}_k - \nabla[\alpha_k \phi_1 + \phi_2] \tag{5.20a}$$

$$\partial \phi_k/\partial t = -\nabla \cdot (\phi_k \mathbf{V}_k) \tag{5.20b}$$

$k = 1, 2$ are the upper and lower layers, respectively; $\mathbf{V}_k = (u, v)$ is the velocity vector, ϕ_k is the geopotential, $f = f_0 + \beta y$ is the Coriolis parameter and the αs are constants, $\alpha_1 = 1, \alpha_2 = \rho_1/\rho_2$, where ρ is density. The implementation of Parrish and Cohn (1985) was for a one-layer barotropic version of Eq. (5.20) in a 6000 km × 6000 km square domain, extending approximately between 15°N and 75°N, with free-slip conditions at the northern and southern boundaries and periodicity in the zonal direction. The equations were linearized about a state with constant zonal velocity $U_0 = 20\ \mathrm{ms}^{-1}$.

These authors carried out computations with resolutions of 20 × 21, 40 × 41, and 60 × 61 grid points on a Cyber 205 vector processor. The latter resolution of 100 km is quite comparable with that of state-of-the-art global and even regional NWP models. Experiments with different bandwidths b for P_k^{f} were carried out: the total number of nonzero entries in any row or column is $2b + 1$. Table III shows the results of these experiments. It is clear that the computation is feasible, and that the efficiency of the algorithm increases with increasing resolution.

Parrish and Cohn (1985) showed that in the absence of model errors ($Q \equiv 0$), it suffices to have observations of velocity and geopotential along a single line of grid points every 12 hr to reduce the analysis error to a level

TABLE III. CPU SECONDS REQUIRED BY THE FORECAST STEP OF THE K-FILTER, EQS. (4.17a,b), IN A 2-D SHALLOW-WATER MODEL AS A FUNCTION OF BANDWIDTH AND MODEL RESOLUTION[a]

	Resolution		
Bandwidth b	20×21	40×41	60×61
1	0.11 (93)	0.27 (154)	0.52 (178)
3	0.48 (94)	1.10 (153)	2.15 (176)
5	1.05 (93)	3.51 (152)	
7	1.84 (93)		
full	3.24 (89)		

[a] Numbers in parentheses are the observed megaflop (MFLOP) rates for each computation. Computations carried out on a CYBER 205 with a peak rate of 200 MFLOPS (from Parrish and Cohn, 1985).

below that of the observational error, diag(R), over most of the domain for all variables in a few days. In the presence of model errors ($Q \neq 0$), the reduction of initial error is still dramatic, and the asymptotic error level of the analysis, achieved in a few days, agrees in its dependence on Q and R with the theory of Ghil *et al.* (1981). In both cases, with and without model errors, analysis errors are symmetric about the midchannel latitude of 45°N for u and v, while they are larger towards higher latitudes in ϕ.

In the more realistic case of $Q \neq 0$, it was shown that a cross-stream meridional measurement line is substantially better than an along-stream zonal line of observations in reducing analysis error (compare Malanotte-Rizzoli and Holland, 1986, 1988, and Section 6.1.1 here). For a meridional line, errors are smallest in the southeastern quadrant of the domain for geopotential (Fig. 9) and in the eastern half for wind, suggesting that the β-effect affects differently the propagation of ϕ-information than of V-information. Comparison of panels (a), (b), and (c) of Fig. 9 shows, and other results confirm, that the banded approximation does not affect greatly the analysis error in either magnitude or spatial distribution. Thus, computational feasibility is achieved at no great detriment to the K-filter's performance. On the other hand, the $\phi-\phi$, $\phi-$V and V$-$V correlations (not shown here) differ markedly from those assumed by OI [Eqs. (5.1, 5.2)], by being neither homogeneous nor symmetric (see also Cohn *et al.*, 1981, and Ghil *et al.*, 1982, for the 1-D case, and Balgovind *et al.*, 1983, for the 3-D, semi-operational case).

Todling and Ghil (1990) implemented the K-filter for a two-layer version of Eq. (5.20) linearized about a zonal flow with strong shear (Fig. 10a) for which front-like patterns (Fig. 10b) develop as a consequence of baroclinic

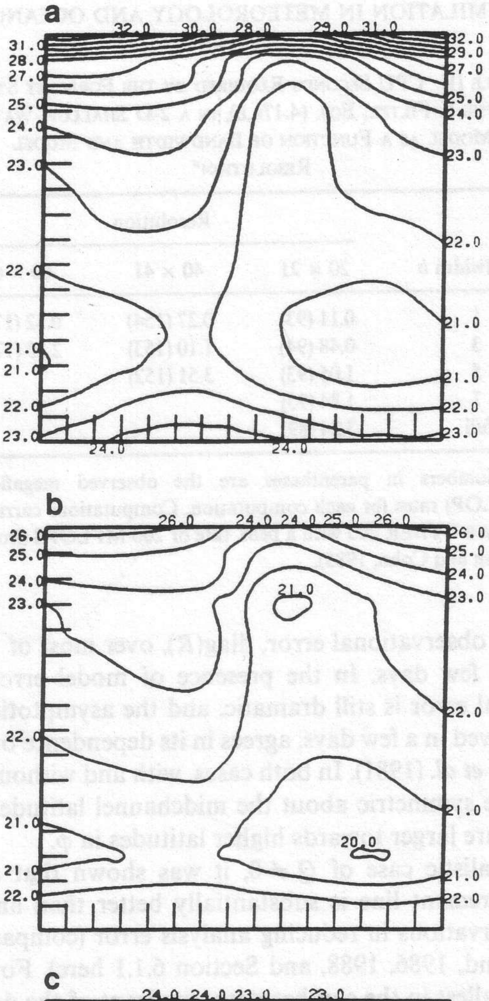


FIG. 9. Forecast error standard deviations in the height field at 10 days for a K-filter experiment with observations along the N–S symmetry axis of a periodic β-channel, perpendicular to the basic zonal flow. (a) Full error covariance matrix used in forecast step; (b) banded approximation of P_k^f, with $b = 5$; (c) bandwidth $b = 3$ (from Parrish and Cohn, 1985).

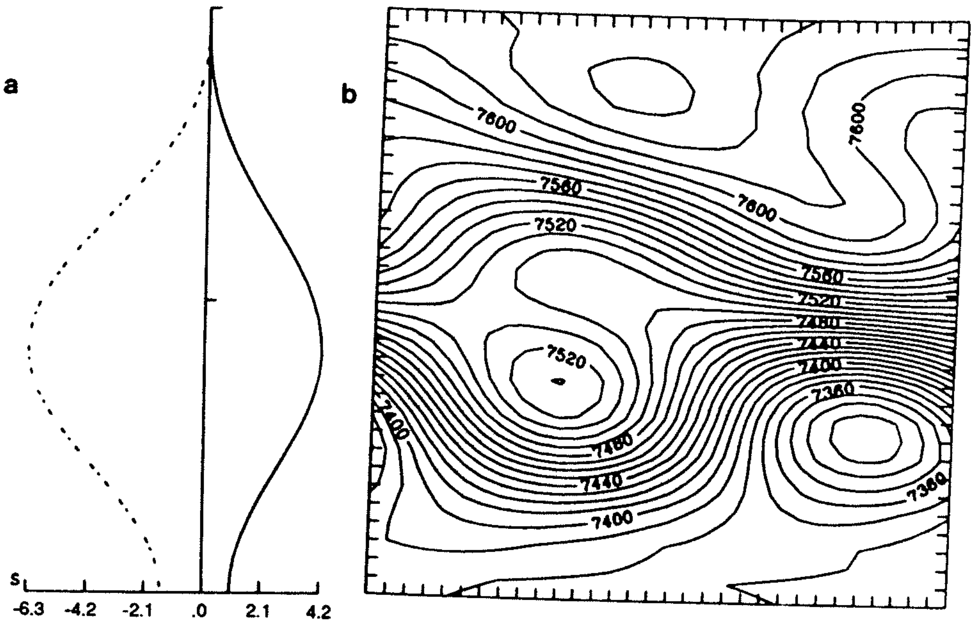


FIG. 10. K-filter experiments for a two-layer shallow-water model in a β-channel. (a) Velocity profile of the basic zonal flow in lower layer (solid) and upper layer (dashed). (b) Surface pressure at 10 days for a density ratio between layers $\alpha_2 = \rho_1/\rho_2 = 0.9$ (after Todling and Ghil, 1990).

instability. Miller (1986) had already shown the ability of the K-filter to track intentionally induced numerical instabilities in the barotropic vorticity equation. The study illustrated in Fig. 10 concentrates on the observability of the front-like patterns in the presence of various error levels, at resolutions from 16 × 17 through 64 × 65, and comparison with OI performance for the same phenomena and parameters.

The fully nonlinear version of Eq. (5.20) on the sphere, with topography, was used by Keppenne (1989) in a study of low-frequency atmospheric variability (Ghil and Childress, 1987, Ch. 6). He examined in detail the dependence of model solutions on viscosity, dissipation, and zonal-jet forcing for spectral resolutions as high as T15, which corresponds to about 1500 real scalar variables. D. Boggs and C. Keppenne (personal communication, 1990) are implementing the K-filter's banded approximation for a higher-resolution, finite-difference version of this model and will carry out observing systems simulation experiments (OSSE), real-data assimilation, and comparisons with OI. The thrust of this work is to remove the loss of positive definiteness of $P_k^{f,a}$ due to the banded approximation, as noted by Parrish and Cohn (1985). This loss is a well-known computational problem for covariance matrices in general (e.g., Kerr, 1990) and can be corrected by the use of stabler square-root filters (Bierman, 1977).

5.3.2. Strong Nonlinearity

No closed-form solutions to the estimation problem for stochastically perturbed, nonlinear systems exist, at least not in any form that is computationally realizable for large systems. Hence, there are many approaches to obtaining approximate, more-or-less suboptimal solutions. One of these is to renounce the physically realistic assumption of an imperfect model and replace it with that of a perfect model, i.e., reduce the stochastic minimum-variance estimation problem to a deterministic least-squares problem (Gelb, 1974, Section 6.3). A computationally efficient implementation of the latter is the adjoint method (see Sections 4.2, 5.4.3, and 6.3.4).

The quadratic advective nonlinearities of GFD, albeit small, are well known to have important consequences for long-time behavior (Lorenz, 1963; Ghil and Childress, 1987; Pedlosky, 1987). But in data assimilation, it is only the short-term behavior that counts, and neither these advective nonlinearities, proportional to the small Rossby number, nor other nonquadratic nonlinearities, associated with small-scale thermodynamic processes, affect greatly the short-term behavior.

Lacarra and Talagrand (1988) studied in detail the contribution of linear and nonlinear terms to flow evolution in an f-plane barotropic shallow-water model, as a function of wave number. They showed that for initial perturbations in total energy per unit mass not exceeding 100 m^2 sec^{-2}, the linear terms dominate error growth up to 24 hr, more so in the large-scale Rossby modes than in the gravity waves and the shorter scales. Their results are in agreement with those of Daley (1980), for a nonlinear shallow-water model on the sphere, and of Balgovind *et al.* (1983) for a semi-operational NWP model. Lacarra and Talagrand showed further that a constant-coefficient approximation of linearizations about an arbitrary state reproduces rather faithfully the behavior of the fastest-growing, large-scale waves for up to 48 hr.

These results confirm that a promising approach to nonlinear estimation in GFD is the extended Kalman filter (EKF), which proceeds by successive linearizations of the flow equations about the current estimate of the flow field [cf. Eqs. (4.18) and (4.19)]. In most engineering applications, linearizations are performed at every update time. The theoretical results just discussed and practical experience with OI imply that updates of the linearization [Eq. (4.19)] should only be necessary every 12–24 hr in meteorology and at increasingly larger intervals in midlatitude and tropical oceanography.

This still leaves the question of whether the EKF can track successfully the flow when its evolution is not smooth, but shifting from one type of behavior to another, e.g., from zonal to blocked flow in the atmosphere (Charney and

DeVore, 1979; Ghil and Childress, 1987, Ch. 6) or from a small meander to a large meander state in the Kuroshio (Taft, 1978; Chao, 1984; Miller and Ghil, 1990). To illustrate the issues which arise in tracking such transitions between different attractor basins, we give a prototypical minimal example. Stochastically perturbed motion in a double-well potential has been used as an example of the interaction between nonlinearity and random perturbations in long-term climate theory by Sutera (1981), and a careful extension of such ideas to the Charney–DeVore model has been given by De Swart and Grasman (1987, and references therein).

In its simplest form, the example can be written as

$$\dot{x} = f(x) + \sigma\eta(t) \tag{5.21a}$$

where $\eta(t)$ is Gaussian white noise of unit variance, so that σ^2 is the variance of the random forcing, and

$$f(x) = -V'(x), \qquad V(x) = x^2(x^2 - 2) \tag{5.21b,c}$$

Here V is the potential, having the minima of its two wells at $x = \pm 1$, and a (local) maximum at $x = 0$, with $V \to +\infty$ as $x \to \pm\infty$. The two minima are stable equilibria of Eq. (5.21), and the origin is an unstable equilibrium (see also Ghil and Childress, 1987, Section 10.3). The random forcing pushes the point $x(t)$ away from the stable equilibria, and a succession of pushes in the same direction will effect a transition from the left into the right well, or vice versa.

Miller and Ghil (1990) implemented the EKF for (Eq. 5.21). The results are shown in Fig. 11. It is clear that the estimated position $x_k^{f,a}$ will follow the true position x_k^t from one well into the other, provided the observations are accurate enough or frequent enough or both. Miller and Ghil are also carrying out EKF studies for the Lorenz model and for a finite-element barotropic model of the Kuroshio, similar to that of Chao (1984).

5.4. Applications of Variational Methods

5.4.1. Duality

As mentioned already at the end of Section 2, the beginning of Section 4, Eqs. (4.5–4.7), and many times since, variational methods and sequential estimation methods have strong connections to each other. The basis of these connections is the duality principle established by Kalman (1960) between deterministic control and stochastic estimation. Using the continuous-time notation introduced in Section 4.2, a linear deterministic control problem can be written as

$$\dot{\mathbf{w}} = \tilde{F}(t)\mathbf{w} + \tilde{H}(t)\mathbf{u} \tag{5.22}$$

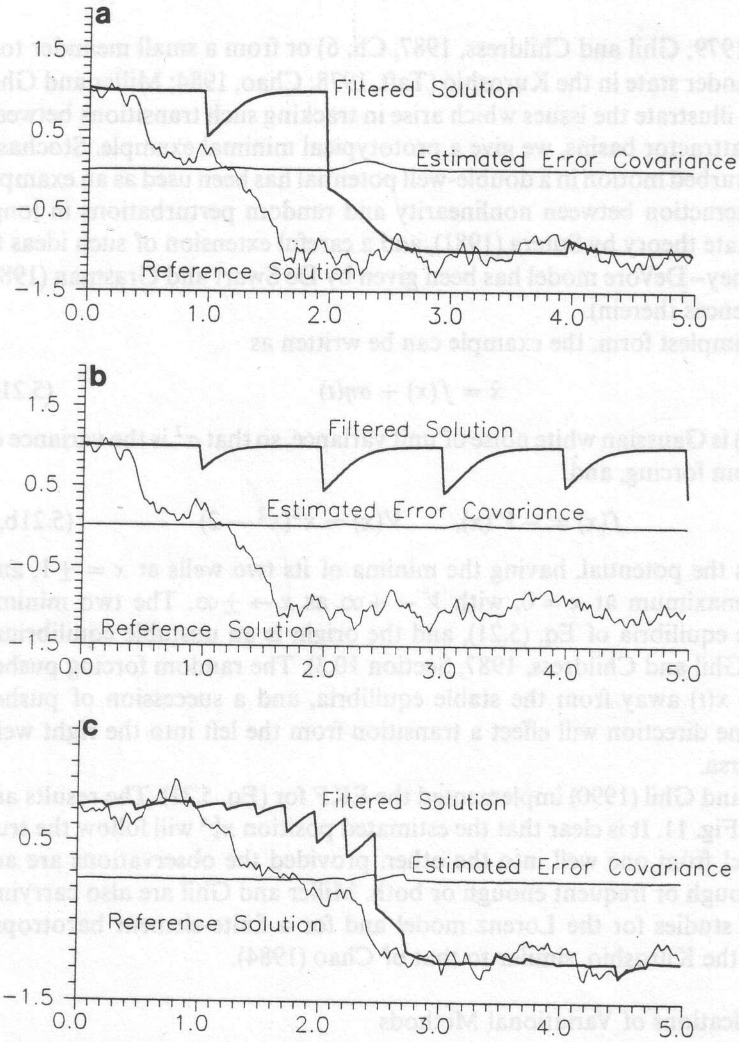


FIG. 11. K-filter experiments for stochastically perturbed motion in a double-well potential, Eq. (5.21); $\sigma^2 = 0.24$. (a) Observational noise variance $r^2 = 0.01$, observations taken at time intervals $\Delta t = 1$; (b) $r^2 = 0.04$, $\Delta t = 1$; (c) $r^2 = 0.04$, $\Delta t = 0.25$ (from Miller and Ghil, 1990).

The idea is to use the control $\mathbf{u}(t)$ so that the state vector $\mathbf{w}(t)$ reach a prescribed value $\mathbf{w}_f$ at final time t_f from an arbitrary state at initial time. In a linear system, one can choose, without loss of generality, the final state to be zero and the initial time to be zero. A simple GFD example is for Eq. (5.22) to be a linear, say tropical, ocean model, and for $\mathbf{u}(t)$ to be an arbitrarily prescribed wind stress. We shall have to restrict ourselves here to the openloop case, in which $\mathbf{u}(t)$ is independent of the state $\mathbf{w}(t)$. The closed-loop case, in which $\mathbf{u}$ depends on $\mathbf{w}$, is treated in the extensive control literature (e.g., Wunsch, 1988, and references therein).

The objective of driving $\mathbf{w}(t)$ to zero in finite time is stated more precisely by requiring that $\mathbf{u}(t)$ minimize a performance index or cost functional, $J[\mathbf{w}, \mathbf{u}]$, measuring the size of $\mathbf{w}(t)$, subject to the intuitively obvious additional condition that the control energy spent in the process be also minimized,

$$J[\mathbf{w},\mathbf{u}] \equiv \mathbf{w}_f^T\tilde{Q}_f\mathbf{w}_f + \int_0^{t_f} [\mathbf{w}^T(t)\tilde{Q}(t)\mathbf{w}(t) + \mathbf{u}^T(t)\tilde{R}(t)\mathbf{u}(t)]\,dt \qquad (5.23)$$

The formal similarity between Eq. (5.23) and the functional Eq. (4.20) for variational data assimilation should be obvious.

The key assumption in solving the minimization problem given by Eq. (5.23) is that the solution $\mathbf{u}(t)$ should be sought in the linear form

$$\mathbf{u} = -\tilde{K}(t)\mathbf{w}(t) \qquad (5.24)$$

$\tilde{K}$ remains to be determined. The Ansatz of Eq. (5.24) is similar to that of linear unbiased data assimilation, Eqs. (4.2a), (4.8b), or (5.5b). With this analogy, Eq. (5.22) looks much like the time-continuous version of the data assimilation step of the K-filter [Eq. (4.17e)], with the control $\mathbf{u}$ replacing the observational residual $\boldsymbol{\eta} = \mathbf{w}^o - H\mathbf{w}^f$ and $\tilde{K}$ replacing the Kalman gain matrix K^*. Intuitively, this similarity of Eqs. (5.22) and (4.17e) corresponds to the idea that data can force the model to the correct solution. The simplest expression of this idea in practical data assimilation is given by the nudging method (Anthes, 1974; Holland and Malanotte-Rizzoli, 1989; Section 6.3.2 here).

The analogy has in fact a solid mathematical foundation in the duality theorem proved by Kalman (1960) in the discrete-time case and by Kalman and Bucy (1961) in the continuous-time case. It states, roughly speaking, that the optimal control problem in Eqs. (5.22) and (5.23) and the optimal estimation problem given by the time-continuous form of Eqs. (4.9) and (4.10) are dual to each other, i.e., their solutions $\mathbf{u}(t)$ and $\mathbf{w}^a(t)$ are both obtained by solving analogous evolution equations. The equivalence includes solving a Riccati equation for $\tilde{P}(t)$, the quadratic performance under optimization, subject to $\tilde{P}(t_f) = \tilde{Q}_f$. In fact, final time t_f in the deterministic control problem is equivalent to initial time t_0 in the stochastic estimation problem.

The complete list of equivalences is given in Table IV and is restricted to linear systems and observations. Optimal control for a system (5.22) with cost (5.23) goes back to the Bolza problem of the calculus of variations, requiring minimization of a functional in the presence of differential equation constraints. A review of the mathematical literature on problem (5.22) with (5.23) is given by Berkovitz (1974, pp. 294–297). Important contributions were made by J. P. La Salle and E. J. McShane in the 1940s and early 1950s and by Soviet mathematicians (A. F. Filippov, R. V. Gamkrelidze, and

TABLE IV. DUALITY RELATIONSHIPS BETWEEN STOCHASTIC ESTIMATION AND DETERMINISTIC CONTROL[a]

A. Continuous (linear) Kalman Filter

System Model	$\dot{\mathbf{w}}^{\mathrm{t}}(t) = F(t)\mathbf{w}^{\mathrm{t}}(t) + G(t)\mathbf{b}^{\mathrm{t}}(t), \quad \mathbf{b}^{\mathrm{t}}(t) \sim \mathrm{N}[0, Q(t)]$
Measurement Model	$\mathbf{w}^{0}(t) = H(t)\mathbf{w}^{\mathrm{t}}(t) + \mathbf{b}^{0}(t), \quad \mathbf{b}^{0}(t) \sim \mathrm{N}[0, R(t)]$
State estimation	$\dot{\mathbf{w}}^{\mathrm{a}}(t) = F(t)\mathbf{w}^{\mathrm{a}}(t) + K(t)[\mathbf{w}^{0}(t) - H(t)\mathbf{w}^{\mathrm{a}}(t)], \quad \mathbf{w}^{\mathrm{a}}(0) = \mathbf{w}^{\mathrm{a}}_0$
Error covariance propagation (Riccati Equation)	$\dot{P}(t) = F(t)P(t) + P(t)F^{\mathrm{T}}(t) + G(t)Q(t)G^{\mathrm{T}}(t) - K(t)R(t)K^{\mathrm{T}}(t), \quad P(0) = P_0$
Kalman Gain	$K(t) = P(t)H^{\mathrm{T}}(t)R^{-1}(t)$
Initial conditions	$E[\mathbf{w}^{\mathrm{t}}(0)] = \mathbf{w}^{\mathrm{a}}_0, \quad E\{[\mathbf{w}^{\mathrm{t}}(0) - \mathbf{w}^{\mathrm{a}}_0][\mathbf{w}^{\mathrm{t}}(0) - \mathbf{w}^{\mathrm{a}}_0]^{\mathrm{T}}\} = P_0$
Assumptions	$R^{-1}(t)$ exists $E\{\mathbf{b}^{\mathrm{t}}(t)[\mathbf{b}^{0}(t')]^{\mathrm{T}}\} = 0$
Performance Index	$p^{\mathrm{f,a}}(t) = E\{[\mathbf{w}^{\mathrm{f,a}} - \mathbf{w}^{\mathrm{t}}][\mathbf{w}^{\mathrm{f,a}} - \mathbf{w}^{\mathrm{t}}]^{\mathrm{T}}\}$

B. Continuous (linear) Optimal Control

System Model	$\dot{\mathbf{w}}^{\mathrm{t}}(t) = \tilde{F}(t)\mathbf{w}(t) + \tilde{H}(t)\mathbf{u}(t)$
Measurement Model	$\mathbf{w}^{0}(t) = \mathbf{w}(t)$ (all system variables are measured)
Performing control	$\mathbf{u}(t) = -\tilde{K}(t)\mathbf{w}(t)$
Performance propagation (Riccati Equation)	$\tilde{P}(t) = -\tilde{F}^{\mathrm{T}}(t)\tilde{P}(t) - \tilde{P}(t)\tilde{F}(t) - \tilde{Q}(t) + \tilde{P}(t)\tilde{H}(t)\tilde{K}(t)$
Control Gain	$\tilde{K}(t) = \tilde{R}^{-1}(t)\tilde{H}(t)\tilde{P}(t)$
Terminal conditions	$\mathbf{w}(t_{\mathrm{f}}) = 0$ $\mathbf{P}(t_{\mathrm{f}}) = \tilde{Q}_{\mathrm{f}}$
Cost function	$J[\mathbf{w}, \mathbf{u}] = \mathbf{w}_{\mathrm{f}}^{\mathrm{T}}\tilde{Q}_{\mathrm{f}}\mathbf{w}_{\mathrm{f}} + \int_0^{t_{\mathrm{f}}} [\mathbf{w}^{\mathrm{T}}(t)\tilde{Q}(t)\mathbf{w}(t) + \mathbf{u}^{\mathrm{T}}(t)\tilde{R}(t)\mathbf{u}(t)]\, dt$

C. Estimation-Control Duality

Estimation	Control
t_0 initial time	t_{f} final time
$\mathbf{w}(t)$ unobservable state variable of random process	$\mathbf{w}(t)$ observable state variable to be controlled
$\mathbf{w}^{0}(t)$ random observations	$\mathbf{u}(t)$ deterministic control
$F(t)$ dynamic matrix	$\tilde{F}^{\mathrm{T}}(t)$ dynamic matrix
$Q(t)$ covariance matrix for the model errors	$\tilde{Q}(t)$ quadratic matrix defining acceptable errors on model variables
$H(t)$ effect of observations on state variables	$\tilde{H}(t)$ effect of control on state variables
$P(t)$ covariance of estimation error under optimization	$\tilde{P}(t)$ quadratic performance under optimization
$K(t)$ weighting on observation for optimal estimation	$\tilde{K}(t)$ weighting on state for optimal control

[a] (A), Kalman filter as the optimal solution for the former problem; (B), optimal solution for the latter problem; (C), equivalences between the two (after Kalman, 1960, and Gelb, 1974, Section 9.5; courtesy of R. Todling).

L. S. Pontryagin) in the late 1950s. A brief review of the engineering literature, including other connections between recursive least squares, least variance, and control as well as nonlinear results is given by Jazwinski (1970, pp. 151–158). Of particular interest for nonlinear problems are Bellman's quasi-linearization and invariant imbedding techniques (e.g., Bellman *et al.*, 1966). The optimal control work of Lions and his associates in France (e.g., Lions, 1971) clearly motivated and influenced the development and applications of the adjoint method by Le Dimet, Talagrand, and their associates (see Section 5.4.3). The solution of the so-called linear quadratic tracking problem in discrete time had been derived previously by Kalman and Koepcke (1958). Notice that for a truly optimal solution of this problem, the number of operations for the adjoint Riccati equation yielding $\tilde{P}(t)$ backward in time is still $0(N^2)$.

Various forms of the well-known linear duality between deterministic control and stochastic estimation have been given in the GFD literature (e.g., Thacker, 1986; Wunsch, 1988, and references therein). In particular, the adjoint method for variational minimization of the distance between a perfect-model trajectory and given data is closely related to the optimal control problem (5.22) and (5.23), although it is not the exact dual of Eq. (4.17).

To allow for errors in model and data, one has to consider the stochastic control problem. In this case, one does not assume that the model is perfect nor that the trajectory is known, as in Eqs. (5.22) and (5.23), but allows instead for system noise $Q \neq 0$ and an observation model [Eq. (4.10)]. For linear systems, this problem is solved by the separation principle (e.g., Gelb, 1974, pp. 361–365).

One still wishes to minimize Eq. (5.23), but $\mathbf{w}(t)$ is no longer known with certainty, only via incomplete and noisy observations [cf. Table IV(a)]. The crucial observation is that $\mathbf{w}^a(t)$ is still determined by a K-filter, independently of the control $\mathbf{u}(t)$, which is assumed to be known, and that minimization of the modified cost function for the control does not depend on the noise covariance Q. Thus, the optimal control in the stochastic problem can be obtained by the cascading of two steps: the state $\mathbf{w}^a(t)$ is estimated first by a K-filter; then the control $\mathbf{u}(t)$ is calculated from Eq. (5.24) by determining the appropriate gain $\tilde{K}(t)$ from the deterministic procedure in Table IV(b).

The actual application of stochastic optimal control ideas to variational data assimilation with weak constraints in meteorology and oceanography is an open research problem and hence beyond the scope of this review. But the separation principle discussed earlier suggests the following approach: First determine the statistically appropriate weights $A(t)$ and $\Gamma(t)$ in Eq. (4.20) by a sufficiently long application of the adaptive K-filter (Dee *et al.*, 1985), as $A \cong R^{-1}$ and $\Gamma \cong Q^{-1}$ (cf. Section 4.2). Then, keep A and Γ fixed and continue assimilation by using the adjoint method. Reapply the adaptive K-filter when the large-scale flow, and hence its subgrid-scale manifestations, have changed substantially to modify A and Γ and so on. This could combine some of the advantages of both methods in terms of the error estimates provided by the K-filter and the relative simplicity of the adjoint method.

With this theoretical background, we are prepared to consider actual examples of variational data assimilation in meteorology.

5.4.2. Direct Minimization

As indicated in Section 4.2, it is desirable for the purposes of data assimilation to circumvent the classical variational approach of deriving and solving Euler–Lagrange equations for a given quadratic functional [Eqs. (4.20), (4.21), and (5.23)]. Instead, modern computing devices permit the use of direct minimization for sizable state vectors and data sets. In particular, such an approach might be desirable for novel types of satellite data, for which the empirical statistics required by linear regression methods, such as OI, are difficult to accumulate.

This idea was applied first by Ghil and Mosebach (1978) to temperature retrievals from NASA's DST-6 experiment (see Section 5.1). The functional chosen was

$$J[T, \mathbf{V}, p_s] = \int_0^1 \int_\Sigma \{\alpha(T - T^0)^2 + \beta|\mathbf{V} - \mathbf{V}_g^0|^2 + \gamma(p_s - p_s^0)^2 + \delta(\partial p_s/\partial t)^2\}\, d\Sigma\, d\sigma \tag{5.25a}$$

using notation similar to that of Eq. (4.21). The volume over which minimization was carried out extended over several grid points and model σ-levels; within this volume lay a number of vertical temperature profiles derived from a polar-orbiting satellite. Direct measurements used were satellite-retrieved temperatures T^0 and surface pressures p_s^0, with pseudo-observations of wind $\mathbf{V}_g^0$ derived from the former by the geostrophic relationship. The weak constraint used was the continuity equation in σ-coordinates,

$$\partial p_s/\partial t = -\nabla \cdot \int_0^1 p_s \mathbf{V}\, d\sigma \tag{5.25b}$$

α, β, γ, and δ were all prescribed positive constants. Equation (5.25b), along with the geostrophic relation between $\mathbf{V}^0$ and T^0, couple all the variables in Eq. (5.25a).

A conjugate-gradient algorithm was used to minimize Eq. (5.25a) subject to Eq. (5.25b). Substantial changes from the forecast model's first guess for the wind field were obtained over the otherwise data-void southern oceans, with convergence of the algorithms in a few iterations. Complete implementation of the method for semi-operational use was not feasible at that time because of its computational cost.

Direct minimization for a local patch of satellite data was applied by

Hoffman (1982) to the interesting problem of dealiasing wind data from the Seasat-A satellite scatterometer (SASS). The SASS instrument provided surface winds from wave backscatter with reasonably accurate magnitudes, but with directions having to be selected from 1, 2, 3, or 4 possible values, called aliases. Hence the need for an objective method of directional ambiguity removal, or dealiasing.

Hoffman (1982) used a quadratic functional based on the sum of four terms, penalizing errors separately with respect to SASS wind vectors and wind speeds, as well as and with respect to conventional data and a forecast. He applied a multistep procedure in which the four weights are changed gradually from emphasizing the first guess, in this case the forecast, to emphasizing the SASS velocity data. The procedure was applied to a region in the North Atlantic surrounding the storm that damaged the luxury liner Queen Elizabeth II (QE II) on 10 September 1978 and which had been underpredicted by forecasts not using the SASS data. The region contained about 1000 grid points and about 6500 data points. The resulting objective analysis was satisfactory, with the multistep procedure providing a lower minimum than a single-step procedure. But the aliases of the SASS winds still showed meteorologically unrealistic small-scale features.

Hoffman (1984) added smoothness and vorticity-advection constraints and analyzed again the QE II storm, as well as an additional one centered south of Japan on 6 September 1978. The constrained analysis was more robust to various changes in parameters, but it still had difficulties in accommodating systematic errors in various data sources and in defining sharp fronts. Both types of problems are common to all objective-analysis methods not incorporating data-adaptive error structures.

Both variational analyses of SASS data (Hoffman, 1982, 1984) used only data lumped together at one synoptic time, 1200 GMT for the Atlantic QE II storm and 0000 GMT for the Pacific storm south of Japan. Harlan and O'Brien (1986) applied a simpler variational method for the assimilation of SASS winds over 24 hr, centered at 1200 GMT for the QE II storm; their emphasis was on modifying NMC's surface pressure field analysis for the purposes of ocean prediction. Direct minimization with respect to simulated data distributed over a 42-hr time interval in 6-hr steps was carried out by Hoffmann (1986) using two versions of a two-layer spectral model on an f-plane. The PE version was used to generate the data, and the QG version was used to assimilate them on the interval $-t^* \leq t \leq 0$ and forecast for $t > 0$ (see also discussion of observing system simulation experiments in Section 6). The weights given to the data decreased monotonically in time from $t = 0$ to $t = -t^*$. The results are shown in Fig. 12.

The variational 4-D assimilation method produced the best estimate of the atmospheric state towards the middle of the interval over which data are

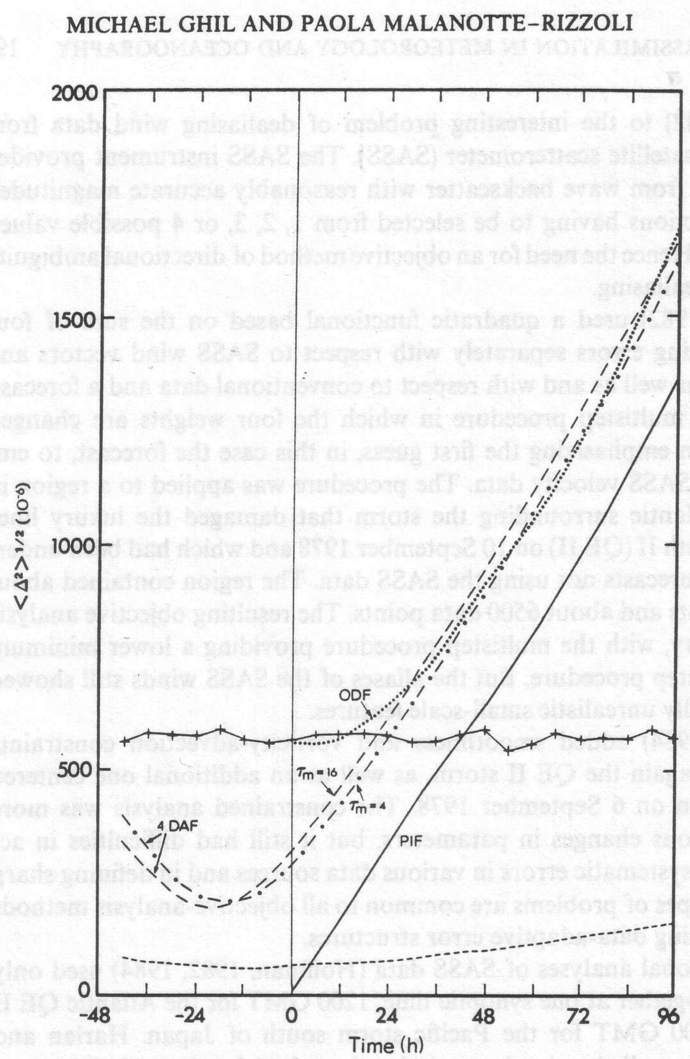


FIG. 12. Rms global error $\langle \Delta^2 \rangle^{1/2}$ of an ensemble of 50 independent analysis-forecast cycles, as a function of time: ODF (dotted line), ordinary dynamic forecast, using data at $t = 0$ only; PIF (solid line), perfect initial-data forecast; 4DAF (dashed and dash-dotted lines), 4-D analysis and forecast (using slightly different weights for data at different times). The rms measurement error is shown by the solid line with plus signs (from Hoffman, 1986).

available, in contradistinction from the K-filter which produces it towards the end (e.g., Fig. 8). The exact position of the minimum rms error in Fig. 12 depends slightly on the weights given to the data: it shifts to the right as more weight is given to the most recent data (dash-dotted line in the figure). But in any case, the forecast started at $t = 0$ using the 4-D variational estimate is better than the forecast using exclusively data at $t = 0$ only for a short while—about 36 hr.

Since this result does not appear to depend on the distribution of weights, it is also unlikely to depend on the particular minimization method used: direct or adjoint. Still, the adjoint method is computationally more efficient for the 4-D assimilation of data over a large domain, using the time-dependent model equations themselves as a strong constraint.

5.4.3. The Adjoint Method

As outlined in Section 4.2, this method provides simply an efficient way for computing the gradient of a quadratic functional [Eq. (4.22)] with respect to the initial data for the exact solution of an evolution equation (4.23), which minimizes the distance to the data over an interval $0 \leq t \leq t^*$. The computation of this gradient involves linearizing about the current iterate of the trajectory and solving the adjoint of this linearization backwards in time.

The adjoint method falls within a broader class of methods for solving constrained minimization problems. In the meteorological literature these are reviewed succinctly and clearly by Le Dimet and Talagrand (1986). Using W for the meteorological fields discretized in space and time as $\mathbf{w}_k$ before, one wishes to minimize the cost functional

$$J[W] = \int_\Sigma \|W - W^0\|^2 d\Sigma \tag{5.26a}$$

where $\|\cdot\|$ is a suitable norm, subject to the dynamical constraint

$$F[W] = 0 \tag{5.26b}$$

The latter is a strong constraint in the terminology of Sasaki (1970). In the mathematical optimization literature, minimization with a weak constraint is still referred to as unconstrained optimization. The class of methods reviewed by Le Dimet and Talagrand (1986) reduces, in fact, a constrained to an unconstrained minimization problem.

Introducing an inner product $\{\cdot,\cdot\}$ compatible with the norm $\|\cdot\|$ into the function space appropriate for W, one defines the Lagrangian functional

$$\mathscr{L}_\delta[W, \Lambda] = J[W] + \delta\{\Lambda, F[W]\} \tag{5.27}$$

for the problem in Eq. (5.26a, b), where Λ is the Lagrange multiplier and $\delta = 1$; compatibility of the norm and inner product simply means that $\{W, W\} = \|W\|^2$. The augmented Lagrangian $\mathscr{L}_{\varepsilon,\delta}$ is then

$$\mathscr{L}_{\varepsilon,\delta} = \mathscr{L}_\delta + \frac{1}{\varepsilon}|F[W]|^2 \tag{5.28}$$

where $0 < \varepsilon \ll 1$ is the control parameter while W is the control variable. The solution W^* of the constrained problem in Eq. (5.26) is sought as the limit $W_k \to W^*$ of a sequence of unconstrained problems [Eq. (5.28)], as follows: Given a triplet $(W_k, \Lambda_k, \varepsilon_k)$, W_{k+1} is determined by minimizing $\mathscr{L}_{\varepsilon_k,\delta}[W, \Lambda_k]$, with W_k as a first guess; Λ_k and ε_k are updated by

$$\Lambda_{k+1} = \Lambda_k + \frac{1}{\varepsilon_k}|F[W_{k+1}]| \tag{5.29}$$

$$\varepsilon_{k+1} = c_k \varepsilon_k \tag{5.30}$$

with $0 < c_k < 1$. Bertsekas (1982) gives a proof of the convergence of this augmented Lagrangian algorithm and practical indications for choosing the sequence of c_k's. The latter are valuable in accelerating convergence, because in practice one always stops short of $\varepsilon = 0$, and hence one never satisfies Eq. (5.26b) exactly. The use of a sequence ε_k is somewhat analogous to Hoffman's (1982) shifting the weights of the summands in the cost functional from emphasizing the first guess to emphasizing the SASS data.

This algorithm is the most efficient and general of its class. The *penalty algorithm* is obtained from Eq. (5.28) by letting $\delta = 0$, i.e., $\Lambda \equiv 0$, while the *duality algorithm* follows by letting $\varepsilon \to \infty$. The latter is unrelated to the duality principle discussed in Section 5.4.1; it involves instead the alternate use of ascent and descent steps in determining the saddle points of the Lagrangian $\mathscr{L}_1$ in Eq. (5.27). The augmented Lagrangian algorithm avoids the ill conditioning that occurs in the penalty algorithm as ε becomes very small and provides greater freedom in choosing the first-guess Λ_0 than in the duality algorithm, avoiding the problems the latter encounters when the Lagrangian $\mathscr{L}_1[W, \Lambda]$ is not globally convex with respect to W. Many practical considerations on the convergence and relative efficiency of optimization algorithms can be found in Fletcher (1987) and Gill *et al.* (1982).

The augmented-Lagrangian algorithm was used by Navon and De Villiers (1983) to maintain global energy and enstrophy constraints in the time integration of a shallow-water model. Le Dimet and Talagrand (1986) applied it to the minimization of the functional (4.21), with the constraint $\mathbf{s} = 0$ being the steady-state form of the shallow-water equations, using a one-layer version of Eq. (5.20). The weights α and β in Eq. (4.21) were chosen as $\alpha(\mathbf{x}, t) \equiv 1$ and $\beta(\mathbf{x}, t) \equiv$ const. $\neq 1$. The domain Σ for which they sought a variational analysis was a square with a side of 2500 km centered at 45°N and 5°W, and observations of both ϕ and $\mathbf{V}$ were provided at all points of a 25×25 grid covering Σ. The purpose of the minimization was thus to reduce the error in the observations, rather than to fill data gaps.

The results are shown in Table V. The quantities E_u, E_v, and E_ϕ listed are rms values of the residual dynamic imbalances in the right side of Eq. (5.20a,b), in m sec^{-2} and m sec^{-1}, respectively. Convergence to rms

TABLE V. VARIATIONS OF THE RESIDUAL IMBALANCES E_u, E_v, E_ϕ WITH THE ITERATION NUMBER k OF THE AUGMENTED-LAGRANGIAN ALGORITHM[a]

k	1	2	5	10
E_u	0.29×10^{-3}	0.19×10^{-3}	0.42×10^{-5}	0.18×10^{-5}
E_v	0.12×10^{-3}	0.11×10^{-3}	0.24×10^{-5}	0.24×10^{-5}
E_ϕ	0.13×10^{-1}	0.10×10^{-1}	0.61×10^{-3}	0.21×10^{-3}

[a] From Le Dimet and Talagrand (1986).

values smaller by two orders of magnitude than the initial imbalances occurred in 10 steps. The changes in the geopotential heights and the velocity components during the adjustment were 2 m and 1 m $\sec^{-1}$, respectively, i.e., they stayed within the accuracy of the original observations.

The main feature that renders the adjoint method computationally more efficient than the class of optimization methods described so far in this section is that it works with a smaller number of discrete variables. This reduction of the dimension of the control variable is obtained by applying the following key observation of control theory: The minimizing solution W^* is determined uniquely by its initial and boundary values over the generalized boundary, in time and space, of the domain of interest, say $\Sigma \times [0, t_f]$. Thus, instead of minimizing with respect to W, one can minimize with respect to the appropriate initial values. It is this observation that yields the algorithm summarized in Eqs. (4.24)–(4.27).

The adjoint method for solving time-dependent constrained minimization problems [Eqs. (4.22) and (4.23)] was introduced into the meteorological literature by Penenko and Obraztsov (1976). Concrete examples were worked out simultaneously and interdependently by Lewis and Derber (1985) and by Le Dimet and Talagrand (1986). The latter used a purely 1-D ($v \equiv 0$), nonlinear version of the shallow-water equations and simulated observations of geopotential at two times, t_1 and t_2, to obtain a reduction of the cost functional by a factor of two per iteration step. The fprmer used both simulated and real data.

The simulated-data studies of Lewis and Derber (1985) considered the advection equation

$$q_t + uq_x = 0, \qquad t > 0, \qquad 0 \le x < 2\pi \tag{5.31a}$$

with periodic boundary conditions and initial data chosen as a pure sine wave

$$q(x, 0) = \sin x, \qquad 0 \le x < 2\pi \tag{5.31b}$$

The advection velocity was taken first as constant

$$u \equiv \mathrm{c} \equiv \text{const.} \tag{5.32a}$$

then as variable but given,

$$u(x) = \frac{6x}{(2\pi)^2}(2\pi - x) \tag{5.32b}$$

and finally the nonlinear equation with

$$u(x, t) \equiv q(x, t) \tag{5.32c}$$

was studied. The constant-velocity case [Eq. (5.32a)] can be analyzed completely and solved in a single forward–backward iteration step to yield 50% rms error reduction in the limit of small separation between the two observations times. In this case, the transition matrix is orthogonal, yielding these very simple but also very special results (cf. Miller, 1987). The linear, variable-coefficient case [Eq. (5.32b)] requires multiple iterations, and the error reduction for finite separation is less than in the constant-coefficient case. In the nonlinear case [Eq. (5.32c)], the possibility of multiple solutions had to be avoided, and rms error reduction of 20% only was obtained.

Lewis and Derber (1985) also used real data from six analyses over the central United States, three hours apart, from the Atmosperhic Variability Experiment, 6–7 March 1982. The analyses were generated separately from a special network of rawinsonde observations (RAOB) at 2100,0000, and 0300 GMT and from temperature retrievals of the VISSR Atmospheric Sounder (VAS) at 2030, 2330, and 0230 GMT. There were about 35 RAOB and 180 VAS data available for each of the analyses, corresponding to an average spatial separation of 250 km and 100 km, respectively. The dynamic constraint of single-level geostrophic potential vorticity conservation was applied separately to each set of three RAOB and of three VAS analyses, at 700 mb and at 250 mb. The results are shown in Fig. 13.

There are clear differences between the input RAOB and input VAS analyses. The rms discrepancy in geopotential heights was 14 m at 700 mb and 42 m at 250 mb, respectively. Variational adjustments to the heights using the conjugate-gradient method for Eq. (4.27) resulted in rms changes of 3.7 m for the RAOB and 7.9 m for the VAS analyses at 700 mb; the corresponding changes at 250 mb were 13.5 m and 22.0 m, respectively. As a consequence, the rms difference between the RAOB and VAS analyses was reduced to 12 m at 700 mb and increased to 53 m at 250 mb. In particular, a systematic positive bias of 6 m in the VAS versus the RAOB analyses at 250 mb was not removed by the adjustment. This problem of warm biases in satellite temperature retrievals was treated, for instance, by Ghil *et al.* (1979) in the context of combining conventional and remote-sounding data during a sequential estimation process. Similar biases, resulting in nonzero values of long-range correlations for OI also occur in the operational data assimilation systems

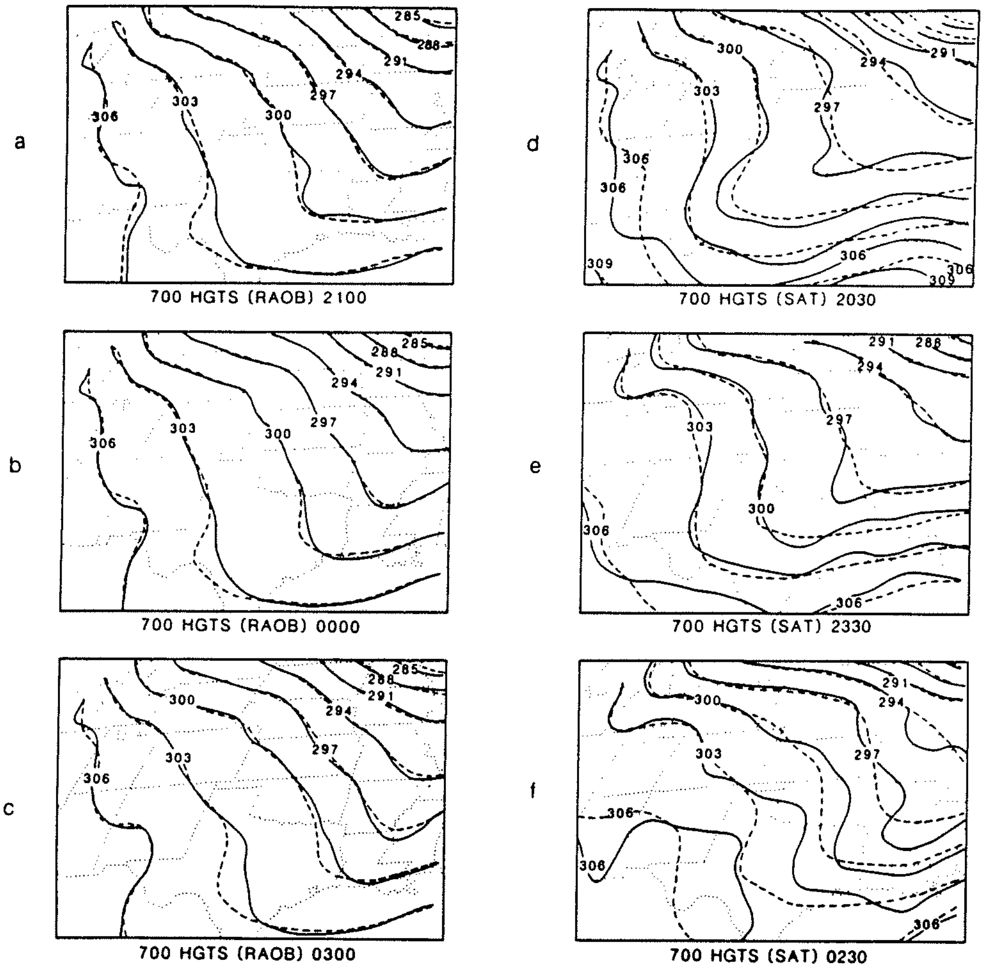


FIG. 13. Height analyses at 700 mb on 6–7 March 1982. Solid lines are contours of the input analysis; dashed lines are contours after adjustment by the adjoint method. Both input and output analysis are on a 1° lat. × 1° long. grid. (a)–(c) based on RAOB; (d)–(f) based on VAS retrievals (from Lewis and Derber, 1985).

at ECMWF and the French Direction de la Météorologie Nationale (P. Courtier, personal communication, 1990).

Talagrand and Courtier (1987) used the barotropic vorticity equation

$$\zeta_t = J(\zeta + f, \Psi) \qquad (5.33a)$$

$$\zeta = \Delta\Psi \qquad (5.33b)$$

on the sphere to study the adjoint method for simulated data. They truncated at total wave number $n = 21$ and kept only the $N = 231$ real components antisymmetric about the equator for both vorticity ζ and streamfunction Ψ, which were hence symmetric for the zonal velocity and antisymmetric for the meridional velocity. The test case was a Rossby–Haurwitz wave with $n = 5$ and zonal wave number $m = 4$, propagating eastward without change of shape by 9.55° per day. Minimization of the quadratic distance between the analyzed vorticity $\zeta(\mathbf{x}, t)$ and the observed vorticity $\zeta^o(\mathbf{x}, t)$ over a 12 h time interval in the Northern Hemisphere was started from a state of rest, $\zeta^{(0)}(\mathbf{x}, t) \equiv 0$. When complete data ζ^o were provided every time step, the initial

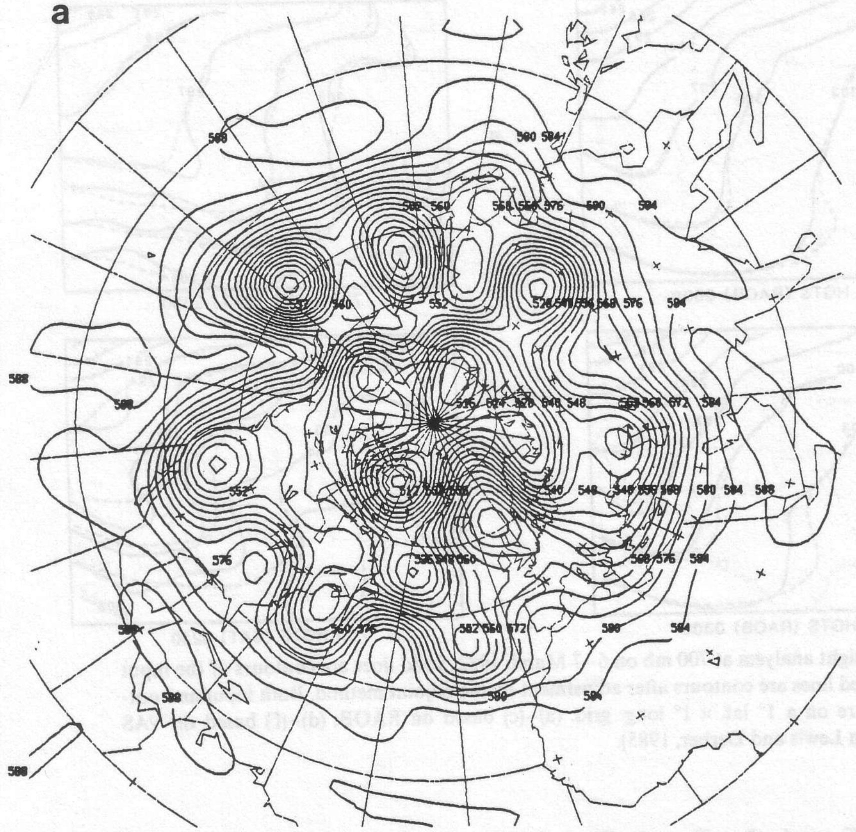


FIG. 14. 500 mb height field for 0000 GMT 26 April 1984. (a) Variational analysis minimizing distance to data over 24-hr intervals; (b) differences between this variational analysis and the operational analysis of the Direction de la Météorologie Nationale. Units are dam, and the contour interval is 4 dam (from Courtier and Talagrand, 1987).

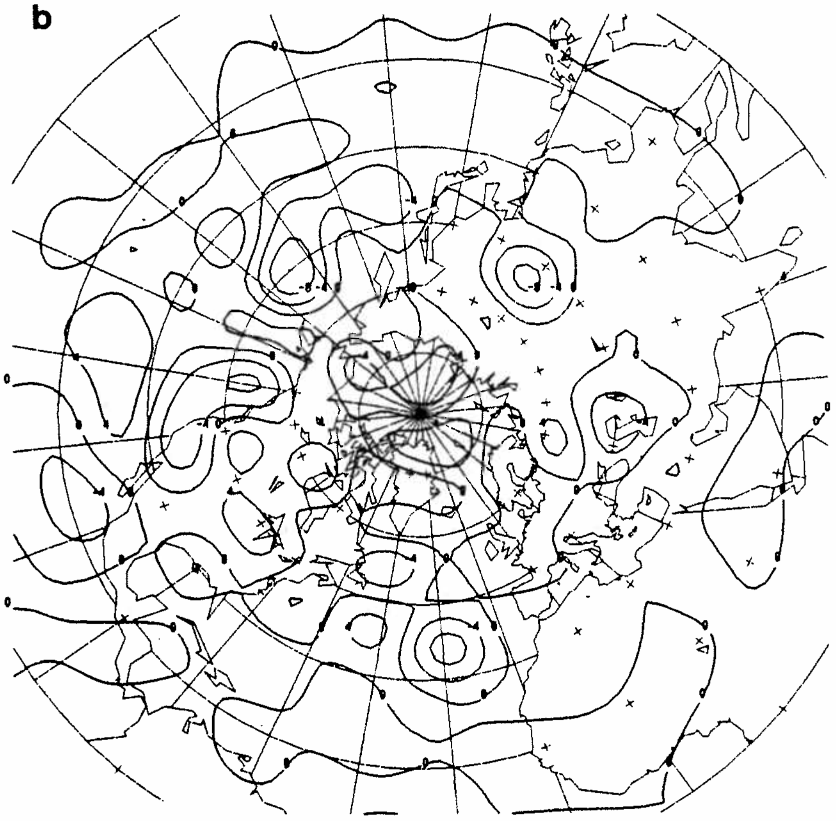


FIG. 14. (*Continued*)

vorticity field $\zeta^o(\mathbf{x}, 0)$ was reconstructed to within an accuracy of 10^{-9} sec^{-1} after five iterations. When only $\zeta^o(\mathbf{x}, t_f)$ at $t_f = 12$ hr was provided, the initial field was reconstructed to within 10^{-8} sec^{-1} in eight descent steps.

Courtier and Talagrand (1987) applied the same constraint of Eq. (5.33) to operational 500 mb data from a 24 hr interval, 0000 GMT 25 April 1984 to 0000 GMT 26 April 1984. The data contained 1653 geopotential and 2 × 1913 horizontal velocity components for a total of $p = 5479$ scalar observations over the Northern Hemisphere; all but 11 reports were from radiosondes. The cost functional was a sum of quadratic residuals for the wind and geopotential observations, with model streamfunction being transformed into geopotential by solving the nonlinear balance equation in the easy direction (from Ψ to ϕ).

Figure 14a shows the height field produced by the minimization process at the final time, 0000 GMT 26 April 1984; Fig. 14b shows the difference between this variational analysis and the operational analysis of the French weather service. The rms difference between the variational analysis and the observations is 29.1 m for the heights and 8.0 m sec^{-1} for the winds, compared to values of 185 m and 17.6 m sec^{-1} for an atmosphere at rest. Using wind observations only yields 33.5 m for heights and 7.7 m sec^{-1} for winds; height observations only yield 28 m and 8.8 m sec^{-1}. These values are much smaller than the corresponding climatological standard deviations, and the variational analysis also reconstructs certain features not observed directly over the data-sparse Pacific Ocean. This work was extended by Courtier and Talagrand (1990) to the nonlinear shallow-water equations at 500 mb, with truncation $n = 21$ and $n = 42$.

Using a shorter update interval of 4 hr reduces the cost functional in the highly overdetermined, $p \gg N$, problem [Eq. (5.33)], but reconstructon of unobserved features is no longer possible. Courtier and Talagrand (1987) noticed that in the adjoint approach, advection of information occurs not only downstream, as in sequential estimation (Ghil *et al.*, 1981, 1982; Ghil, 1989), but also upstream with the forced adjoint in Eq. (4.26). The reach of this advection, however, is still limited by its speed of propagation. The use of longer time intervals over which to minimize the distance between model trajectory and the data obviously increases the computational burden. It also leads to increased difficulties caused by the instability of the flows, the consequent divergence of forward and backward trajectories, and the appearance of multiple minima of the cost functional (F. Gauthiez, personal communication, 1990; Miller and Ghil, 1990).

Derber (1989) also noted the discontinuity, created by variational methods using multiple levels in time, between analyses based on successive 4-D assimilation intervals. In particular, in the adjoint method, the state at time $t = 0$ determined from data over the interval $[0, t^*]$ is substantially different from the state obtained from a forecast started at $t = -t^*$, using data over $[-t^*, 0]$. In fact, Courtier and Talagrand (1987) attributed to this clash of disjoint sets of observations much of the differences between their variational result and the operational one (Fig. 14b).

To circumvent this difficulty in the context of variational methods, Derber (1989) proposed a variational continuous assimilation (VCA) technique. This technique also tries to move away from the perfect-model assumption of constrained optimization approaches by rewriting Eq. (4.18), which is the discrete-time version of Eq. (4.23), as

$$\mathbf{w}^f_{k+1} = \mathbf{N}_k(\mathbf{w}^f_k) + \lambda_k \phi \qquad (5.34)$$

where λ_k is a sequence of scalars determined *a priori*, and ϕ is a spatially dependent vector of the same dimension as $\mathbf{w}$, determined in the VCA process. Minimization of the mean-square observational residual with respect to ϕ replaces in VCA minimization with respect to $\mathbf{w}_0$ in the adjoint method. The determination of ϕ in VCA can be thought of as that of a constant bias, $E\mathbf{b}_k^t \equiv \phi$, in the sequential estimation formulation [Eqs. (4.9) and (4.10)] to which VCA is much closer in spirit than the adjoint method.

Derber (1989) used three guesses for the sequence $\{\lambda_k\}$: (i) $\lambda_0 = 1$, with all other $\lambda_k = 0$; (ii) $\lambda_k = 1/K$, where K is the total number of model time steps over which the minimization is done; and (iii) a parabolic profile, with $\lambda_0 = \lambda_{K-1} = 0$ and its maximum at the middle of the assimilation interval, normalized like (i) and (ii) so that

$$\sum_0^{K-1} \lambda_k = 1$$

Derber's model was a ten-level QG model discretized on a 25 × 30 grid with 200 km spacing. His data were four FGGE analyses from GFDL's level IIIb cycle, 12 hr apart, over the time interval 0000 UTC 18 February to 0000 UTC 20 February 1979, covering the development of the extensively studied President's Day storm.

The results in Fig. 15 compare VCA over a 12 hr interval, using each one of the three choices of $\{\lambda_k\}$ just shown, with a model forecast from the initial FGGE analysis and $\phi \equiv 0$ (solid line), on the one hand, and the adjoint method (dashed line) on the other. The first choice of $\lambda_0 = 1$ (dashes and double dots) behaves almost like the adjoint method, as expected, and better

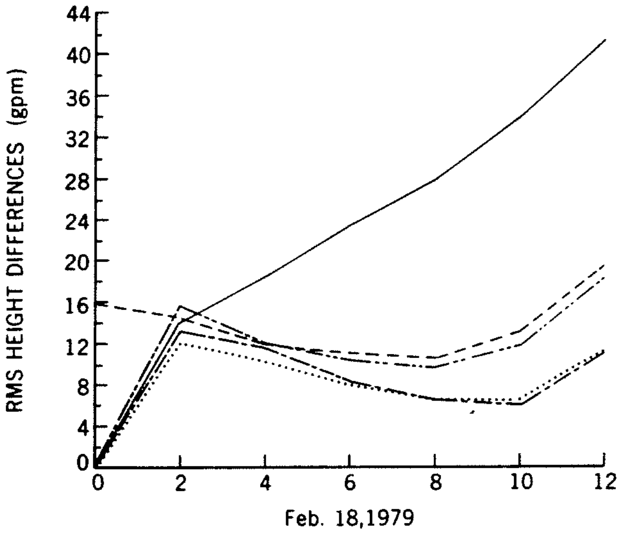


FIG. 15. Rms height differences between FGGE analyses and VCA solutions for a 12-hour assimilation interval. Three variants of VCA, differing by their prescription of λ_k in time, as well as an adjoint-method solution, are shown (from Derber, 1989). (——), Forecast from FGGE analyses; (– – –), adjoint; (—·—) parabolic λ; (—··—), delta function λ; (···), constant λ.

than the pure forecast, obviously. The constant (dotted) and parabolic (dash-dotted) choice of λ_k give mutually similar results, both of them much better than the adjoint method, which also leads to an initial jump in the solution. Forecasts from the adjoint assimilations (not shown here) were considerably less accurate than any of the others.

The ideas, methods, and results reviewed in this section show, aside from the maturity of meteorological data assimilation, a host of remaining problems. The emergence of new observing systems per se will aggravate rather than solve these problems by the complexity of the different observing patterns, the novelty of the distinct error characteristics, and the necessary extension of numerical models to domains and scales little explored so far. At the same time, it will be possible to test and compare much more fully the wealth of new ideas, from optimization and sequential estimation theory, due to rapidly increasing computing power and memory size.

6. Current Status of Oceanographic Data Assimilation

As mentioned in the introduction, the 1990s will mark a profound revolution in the history of oceanography as new technology will for the first time provide oceanographers with large synoptic data sets. Specifically, three new techniques will be of crucial importance. First, altimetry will provide global maps of sea surface height that, in the case of the oceanographically designed Topographic mission Experiment (TOPEX/POSEIDON), starting in 1991) will have a horizontal resolution of about 300 km × 300 km in midlatitudes, corresponding to a 10-day orbital period. The importance of satellite altimetry lies in the fact that the surface elevation of the ocean relative to the geoid can be shown to represent closely the pressure distribution produced by the large-scale general circulation, assumed to be in quasi-geostrophic balance (Pedlosky, 1987).

Second, scatterometry will provide one of the two major surface forcing functions of the ocean circulation, the wind stress field, with a horizontal resolution of 1° longitude × 1° latitude for two-day vector-averaged velocities [World Ocean Circulation Experiment (WOCE), 1989]. Third, ocean acoustic tomography, even though projected further into the future, has the potential of providing a 3-D picture of the interior density and velocity fields of the ocean. The most important potential use of tomography lies in its integrating properties. The tomographic measurement per se is an integral performed over long paths at the sound speed of ~1.5 km/sec. Thus, it is capable of averaging out the energetic mesoscale eddy field and measure averages over the large space and time scales of motion.

This capability of tomography to measure integral properties has already been proved for the total transport (or average velocity) over long distances (Howe *et al.*, 1987) and for relative vorticity (Ko *et al.*, 1989; Chester *et al.*, 1991). However, tomography associated with inverse methods also has the capability of mapping the mesoscale eddy field (Cornuelle *et al.*, 1985). Mesoscale mapping over areas 1000 km × 1000 km is presently being tested in experiments where acoustic moorings have been augmented with a movable ship-based receiver (Cornuelle *et al.*, 1988). Thus, an unprecedented synoptic data set will become available at the end of the 1990s with which to interpret, understand, and predict the evolution of the ocean general circulation.

Numerical models of the ocean general circulation will play a key role in this process. Until now, with a few noteworthy exceptions (Clancy, 1987; Leetmaa and Ji, 1989; Robinson *et al.*, 1989), oceanographic modeling has proceeded rather independently from observations, mainly due to the inadequacy of the observations in providing effective tests for model descriptions and predictions. The lack of effective data feedback has resulted in ocean models that are rather less realistic and sophisticated than their meteorological counterparts, with respect to the parameterizations of internal physics, as well as in the inclusion of realistic geometries and forcing functions. The two previous parallel paths of modeling and observations are, however, on the threshold of converging, thanks to the data sets just described. The models will become more realistic and consequently will provide more reliable estimates of the fields of interest where data sets are sparse.

The process of combining data with models, that is the process of model initialization and data assimilation, is relatively new to oceanographers. As discussed in Sections 1 and 3, the difference in emphasis, applied or theoretical, between the meteorological and oceanographic data assimilation process, as well as the differences between the two geofluids, will oblige oceanographers to reinterpret and adapt assimilation techniques and not simply borrow them from engineering, meteorology, or geophysics. Moreover, even the anticipated synoptic data sets will be very different from those available in meteorology. Meteorological data sets consist by and large of pointwise measurements that are sparse and have an irregular density coverage. They are collected on a global scale at varying vertical levels and at a high sampling frequency as discussed in Section 3.2.

Altimetry will give a global, very regular horizontal network of sea-level height, i.e., of the surface pressure. Apart from the problem of removing the earth's geoid from the altimetric signal, it is still unclear how to make the best use of the altimetric data set, i.e., whether to use the data in pointwise fashion along the actual tracks and sequentially in time or, alternatively, to construct optimally interpolated maps for each complete global coverage. Specific examples related to this question will be discussed in the following paragraphs. Moreover, it is rather uncertain to which degree the surface elevation is tied to the interior flows. A crude scale analysis based upon quasi-geostrophic balance suggests that a horizontal pressure gradient at the sea surface over mesoscale length scales (~100 km) should reflect interior movements down to the depth of the main thermocline (1500 m; Wunsch, 1989a). Even if this were the case, the deep oceanic layers will not be seen by altimetry and will remain void of measurements over almost the entire world ocean.

Tomography has the potential of providing 3-D imagery of the interior water mass. Exploiting the integrating nature of the tomographic measurement implies filtering out the mesoscale eddy field considered as noise, thus seeking estimates only of the large-scale, long-time component of the circulation. However, it is theoretically well known, and demonstrated at least in numerical experiments (Holland and Rhines, 1980), that eddy-momentum fluxes are capable of giving rise to large-scale, quasi-steady components of the circulation, especially in the deep layers. Thus, the question is how to reconstruct a filtered-out eddy field or at least its statistics. Obviously this must be done by using a dynamical tool, such as a dynamical model.

Moreover, the concept that the mesoscale eddy field is noise may be very misleading in specific examples. There are major and very energetic regions of the world ocean, for instance the western boundary currents of which the Gulf Stream system is the prototype, where the dynamics is dominated by the range of complex interactions between the mean flow (the current jet) and the associated Rossby wave radiation. In such systems, the mesoscale is the essential part of the signal one seeks to measure. Entire experiments such as Synoptic Ocean Prediction (SYNOP) are devoted to map and predict the mesoscale in such systems. There, however, tomography is not the best experimental tool because, in a traditional middepth tomographic configuration, the acoustic rays are not capable of penetrating the swift core of the current. Bottom-mounted configurations that might overcome the acoustic problem have not yet been proved successful (Agnon *et al.*, 1989). And even in more quiescent parts of the ocean, such as the Sverdrup interiors of the gyres, the acoustic wave guide usually prevents the acoustic rays from sampling layers below 3000 m or 4000 m depth. Thus, the deep oceanic layers, those still most unknown, will remain unprobed even by tomography that may fail also in important oceanic regions such as the western boundary currents.

Other important data sets available through forthcoming experiments such as the World Ocean Circulation Experiment (WOCE) will consist mainly of long hydrographic sections, highly localized in space and very asynoptic in time. Wide regions of ocean will remain unmeasured between such sections. Thus, in the foreseeable future, oceanographers will still have to rely heavily on very localized and sparse clusters of moorings for long time-series measurements of currents, temperature, and pressure. The previous constraints, imposed by the nature of the existing and expected oceanographic data sets, define the two most challenging dynamical issues to be attacked by the oceanographic data assimilation problem, namely the question of advection of information (Thompson, 1961) and the trade-off between different measured variables (Charney *et al.*, 1969). These two issues were reviewed in the meteorological context by Ghil (1989) and are discussed here in the next section.

6.1. Role of Dynamics in Oceanographic Data Assimilation

6.1.1. Propagation of Information

The limitations of oceanographic data sets discussed earlier will give to the numerical models of the ocean circulation a much more critical role than their meteorological counterpart. The numerical model used must act as a dynamical interpolator or extrapolator from the vertical layer, or the volume where data are available, to the other vast regions of the ocean interior void of measurements. Advection and propagation of information in physical space produces exchange of information in the frequency–wavenumber domain between different time and space scales, ranging from the mesoscale, the weather of the ocean, to the quasi-steady component of the ocean circulation, or ocean climate. Thus, the following two major questions must be addressed through data assimilation:

(1) Can information provided only at the sea surface be transferred dynamically into the deep oceanic layers, thus reconstructing the deep circulation?

(2) Can information provided only locally, in limited oceanic regions, be transferred to ocean areas far away from the data-dense region, but dynamically connected to it? Which time and space scales are better estimated through the assimilation of local data?

One wishes, of course, to answer these questions *before* the actual observing systems, scheduled to be implemented in the early to mid-1990s, are in place. A systematic approach to doing so is provided by observing systems simulation experiments (OSSE). Such OSSE were developed and conducted by meteorologists in preparation for various field experiments of the Global Atmospheric Research Program (GARP; see Section 2 here and Bengtsson, 1975, Chapter 5). In OSSE, a model run is used to provide a history tape of nature. From this simulated history, measurements are taken, in accordance with the observing pattern—in space and time—of the system or systems under consideration. These observations are then assimilated into a model run with different, erroneous initial data. If the assimilation of these observations is done with the same model as the one used to produce the nature run, one speaks of an identical-twin experiment.

The purpose of OSSE experiments is to answer the observability question of estimation theory for the fluid system in hand: do the observations provided determine the state of the system, asymptotically in theory and over a reasonable amount of time in practice (Bucy and Joseph, 1987; Cohn and Dee, 1988; Ghil, 1980; Miller, 1989)? If not, can the observing system be modified at an acceptable cost so as to yield an affirmative answer? Identical-twin experiments are only a first step in the right direction: meteorological experience tends to indicate that their results are overly optimistic. Since the actual data from nature are not available at the time an OSSE is conducted, it is desirable to use at least a history tape from a different model than the one with which the data assimilation is carried out. This provides a simulation of the discrepancy between any model and nature and its effects on forcing the model with data towards the right solution, the one that nature provides.

Most OSSE carried out so far in oceanography have been of the identical-twin type. Still, their results provide valuable information on the benefits of various observing systems and data acquisition rates. Question (1), about propagation of surface information to depth, was first addressed in the pioneering works of Hurlburt (1986) and Thompson (1986), who made use of two-layer models of the Gulf of Mexico. The method used in the two studies just discussed was direct insertion of the observations into the numerical model (see Sections 5.1 and 6.3.2), and the simulated altimetric data were provided at every grid point in space. Kindle (1986) assimilated simulated altimeter data along the satellite tracks in a one-layer model of the same region.

In the active, two-layer PE model with a free surface, the surface-layer pressure p_1 provided by the altimeter is simply related to the sea level η by

$$p_1 = g\eta \qquad (6.1)$$

Using this two-layer model, Hurlburt (1986) focused on the dynamic transfer of information from the surface to the deep layer, demonstrating the success of the numerical ocean model used in recreating the deep circulation.

An extensive series of simulations was carried out, covering a wide variety of dynamical regimes. In Fig. 16, we show the evolution in time of the global normalized rms difference between the assimilation experiment and the control run (the reference ocean) for the pressure of the surface layer p_1, the pressure in the second layer p_2, and the pycnocline depth anomaly h_1 (i.e., its deviation from the overall spatial mean) for an experiment with strong baroclinic instability. Three update intervals were considered: 40 days (upper panel), 30 days (middle panel), and 20 days (lower panel).

If the assimilation is successful, it forces the dynamic evolution to converge to the reference ocean and the degree of success is measured by the rate of decrease of the rms errors. The initial error is not decreased when

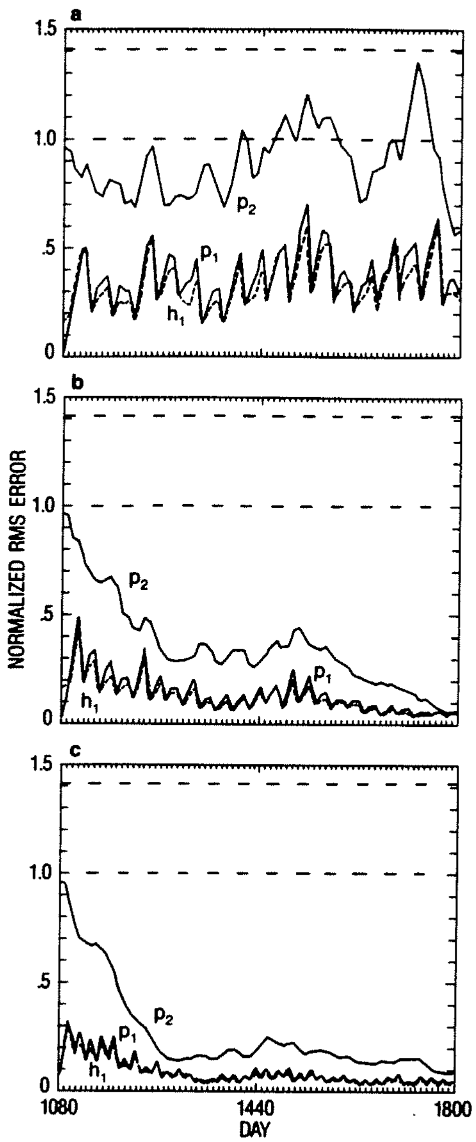


FIG. 16. Normalized rms error versus time for p_1, p_2, and h_1 (dashed lines) forecasts of model experiment *T*1 initiated at day 1080 and with update intervals (a) 40 days, (b) 30 days, and (c) 20 days. The temporal mean (climatology) was used as the initial state for p_2. The rms error is for the whole domain and the normalization factor is the standard deviation of the true field at the same time. The dashed lines at 1 and $2^{1/2}$ indicate the error for a flat field having the true areal mean (zero over the whole domain) and the error for an uncorrelated field with the same variance as the true field, respectively (from Hurlburt, 1986).

updating every 40 days, since the error growth rate due to baroclinic instability is larger than the convergence rate due to data forcing. The dominant time scale for baroclinic eddies in the model was of 57 days, and the important result emerging from Fig. 16 is that approximately two updates per eddy cycle are required for the assimilation to be successful and provide convergence to the reference ocean; the convergence is clearly accelerated further when updating every 20 days. More recently, Hurlburt *et al.* (1990) have used statistical inference to determine weakly correlated subthermocline fields from surface altimeter data.

Kindle (1986) addressed the sampling strategies for a satellite altimeter using a one-layer, reduced-gravity, shallow-water model of the Gulf of Mexico. The major limitations of this model are the absence of baroclinic instability and the fact that there is a one-to-one correspondence between sea-surface height and pycnocline depth. He examined the spatial sampling requirements for the accurate resolution of oceanic eddies. The main result is that an oceanic eddy can be adequately mapped when the altimeter-track spacing equals the radius of the outer contour, and when both ascending and descending tracks are used. The study, however, deals with a single stationary eddy, circular or irregularly shaped, and not with a turbulent mesoscale field where eddies are rapidly moving and interacting.

Thompson's (1986) study focused instead on the geoid error as it affects the assimilation of altimeter data for mesoscale ocean prediction. Assimilation of simulated altimeter data into a quasi-geostrophic eddy resolving ocean model was also carried out by Marshall (1985b), DeMey and Robinson (1987), Verron and Holland (1989), and by Holland and Malanotte-Rizzoli (1989). In QG dynamics with a rigid lid, the surface height variations δh provided by altimetry are related to streamfunction variations $\delta\Psi$ by

$$\delta\Psi = \frac{g}{f}\delta h \tag{6.2}$$

Holland and Malanotte-Rizzoli (1989) examined the space-time resolution to be provided by the forthcoming Topographic Experiment (TOPEX) altimetry according to the two alternatives proposed originally, of a 10-day or 20-day repeat period. These choices correspond to global coverages with spatial resolution in midlatitudes of roughly 2.8° of latitude and longitude, for a track separation of 280 km in the case of a 10-day repeat orbit, and 1.4°, with track separation of 140 km for a 20-day repeat orbit. The decision has now been made to adopt the 10-day repeat period.

When the altimetric data are assimilated along the actual tracks, that is only at the track grid points and at the actual time of arrival, by a nudging technique (see Section 6.3.2), assimilation results achieved with the satellite repeat periods of either 10 or 20 days are about equally unsatisfactory for improving the model estimates of the circulation. The residual rms errors after six years of time-continuous assimilation are between 60 and 70% of their initial values for both repeat periods. The results for this assimilation method show that under the best of conditions, of a perfect model and errorless data, a single satellite makes only small improvements in rms field estimates, and it cannot reconstruct the details of the mesoscale eddy field. This is clear when comparing Figs. 17 and 18 in which altimetric data are inserted

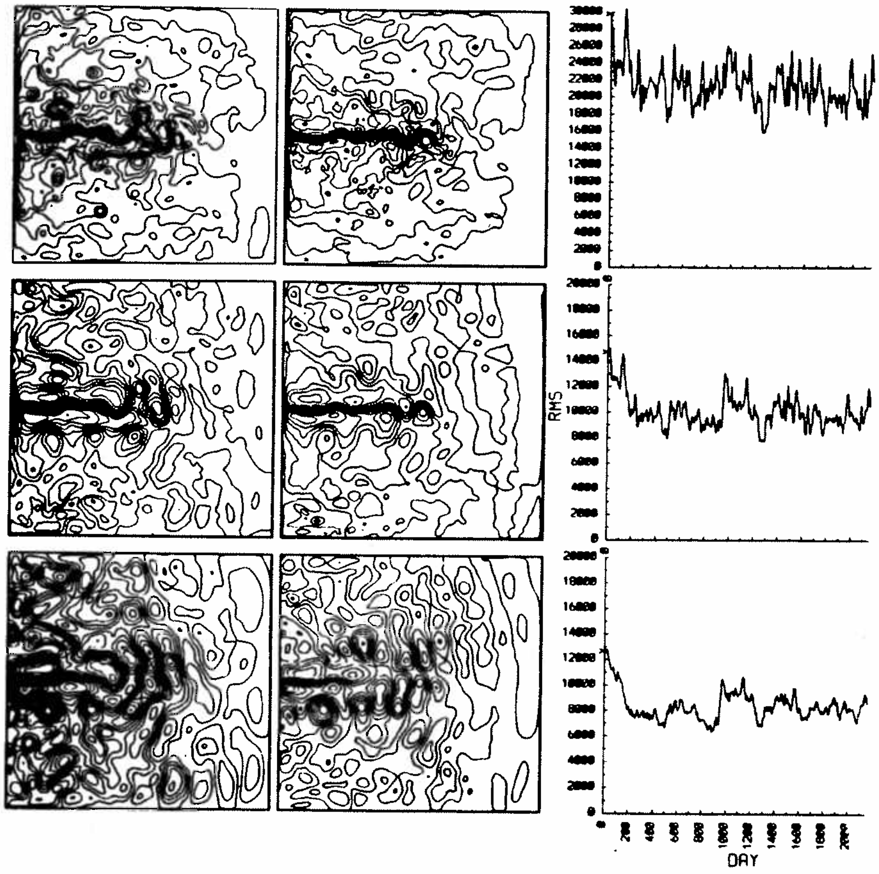


FIG. 17. Results of a TOPEX-type assimilation experiment in which the data are inserted on tracks separated by 140 km in a realistic time sequence, each track repeated every 20 days. The along-track component of vorticity is assimilated using a Gaussian time function to weight the data symmetrically about the actual time of arrival. A decorrelation time constant $\tau = 2$ days has been used. The upper, intermediate, and lower layer results are shown from top to bottom, (i) left panels: the control-run streamfunctions at day 2160; (ii) middle panels: the assimilation run streamfunctions at day 2160; (iii) right panels: the time evolution of the global rms differences between the control- and assimilation-run streamfunctions (from Holland and Malanotte-Rizzoli, 1989).

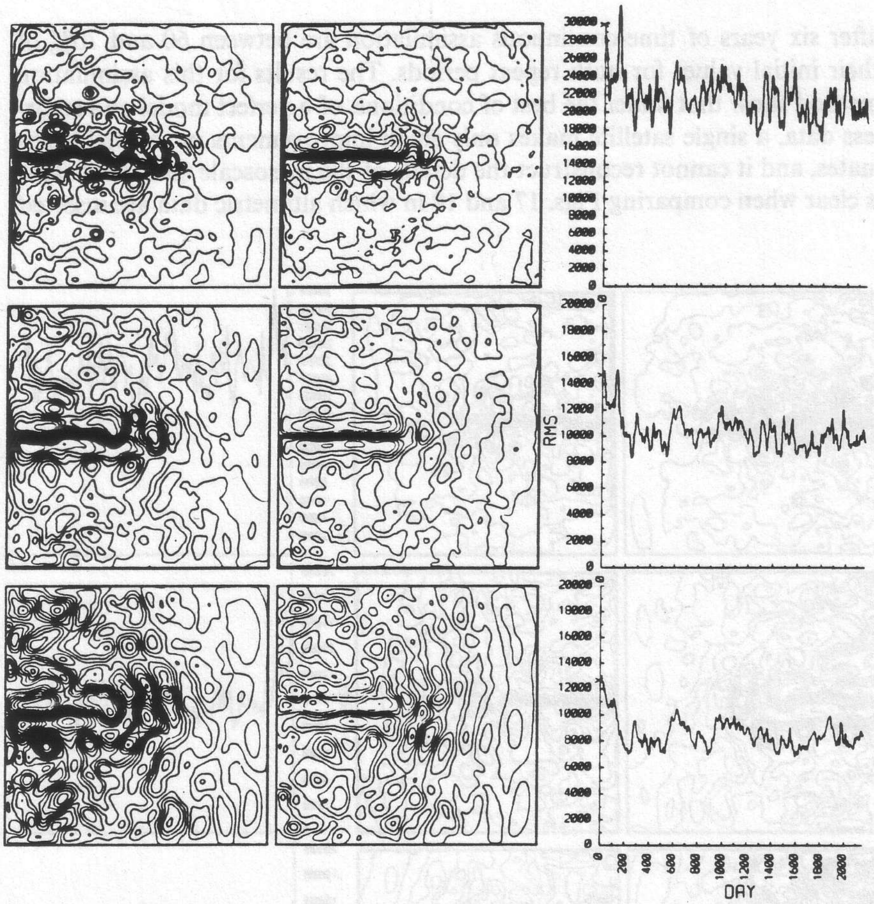


FIG. 18. As in Fig. 17, but for a track separation of 280 km and a repeat time of 10 days (from Holland and Malanotte-Rizzoli, 1989).

on tracks separated by 140 km, according to a 20-day repeat period (Fig. 17), and by 280 km, according to a 10-day repeat period (Fig. 18).

Holland and Malanotte-Rizzoli also examined the space-time resolution issue by providing altimetric data as a gridded map to the model (in the real case, one optimally interpolated map every T days, if T is the repeat period). Two strategies were followed. First the gridded map had the same space resolution as the model, i.e., data were assimilated at every model grid point, but the time interval between successive maps was changed from 2 to 10 and then to 20 days. Second, gridded maps were assimilated continuously in time but the spatial resolution was progressively coarsened, passing from track spacing of 42 to 99 km, and to 198 km. The results are much better when a gridded map is assimilated; in this case, a finer spatial resolution is more critical for the success of the assimilation than increased time sampling.

Some cautionary comments are in order at this point. First, one must differentiate among oceanic regions with different dynamics in order to assess the prospective usefulness of single-satellite coverage. In the previously mentioned experiments, the gyre situation is enlightening since it is clear from visual examination of the streamfunction fields that the relatively smooth Sverdrup interior is sampled reasonably well, with the space-time sampling strategies discussed earlier, and can thus be reproduced reasonably well by the assimilation. The Gulf Stream system, on the other hand, is not well sampled, being characterized by relatively small-scale and rapidly-changing mesoscale phenomena.

Still, an equally energetic mesoscale eddy field, but with different statistics, can be adequately sampled and thus successfully mapped through assimilation even with a single satellite. Altimetric data from GEOSAT during the period of 18 December 1986 to 18 September 1987 were assimilated by Holland (1989) for the Agulhas Retroflection Region south of South Africa. There the mesoscale eddy field, characterized by longer space and time scales, is adequately sampled by GEOSAT, and the assimilation experiments were quite successful. Verron (1990) has investigated the issue of sensitivity to orbital parameters when assimilating altimeter data into ocean models.

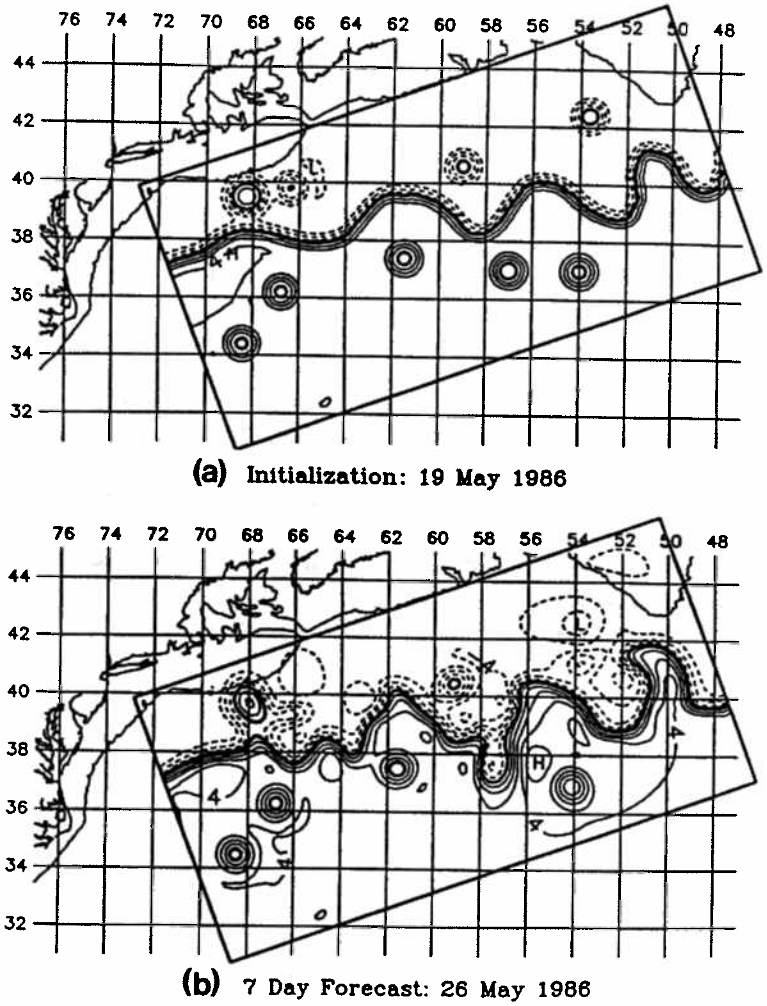
The assimilation of altimetric data, i.e., surface pressure, has been carried out mainly using two types of techniques, nudging and direct insertion (see Section 6.3.2). Holland and Malanotte-Rizzoli (1989), Holland (1989), and Verron and Holland (1989) used the nudging technique. Direct insertion of top-layer streamfunction information was used by Berry and Marshall (1989), who describe the dynamical mechanism transferring the surface information into the deep layers in a quasi-geostrophic (QG) two- and three-layer model. They show that the correction made to the surface-layer streamfunction through the direct insertion of the observed values provides an additional interfacial velocity w between the top two layers that acts like a wave-maker for the second layer and forces the deep circulation to converge towards the true, but unobserved pattern of the reference ocean. This dynamical mechanism is quite efficient for a two-layer model, since the corrected value of the surface streamfunction appears directly as a forcing term on the right side of the ω equation for the interfacial vertical velocity w_{12}. In multilayer models, however, such a correction term does not force directly the interfacial velocities $w_{i,i+1}$ of the deeper layers, $i \geq 2$. These, accordingly, spin up to the reference ocean very slowly, and rms errors decay over a multiyear time scale (Berry and Marshall, 1989).

The nudging technique, however, provides an immediate response of the deep layers to the constraint imposed by the knowledge of the surface pressure and hence relative vorticity; it is limited, therefore, only by the time taken to nudge the surface fields towards the observations. The rms differences between the assimilation experiments and the control run show the same *e*-folding decay time scale of roughly six months in a multi-layer, as in a two-layer model (Holland and Malanotte-Rizzoli, 1989). Thus, for multi-layer models, the nudging technique seems to be much more efficient, given surface data only, than direct insertion of these data in the upper layer only, at least in QG models. We shall return to the dynamical mechanisms underlying the nudging versus direct blending techniques in Section 6.3.2. An explanation for the differences given by Haines (1991) is discussed along with a new method of direct insertion, which has important similarities to the nudging technique.

A different type of transfer of information from the surface to the deep layers is that exploited in the assimilation studies of Robinson and collaborators (Robinson and Leslie, 1985; Robinson *et al.*, 1986, 1987, 1988, 1989; Robinson, 1987; Mooers *et al.*, 1987; Robinson and Walstad, 1987; DeMey and Robinson, 1987). In these studies, sea-surface infrared temperature images are used as initial and update data for an open-boundary regional QG model through the process of reconstruction of ocean features (feature model), such as the Gulf Stream mean path and warm- or cold-core rings. The surface information is projected along the vertical onto the deeper layers through the Empirical Orthogonal Functions (EOFs), which characterize the second-order statistics (covariance matrix) of the flow fields in the region being studied. This assimilation procedure has led to the development of Gulfcast (Robinson *et al.*, 1989).

Gulfcast is an analysis and forecast system for the Gulf Stream meander and ring region consisting of the Harvard dynamical open-ocean model (Robinson and Walstad, 1987) and an observational network (Robinson *et al.*, 1989). The network is comprised of remotely sensed sea-surface temperatures, obtained every other day, and of critically located air-dropped expendable bathythermograph (AXBT) data, obtained once a week. The AXBT drops have the dual role of verifying the previous seven-day forecast and of helping determine the initial state for the next forecast. The Gulfcast system was tested under a wide range of circumstances. The phenomena predicted by the forecasting procedure include Gulf Stream meander growth and propagation, straightening out a previously meandering stream, ring formation, and ring-stream interactions and movements. Figure 19 shows a characteristic forecast experiment, with the 100-m streamfunction field used as initial data (Fig. 19a) and the same field after a seven-day model forecast (Fig. 19b).

Question (2), concerning the transfer of information by advection or wave propagation from data localized in space to other regions of the ocean, has been addressed in tropical oceanography by Miller and Cane (1989), who assimilated real tide-gauge data from six island stations into a simple model



(a) Initialization: 19 May 1986

(b) 7 Day Forecast: 26 May 1986

FIG. 19. (a) Initial 100-m streamfunction field on 19 May 1986. (b) 100-m streamfunction field after a seven-day forecast on 26 May 1986 (from Robinson *et al.*, 1989).

of the tropical Pacific. Using the Kalman filter, this small amount of data produced monthly sea-level height anomaly maps for the equatorial wave guide with reduced rms error and increased spatial detail, even away from the data points (see Section 6.3.3 for details on this study).

In midlatitude oceanography, this question was addressed by the two papers of Malanotte-Rizzoli and Holland (1986, 1988). Holland's (1978) QG model of the general circulation was used to study the effect of local hydrographic, or tomographic, sections in improving the model estimates for ocean areas far away from the data region, but dynamically connected to it. Malanotte–Rizzoli and Holland used a simple data insertion technique, weighting observations by their distance from the grid point being updated, as discussed in Section 6.3.2, and allowed for model errors in this OSSE-type study.

They found that a local section can be quite effective in determining the flow in far away regions if the model is very simple, steady, and quasi-linear (Malanotte–Rizzoli and Holland, 1986) and that the most effective sections are meridional, long and far away from the ocean's western boundary (see also Parrish and Cohn, 1985, and Section 5.3.1 here). On the other side, for fully time-dependent and eddy-resolving simulations, a simple data section is completely ineffective, unless decade-long time series of measurements are available (Malanotte–Rizzoli and Holland, 1988). In the latter case, a considerable improvement in the estimate of the model's climatology, i.e., its long-time average circulation is still obtained.

The question of the effectiveness of oceanographic data collected in sparse localized clusters of mooring arrays has been addressed by Malanotte–Rizzoli and Young (1991). This question is motivated by the forthcoming availability of three-year long time series of current velocities and temperatures measured at two clusters of current meter moorings as part of the already-mentioned SYNOP experiment, which focuses on process studies in the Gulf Stream system. The two current-meter mooring arrays are located one east and the other west of the New England Sea Mountain Chain (Rossby, 1990).

Malanotte–Rizzoli and Young (1991) simulate the two localized clusters in a Semi-Spectral Primitive Equation (S.P.E.M.) model with active thermodynamics originally developed by Haidvogel *et al.* (1991). They use the nudging technique discussed in Section 6.3.2 to relax the prognostic variables (u, v, p) towards their observed values with an identical-twin approach. The results are quite encouraging even though the data are provided only at a very small number of model grid points: the assimilation process is quite successful in reconstructing the jet behavior of the control run, including the bending of meanders and ring pinch-off, in the region downstream of the mooring arrays, especially between the two clusters, after only two months of continuous time assimilation. This is due to the advection of the assimilated information downstream from the measurement points, provided by the strongly nonlinear, idealized mean flow representing the Gulf Stream jet.

6.1.2. Trade-Off between Variables

Few investigations have been devoted until now in oceanography to the problem of the relative usefulness of different variables for the data assimilation process, while in meteorology a substantial literature on this problem exists (Charney *et al.*, 1969; Smagorinsky *et al.*, 1970; Ghil, 1989, and references therein). Most of the existing studies are in the tropical ocean. Moore *et al.* (1987) investigated the effect of updating models of the Indian Ocean using simulated temperature or velocity data. Temperature data were found to be better than velocity data in determining the model state. Further experiments showed, however, that increasing the model's diffusion and decreasing its eddy viscosity results in velocity data determining the state better. These results were ascribed to changes in the energy distribution from one case to the other, with the proportion of kinetic energy being greater in the later experiments. Simulated data from the proposed Tropical-Ocean/Global-Atmosphere (TOGA) Indian Ocean XBT network were also assimilated. The importance of salinity for data assimilation in tropical ocean models was first considered by Cooper (1988).

Other assimilation studies related to the tropical ocean are those by Leetmaa and Ji (1989), who have been using an operational ocean model in the hindcasting mode, Anderson and Moore (1989), Moore and Anderson (1989), and Miller (1990). Moore and Anderson (1989) formulate initial data for and then update a one-layer reduced-gravity model of the tropical ocean at regular intervals with XBT observations of the 16°C isotherm collected as part of the Tropical-Ocean/Global-Atmosphere (TOGA) ship-of-opportunity program. The XBT data were interpolated in space by combining each observation with the model first guess, using a scheme based on the successive correction method (SCM) (see Section 2, especially Table I, and Section 5.1, especially Table II), as used by Moore *et al.* (1987).

The method is a special form of the linear unbiased data assimilation scheme [Eq. (4.8b)] and can be written explicitly in this case as

$$h_k^{\mathrm{a}} = h_k^{\mathrm{f}} + \sum_{i=1}^{N} \alpha_i (h_i^{\mathrm{o}} - h_i^{\mathrm{f}}) \Bigg/ \left(\alpha_p + \sum_{i=1}^{N} \alpha_i \right) \tag{6.3}$$

where *h* is the depth of a given isotherm and the superscripts o, f and a indicate, as in Section 4, observed, first-guess, and analyzed values, respectively. Subscript *i* refers to an observation point and *k* to the model grid point under analysis; α_i is a weighting function assigned to each observation point *i*, and α_p is an additional weight that can be assigned to the first-guess field. Figure 20 shows the normalized rms errors (o-f) and (o-a) in the displacement of the model 16°C isotherm about its 200-m mean depth in four regions of the equatorial Pacific. The effect of the data assimilation process in reducing the rms errors is especially pronounced in the Eastern Pacific, where the control solution is substantially modified. Miller (1990) carried out much the same exercise with a K-filter, obtaining detailed analysis error maps for different observational arrays (see his Fig. 11), using the same model of the tropical Pacific as Miller and Cane (1989).

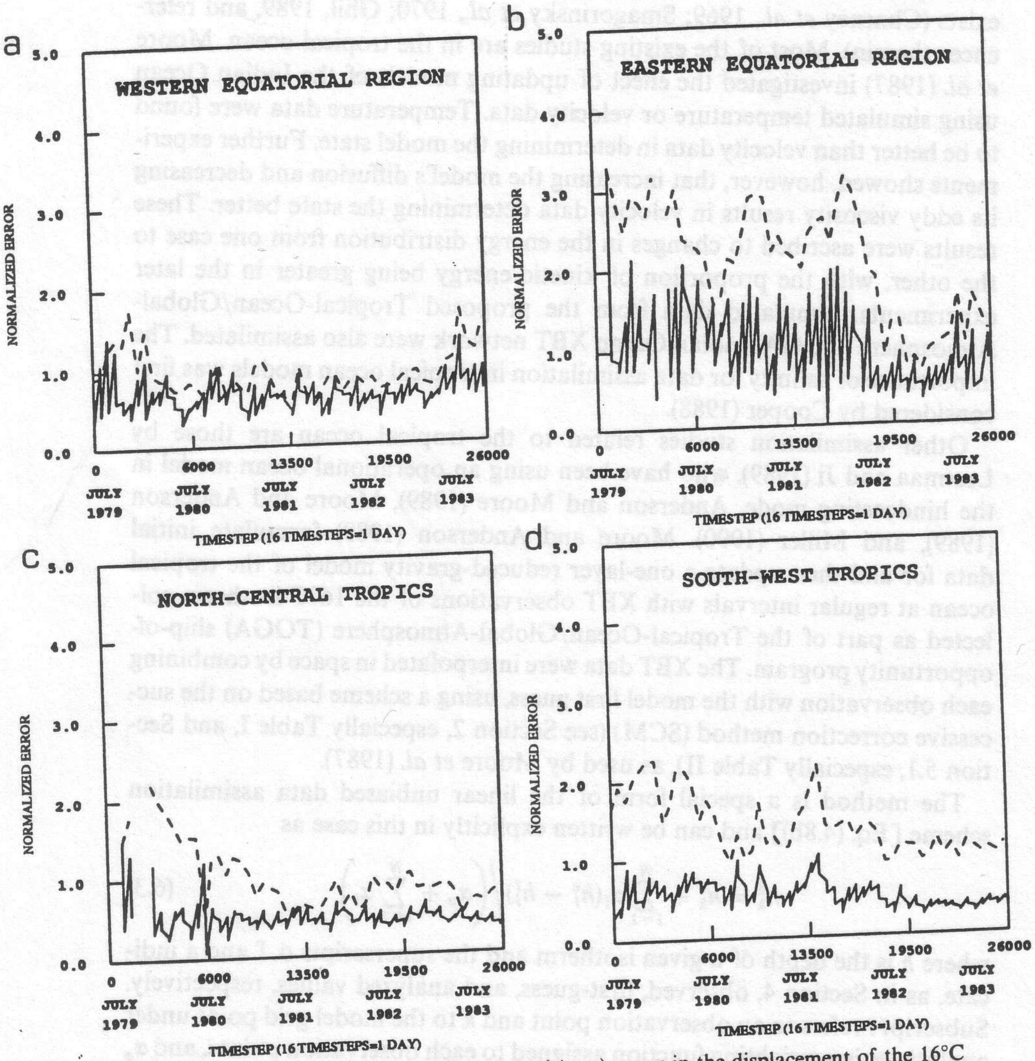


FIG. 20. The normalized rms $(o-f)$ errors and $(o-a)$ errors in the displacement of the 16°C isotherm about its 200-m mean depth in a reduced-gravity model of the tropical ocean. The control integration (dashed curve) and an integration during which the model is updated every month with XBT data (solid curve) are both shown. Data points are plotted just before and just after each model update. The errors in each figure have been normalized by the variances (from Moore and Anderson, 1989).

The question of the effectiveness of different data types in midlatitudes was addressed by Malanotte-Rizzoli *et al.* (1989) in an eddy-resolving multilevel PE model. Two types of data fields were compared:

1. Knowledge of the depth-integrated flow only, i.e., of the barotropic external mode. This type of information would correspond to measurements of the total transport provided, for instance, by tomographic arrays across different sections of the Gulf Stream jet.

2. Knowledge of the density field only, i.e., of the baroclinic, internal modes or, equivalently, of the velocity shear through the thermal-wind relationship. This information would be provided by traditional hydrographic surveys covering the region of study.

Notice that the knowledge of the barotropic streamfunction provides information that is constant with depth, i.e., a 2-D field of data, while the density field provides a 3-D data set. The assimilation experiments for data types (1) and (2) thus allow a comparison of the effectiveness of 2-D horizontal versus fully 3-D data maps.

The results are illustrated by showing the rms errors in different fields, both the global rms over the entire 3-D network and the rms for each of 5 horizontal levels considered in the model. Figure 21 shows the rms errors when assimilating the density field ρ only, at every model grid point; Fig. 22 shows the results for the assimilation of the barotropic streamfunction Ψ_B only, also at every model grid point.

Experiments with progressively coarser horizontal resolutions were also carried out. The following conclusions can be drawn from this set of assimilation experiments:

(1) The knowledge of the depth-integrated flow Ψ_B, or total transport, is much less effective than knowledge of the interior density field ρ. In the second case, the decrease in rms error, i.e., the convergence to the reference ocean, is much faster. (2) If the baroclinic structure is known, i.e., if data are inserted three dimensionally, a decrease in the horizontal resolution of data insertion is not very deleterious. Coarse horizontal resolution can be reached before significantly worsening the model estimates. (3) If only the barotropic component is known, a decrease in the horizontal resolution has an immediate and profound effect on the assimilation: the rms errors sharply increase and the assimilation run diverges from the reference ocean. (4) Even with dense insertion at every grid point of the barotropic flow, the errors in the deep layers always show an increasing trend, which indicates a worsening of the estimate of the deep circulation, as evident from the examination of Fig. 22.

The simplest and natural interpretation is that the assimilation of the density field is much more successful because a 3-D rather than a 2-D data

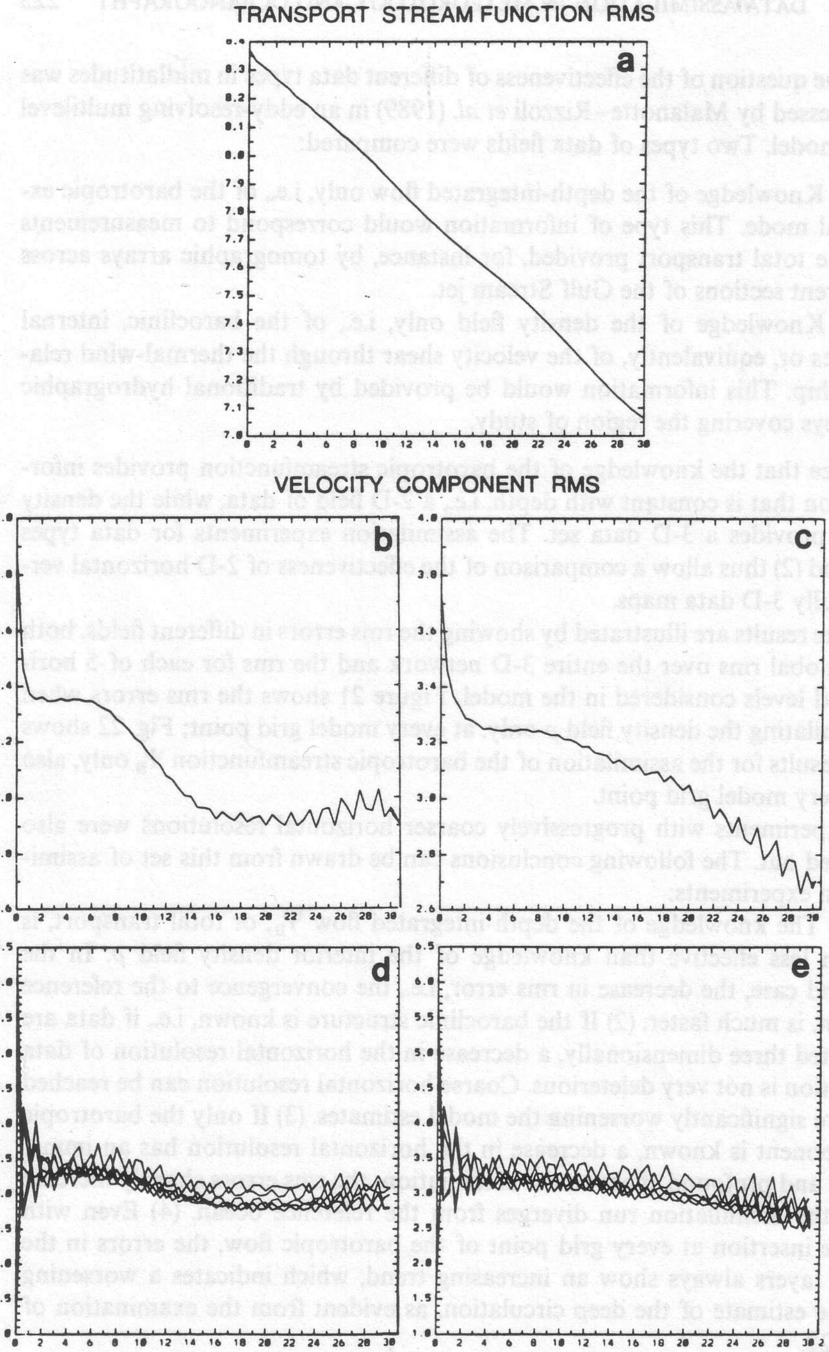
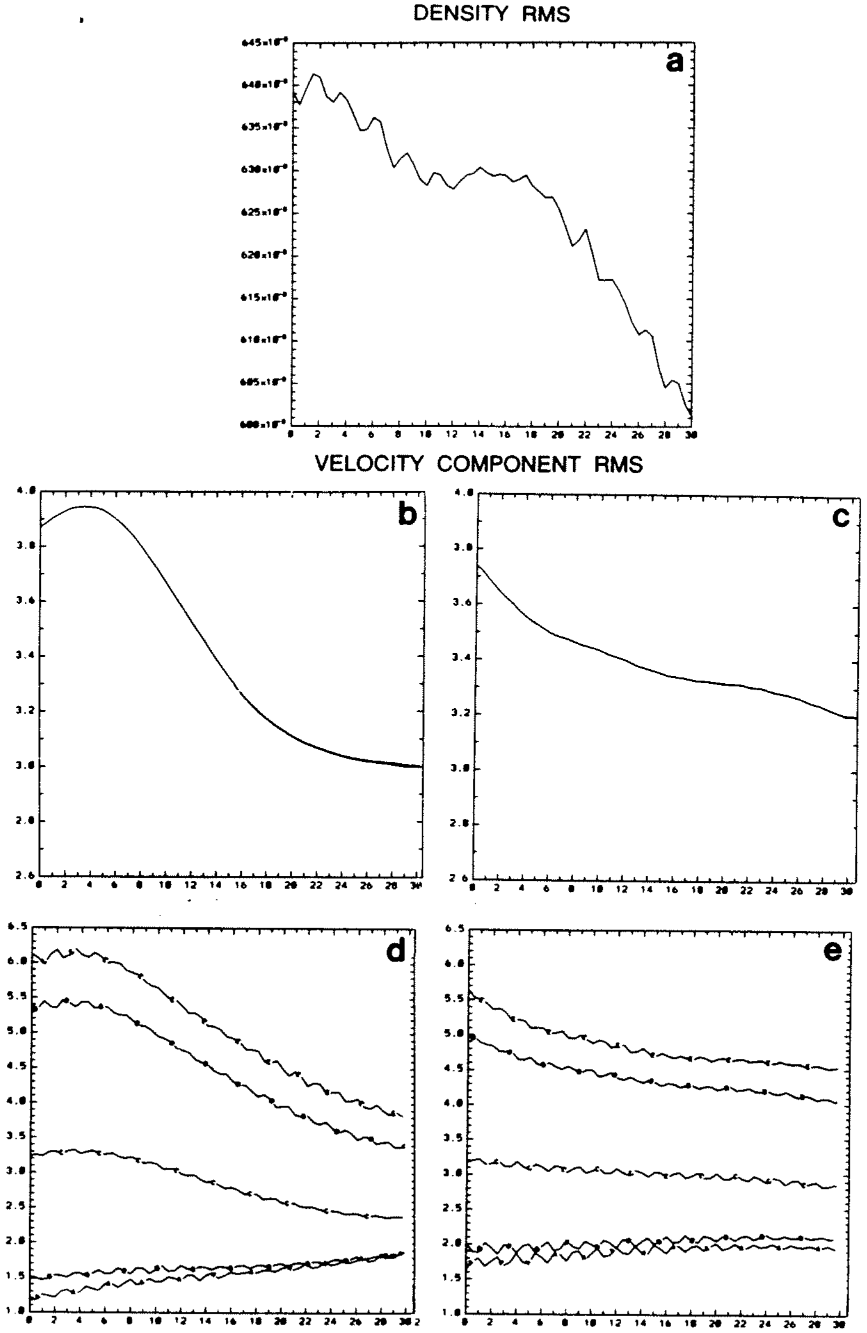


FIG. 21. Assimilation of the density field in an eddy-resolving multi-level PE model. Time evolution over one month of rms errors. Abscissa: time in days. Ordinate: (a) units scaled by 10^{-12} cm^3 sec^{-1}, and (b)–(e) units in cm sec^{-1}. Vertically integrated rms error for: (a) transport streamfunction Ψ_B (maximum value, 8.4×10^6 m^3 sec^{-1}); (b) *u*-velocity component; (c) *v*-velocity component. Rms error by level for: (d) *u*-velocity component; (e) *v*-velocity component (from Malanotte-Rizzoli *et al.*, 1989).



DENSITY RMS

VELOCITY COMPONENT RMS

FIG. 22. Assimilation of the barotropic streamfunction Ψ_B in the same model as used for Fig. 21. Time evolution over one month of rms errors. Abscissa: time in days. Ordinate: (a) density in gcm^{-3} (b–e) velocity in cm sec^{-1}. Vertically integrated rms errors for (a) perturbation density field ρ, (b) u-velocity component, (c) v-velocity component; rms errors by level for: (d) u-velocity component; (e) v-velocity component. Letters on separate curves indicate different levels: (A) 3600 m (deepest level), (B) 2400 m, (C) 1600 m, (D) 800 m, (E) Surface (from Malanotte-Rizzoli *et al.*, 1989).

set is provided to the model, i.e., more data are used (see also Ghil, 1989, Section 4.3.3). However, a rationalization can also be made in the context of simple geostrophic adjustment theory. When providing baroclinic information, the data insertion process must force the (unknown) barotropic mode. The response time of the model will then be the short adjustment time required by the (fast) barotropic component of the flow. That this is so emerges clearly from Fig. 21, in which the baroclinic component is assimilated, and a great percentage of rms error decrease occurs during the first two-to-three days of assimilation. Alternatively, when only the barotropic mode is known, the baroclinic component of the flow must be forced and the adjustment time of the model will occur on the (long) time-scale of the (slow) baroclinic modes. This is also clear in the much reduced rate of error reduction of Fig. 22, when only Ψ_B is assimilated, and comparable error decreases are achieved only after one month of continuous assimilation.

6.2. Initialization Problem in Oceanography

The issue of model initialization was addressed briefly in Section 2 and more extensively in Section 5.2, where meteorological applications were emphasized. We now summarize the oceanographic applications, referring to Section 5.2 for the details of the different procedures.

The problem of initial conditions for ocean GCMs in relation to data assimilation and forecasting was first addressed in the tropics by Philander *et al.* (1987). Initialization in the tropical ocean was studied by Moore (1990). The initialization problem was investigated systematically in the oceanographic context for midlatitude systems by Malanotte-Rizzoli *et al.* (1989), who used the NNMI procedure introduced by Machenhauer (1977) and connected by Leith (1980) to quasi-geostrophic theory (see also Daley, 1981). The method applied by Malanotte-Rizzoli and colleagues represents a computationally efficient, first-order approximation of Machenhauer's NNMI, using a QG streamfunction, and consists of the following two steps:

(1) Geostrophic and hydrostatic initialization,

$$\Psi|_{t=0} = \Psi^{(o)}$$

$$u^{(o)} = -\Psi_y^{(o)}, \qquad v^{(o)} = +\Psi_x^{(o)}, \qquad w^{(o)} = 0$$

$$\phi|_{t=0} = \phi^{(o)} = f_0 \Psi^{(o)}$$

for the pressure field, and

$$\rho|_{t=0} = \rho_0 = -\frac{\bar{\rho}}{g}\phi_z^{(o)} \tag{6.4a}$$

where ρ_0 is the perturbation density field for a constant midlatitude value of the Coriolis parameter, and the superscript (o) indicates geostrophic fields, valid at initial time $t = 0$.

(2) Quasi-geostrophic tendency,

$$\begin{aligned} \Psi_1 &= \Psi^{(o)};\ \phi_1 = f\Psi_1 \\ u_1 &= -\Psi_{1,y};\ v_1 = +\Psi_{1,x} \end{aligned} \tag{6.4b}$$

and

$$w_1 = \frac{1}{f_0 S}\frac{D}{Dt}\Psi_z^{(o)}$$

where $S = \bar{N}^2(z)/f_0^2$. The subscript 1 indicates fields at the first time step. Step (2) constrains the horizontal velocity field to be given by geostrophy and the vertical velocity to be quasi-geostrophic. This second step is equivalent to letting the tendency of the fast gravity modes vanish, as discussed in detail in Section 5.2.

The following conclusions were drawn from the initialization experiments of Malanotte-Rizzoli *et al.* (1989): (1) In the absence of initialization, a large amount of imbalance in the initial data is required to produce substantial levels of internal gravity-wave noise that radiate away from the initial jet. (2) First-order balancing for some of the initial fields only is sufficient to suppress the greatest part of gravity-wave noise. In this balancing, the barotropic velocity or the velocity shear are evaluated geostrophically from the barotropic streamfunction or the density field, respectively. (3) A geostrophically balanced initialization [Step (1), Eq. (6.4a)] for all the fields is sufficient for long evolutions of the jet without any significant appearance of gravity waves. Hence, there does not seem to be a need for constraining the model tendency in the initialization, as given by Eq. (6.4b).

These results obtained by Malanotte-Rizzoli *et al.* (1989) confirm the effectiveness of the simpler balanced initializations carried out by Hurlburt (1986), Kindle (1986), and Thompson (1986) (see also Section 6.1.1) and the results of Carter (1989), who used Lagrangian-float data. In particular, Thompson's (1986) results showed the robustness of geostrophic initializations under the noisy conditions provided by errors from a poorly known geoid. Hurlburt (1986) also showed that a simple geostrophic initialization is quite appropriate as long as the determination of the subsurface pressure field can be properly made. However, forecasting experiments for the Gulf Stream system with a PE model indicate that poorly known deep pressure fields at the initial time are a major source of error in forecast accuracy, and that shocks from these imbalanced initial states are in part responsible for the degraded forecasts (Hurlburt *et al.*, 1990). Thus, a dynamical balance between the surface and subsurface initial pressure fields is an important requirement for accurate ocean forecasting.

In conclusion, the initialization problem does not seem to be as crucial an issue in large-scale oceanography as it is in NWP. Part of the reason may be ascribed to the difference between the two fluids, specifically their different characteristic Burger numbers and the related implications for energy distribution between waves, as discussed in Section 3.1.

6.3. Assimilation Methods

Two approaches to data assimilation are emerging in the oceanographic community. The first is the development and use of sophisticated assimilation techniques. The computational feasibility of this approach has limited its use so far to relatively simple dynamical models, with hundreds to thousands of variables. To this category belong (a) sequential estimation methods, introduced in Section 4.1, of which the K-filter is the prototype; and (b) variational methods, introduced in Section 4.2, especially those based upon the use of the adjoint equations. The meteorological applications of these two approaches were discussed in Sections 5.3 and 5.4, respectively. Their oceanographic counterparts will be presented in the following Sections 6.3.3 and 6.3.4.

The second approach focuses upon the use of more complex and realistic dynamical models, capable of simulating ocean processes in greater detail. In this approach, the data assimilation schemes are, per force, methodologically simple and computationally efficient. Two important schemes in the latter category are the blending and the nudging methods discussed in Section 6.3.2. The methods whose oceanographic applications are presented next are those based on optimal interpolation (OI), and in general on optimization schemes. The former stem most naturally from the practice of meteorological forecasting, the latter from geophysical inverse theory (see Section 1). The theory underlying OI was introduced in Section 5.1. A short review of its oceanographic applications is also given by Webb (1989).

6.3.1. Optimal Interpolation and Inverse Methods

Optimal interpolation is a simplified version of the K-filter in which the interpolation weights for observations are determined using an approximate form of the forecast error covariance matrix, cf. Eqs. (5.1)–(5.5). Optimal interpolation is a practical and internally consistent approach for treating a large set of heterogeneous observations, and it is at present the technique that produces the best results for objective analysis at one given time level in NWP. Several problems occur, however, when this method is applied to the temporal evolution of a nonlinear unstable flow. In fact the procedures now in operational use in meteorology proceed sequentially in time, but the weights assigned at successive analysis steps are still not entirely consistent with the evolution equations (Cohn *et al.*, 1981; Ghil *et al.*, 1982; Webb, 1989; Section 5.1 here).

Statistical interpolation in general tends to smooth the objectively analyzed fields excessively. Excessive smoothing is particularly troublesome for mesoscale models, since it may inhibit developments that are unlikely from a statistical point of view, just because of the rarity of their occurrences, but are very important to correctly simulate the system. A well-known example of such a relatively rare but important phenomenon is ring formation from the Gulf Stream jet. An interesting approach to avoid this excessive smoothing was proposed by Mariano (1990). He estimates the position of dynamical features with known characteristics, rather than grid-point values for complete fields.

We first recall that objective analysis, in general (see Section 2), is also widely used to analyze synoptic or quasi-synoptic data with climatology, rather than a model forecast, as a background field (Bretherton *et al.*,1976; Freeland and Gould, 1976; Gandin, 1963). Data from different time levels can be used in this way as well (Miyakoda and Talagrand, 1971; Kindle, 1986; McWilliams *et al.*, 1986). Various examples on the use of OI are also given in Wunsch (1989a).

OI as a form of objective analysis is most often used nowadays in conjunction with a dynamical model. Marshall (1985a) applied OI to the problem of determining the ocean circulation by assimilating satellite altimetry data and, at the same time, improving the geoid estimate. He used a barotropic (2-D) QG ocean model with the identical-twin approach, i.e., with model-simulated altimetric measurements. The technique was used in a simulation study of Gulf Stream variability in which the surface geostrophic flow, or ocean topography (OT) in Marshall's designation, is degraded by the noise introduced through the uncertainty in the geoid's estimate. Figure 23 shows results of this application.

In Fig. 23a, the true six-month time mean OT is shown as constructed by the control run of the identical-twin experiment. Figure 23b shows the six-month mean OT estimate obtained through data assimilation into the ocean model when updating continuously both the OT and the geoid estimate. The rms error of Fig. 23b with respect to 23a is only 4.1 cm. Figure 23c is also evaluated through the data assimilation process, but only the OT estimate is continuously updated, not the geoid. The comparison of the assimilation experiments shown in Figs. 23b,c with the true ocean of Fig. 23a clearly demonstrates that solving simultaneously for the geoid and OT is preferable. The analysis of Fig. 23c is visibly poorer than the analysis of Fig. 23b, and the rms error of Fig. 23c with respect to Fig. 23a is in fact 8.4 cm.

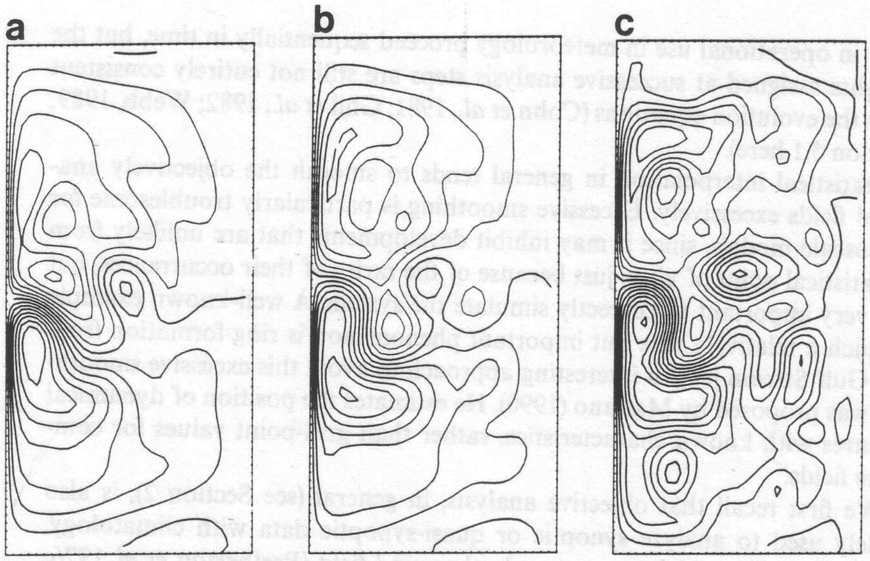


FIG. 23. Surface circulation (ocean topography, OT) evaluated by a quasi-geostrophic 2-D barotropic ocean model. (a) True six-months time mean OT from the control run. (b) Six-months time mean OT reconstructed in the assimilation experiment when continuously updating both the OT estimate and the geoid estimate. (c) Six-months time mean OT reconstructed in the assimilation experiment when continuously updating the OT estimate, but not the geoid estimate (from Marshall, 1985a).

Marshall (1985b) applied the method further to study the efficiency of different altimeter sampling strategies. White *et al.* (1990a,b,c) also used OI to assimilate first simulated and then GEOSAT altimetric sea-level observations continuously into a QG eddy-resolving ocean model. Multivariate statistical objective analysis of the OI type was also applied by Carton and Hackert (1989) to the circulation of the tropical Atlantic Ocean. Statistical regression techniques blended with a deterministic modeling approach and projection of the surface information onto the deep layers through vertical EOFs have also been used by Robinson and collaborators in the studies quoted in Section 6.1.1 in their use of the Harvard open-ocean QG model in different regional domains of the world ocean. The domains are typically of the order of 10 Rossby radii on the side, as the emphasis of the work of this group is upon the real-time prediction of the mesoscale eddy field.

The 4-D data assimilation approach so successful in meteorology for coping with data sparsity, which uses objective analysis techniques to blend the model-evaluated variables and the observations, was first applied to oceanographic problems in limited-ocean domains (Robinson and Leslie, 1985; Rienecker *et al.*, 1987). Clancy *et al.* (1988) applied the method to larger domains but used a simple mixed-layer model. More recently, Leetmaa and Ji (1989) have used the methodology for operational hindcasting of the tropical Pacific. The most recent oceanographic application of 4-D data assimilation has been made by Derber and Rosati (1989) to the global ocean using a high-resolution PE model. At the National Meteorological Center (NMC), experience with a quasi-operational data assimilation system for a noneddy-resolving model has shown that the largest source of errors is the poor knowledge of atmospheric fluxes (J. Derber, personal communication, 1990; see also Atlas *et al.*, 1987).

Optimization methods based upon the idea of minimizing a deterministic objective function, rather than an expected error [see Eqs. (4.5)–(4.7) and Sec. 5.4.1], were first used by Provost (1983) and by Provost and Salmon (1986) to estimate the geostrophic circulation from hydrographic data. These optimization schemes belong to the category of variational methods discussed in Sections 4.2 and 6.3.4. A nonlinear optimization method was developed by Schröter and Wunsch (1986) to study the effect of observation errors in determining solutions of nonlinear, finite-difference, steady models with one and two layers. They carry out a sensitivity analysis exploring the effect of uncertainties in the surface forcing function, i.e., the wind stress curl, as well as uncertainties in a model parameter, the bottom friction coefficient. The optimization again proceeds by minimizing a diagnostic objective function that represents some feature of the flow, such as the area integral of the potential or kinetic energy (nonlinear objective function) or the transport of the model's western boundary current (linear objective function).

The inverse techniques borrowed from solid-earth geophysics (Backus and Gilbert, 1967; Parker, 1972; Tarantola, 1987) and introduced into physical oceanography by Wunsch (1977) broadly belong to the category of optimization methods. First, the distinction between "inverse methods" and "inverse problems" has been pointed out by Wunsch (1989b), who gives a very simple example of a direct and an inverse problem in the context of the same model, the Poisson equation

$$\nabla^2\phi = \rho \tag{6.5}$$

The direct problem is the solution of Eq. (6.5) within a domain D, with boundary ∂D and boundary conditions $\phi = \phi_0$ on ∂D. This is the classical Dirichlet problem. The inverse problem would be to assign ϕ and its boundary conditions and solve for the source term ρ. An oceanographic inverse problem is Stommel's β-spiral (Stommel and Schött, 1977), which was originally solved by a direct method, least squares (even though Davis, 1978, showed the dynamical connection between the β-spiral and Wunsch's inverse problem).

Inverse methods are directed primarily at problems which in some way are ill-posed and which need to be regularized. For a review of steady and time-dependent ill-posed problems, see Wunsch (1989a). Here we shall give a simple example of a problem best solved by an inverse method. Consider the linear problem

$$\sum_{j=1}^{M} D_{i,j}p_j = b_i \qquad i = 1,\ldots,N \tag{6.6}$$

with unknowns p_j, $j = 1,\ldots,M$. Equation (6.6) has, in general, no solutions if $N > M$ (overdetermined case) and an infinity of solutions if $N < M$ (underdetermined case). In the overdetermined case, if the N equations are linearly independent, solutions can be defined only by discarding $(N - M)$ equations. In the underdetermined case, unique solutions can be found only by assigning additional criteria for their selection from the infinity of solutions.

A powerful inverse method to solve this problem in both cases is singular value decomposition (SVD) (Lanczos, 1961), widely applied to solve inverse problems in geophysics as well as oceanography. For a review of SVD theory and of inverse problems and methods, see Olbers (1989) and Wunsch (1989b). In the latter reference, a short history on the use of inverse methods in ocean circulation studies is also given. They were pioneered by Wunsch (1977, 1978) to study the general circulation of the North Atlantic and solve the problem of determining the classical level of no motion, i.e., to calculate reference-level velocities for the thermal-wind equations of motion. These studies used the SVD method, and further applications were later made by Wunsch and Grant (1982) and Fiadeiro and Veronis (1984). A number of linear inverse calculations of the ocean circulation in different ocean basins followed. A nonlinear inverse method for the general circulation was also proposed by Mercier (1986). The interested reader is referred to Wunsch (1989b) for a complete review.

We conclude here by mentioning that inverse methods have been successfully applied to two other oceanographic inverse problems. The first one is the tomographic problem introduced by Munk and Wunsch (1979), who used SVD in their original study. Statistical inverses based on OI were constructed and applied to acoustic-tomography data by Cornuelle *et al.* (1985). The second oceanographic inverse problem is the tracer problem, also formulated from this perspective by Wunsch (1988, 1989b). In Section 6.3.4, the tracer problem is treated with the adjoint method.

6.3.2. *Blending and Nudging Methods*

The blending technique is a highly simplified and localized version of OI, with purely empirical weights. At assigned times, the observed or forecast field variable f at a given grid point is replaced by a new variable f^{new} which is a blending of the two, f^{model} and f^{obs}

$$f^{\text{new}} = \alpha f^{\text{obs}} + (1 - \alpha) f^{\text{model}} \tag{6.7}$$

α is the weight assigned to the observed value [compare Eq. (4.2a)]. When $\alpha = 1$, the blending method reduces to direct insertion of the observed value in place of the model-predicted value (see Table II in Section 5.1). The direct insertion technique was used by Kindle (1986), Thompson (1986), and Hurlburt (1986), and Malanotte-Rizzoli and Holland (1986, 1988) also used it in their investigation of the assimilation of localized hydrographic sections (see Section 6.1.1), Insertion techniques were also applied by Moore *et al.* (1987) and Moore and Anderson (1989). Their method was a direct updating of the whole temperature or velocity fields (Section 6.1.2). A direct insertion technique was used by Malanotte-Rizzoli *et al.* (1989) in the study discussed in Section 6.1.2 and by Berry and Marshall (1989) in the assimilation of altimeter data. The latter work shows that a time of the order of the baroclinic Rossby adjustment time (i.e., of years) may be necessary with some insertion methods for surface information to spread into the deep layers and provide convergence of the deep circulation to the reference ocean.

The nudging technique, introduced for oceanographic data assimilation by Verron and Holland (1989) and by Holland and Malanotte-Rizzoli (1989) seems to be much more effective in reconstructing the circulation in the deep layers. As noted in Section 6.1.1, the rms error of the assimilation experiments of Holland and Malanotte-Rizzoli (1989) exhibited an *e*-folding decay time scale of about 6 months, in contrast to the much longer rms-decay time scales of Berry and Marshall (1989). Before discussing the reasons for the success of the nudging technique in comparison to the blending one, we first summarize its properties.

The nudging or relaxation scheme was first introduced in meteorology by Anthes and colleagues (Anthes, 1974; Hoke and Anthes, 1976). Following the original formulation by Anthes (1974), consider a prognostic variable f of the model measured at N observational stations. The equation for nudging can be written as

$$\frac{\partial f}{\partial t} = \text{RHS} - \sum_{n=1}^{N} G(\varepsilon_n, \delta t, \delta r, \delta z)(f - f_n^{\text{obs}}) \tag{6.8a}$$

here RHS (Right-Hand Side) contains all the other linear and nonlinear terms of the prognostic evolution equation for f, while f_n^{obs}, $n = 1, \ldots, N$, are the measurements of f at N observational stations. Equation (6.8a) is analogous to Eqs. (4.8b) and (5.22) (see also Table IV and its discussion in Sec. 5.4.1).

The relaxation function G is in principle a function of ε_n, the standard deviation of the nth observation, of δt, which is the separation between the observation times t_n and the model evaluation time t, and of δr and δz, the horizontal and vertical distances, respectively, between the model grid point and the observational point. In practice, however, G is assumed to be a constant in all meteorological applications. Anthes (1974) noted that, from first principles, one can expect G to be positive and decrease with increasing observation error ε, increasing horizontal and vertical distance separation and increasing time separation. In practice, however, he chose

$$G = \text{constant} > 0 \text{ for } \delta r = \delta z = 0 \qquad (6.8b)$$

$$G = 0 \text{ if either } \delta r \text{ or } \delta z \neq 0 \qquad (6.8c)$$

In the altimetric data assimilation carried out with a QG model by Holland and Malanotte–Rizzoli (1989), Eq. (6.8a) specializes to

$$\frac{\partial \zeta_1}{\partial t} = \text{RHS} - r(\zeta_1 - \zeta_1^{\text{obs}}) \qquad (6.9a)$$

where ζ_1 is the relative vorticity in the surface layer, which is related to the quasi-geostrophic streamfunction ψ_1 (the direct altimetric measurement) by $\zeta_1 = \nabla^2\psi_1$. Holland and Malanotte–Rizzoli studied the sensitivity of the assimilation experiments to different choices of r. They considered a general shape for r given by

$$r = r_0 e^{-(x^2+y^2)/L_R^2} e^{-\alpha t} \qquad (6.9b)$$

where r_0 is typically of the order of (2 days)$^{-1}$. The Gaussian shape for the horizontal dependence of r has a decay distance of the order of the first Rossby deformation radius L_R and the (empirically found) best value for the decay time scale is $\alpha = (5 \text{ days})^{-1}$. The nudging method has been tested successfully for the assimilation of altimeter data also in experiments carried out with the Holland and Lin (1975) model (Haines *et al.*, 1991) and in the experiments carried out with the S.P.E.M. Gulf Stream model when assimilating localized data clusters (Malanotte–Rizzoli and Young, 1991, and Section 6.1.1 here).

As remarked in Section 6.1.1, the direct insertion of altimetric data used by Berry and Marshall (1989) causes an additional vertical velocity $w_{1,2}$ between the surface layer and the one immediately below. In the case of a two-layer model, $w_{1,2}$ is determined principally by ψ_1, the observed surface-pressure field, because ψ_2 tends to be small due to bottom friction. This is not true, however, for multilayer models. In multilayer models, the stretching induced by the top two layers does not cause any immediate change in the current structures, and alterations can only occur on much longer time scales.

The description of this process given by Haines (1991) is based on considering the field changes caused by the direct insertion. Assuming that a complete set of surface observations are available, the Berry and Marshall (1989) scheme corresponds to

$$\begin{cases} \Delta\psi_1 = \psi_1 - \psi_1^{\text{obs}} = \text{given} & (6.10\text{a}) \\ \Delta\psi_j = 0 \text{ for } j = 2, \ldots, N & (6.10\text{b}) \end{cases}$$

where $\Delta\psi$ is the field change due to insertion. That is, the pressure change is assigned at the surface only. Of course, pressure changes in the deeper layers do occur over time as the model is integrated, as described by the Berry and Marshall method of considering the w fields in the ω-equation.

In contrast, the nudging scheme indirectly alters the surface pressure field by altering the surface potential vorticity field q_1, since the nudging term appears on the right side of the surface potential vorticity equation. Thus, after one time step Δt, we have

$$\Delta q_1 = \Delta t(q_1 - q_1^{\text{obs}}) = \text{given} \qquad (6.11\text{a})$$

$$\Delta q_j = 0 \text{ for } j = 2, \ldots, N \qquad (6.11\text{b})$$

This has, of course, the disadvantage that q_1^{obs} is not really the observed field, which is ψ_1^{obs}. Disregarding this at present, it was pointed out by Holland and Malanotte-Rizzoli (1989) and more explicitly by Haines (1991) that a change in the surface potential vorticity field alone causes an immediate change in the current structures at all vertical levels; this is consistent with the QG approximation, i.e., with the approximate validity of the Taylor-Proudman theorem. Thus, nudging causes rapid penetration of surface information into the deep ocean, unlike the direct insertion method, which only changes the surface currents instantaneously.

With this contrast in mind, Haines (1991) develops a new method of data assimilation that avoids the disadvantages of both previous methods. He proposes that the following field changes be made at the assimilation time:

$$\begin{cases} \Delta\psi_1 = \psi_1 - \psi_1^{\text{obs}} = \text{given} & (6.12\text{a}) \\ \Delta q_j = 0 \text{ for } j = 2, \ldots, N & (6.12\text{b}) \end{cases}$$

Equation (6.12a) is clearly the same as Eq. (6.10a), while Eq. (6.12b) is the same as Eq. (6.11b). Thus, the known ψ_1 is used in the upper layer, avoiding the use of the unknown q_1, while instantaneous QG adjustment is achieved in the lower layers. To see how this adjustment occurs, Haines demonstrates how the variable changes of Eqs. (6.12a,b) can be inverted to obtain the other variable changes, i.e., Δq_1 and $\Delta\psi_j$ for $j = 2, \ldots, N$.

The method first solves for the $\Delta\psi_j$ using the definitions of potential vorticity in layers $2, \ldots, N$ (in his case $N = 4$),

$$\nabla^2(\Delta\psi_2) - \gamma_{2,3}^2[(\Delta\psi_2) - (\Delta\psi_3)] - \gamma_{2,1}^2(\Delta\psi_2) = -\gamma_{2,1}^2(\Delta\psi_1) \quad (6.13a)$$

$$\nabla^2(\Delta\psi_3) - \gamma_{3,2}^2[(\Delta\psi_3) - (\Delta\psi_2)] - \gamma_{3,4}^2[(\Delta\psi_3) - (\Delta\psi_4)] = 0 \quad (6.13b)$$

$$\nabla^2(\Delta\psi_4) - \gamma_{4,3}^2[(\Delta\psi_4) - (\Delta\psi_3)] = 0 \quad (6.13c)$$

The right sides would normally contain Δq_j, but these are assumed to be zero, and $\Delta\psi_1$ in Eq. (6.13a) is known. The $\gamma_{i,j\pm1}$ are the inverse Rossby deformation radii for the interfaces. The resulting Δq_1 is given by

$$\Delta q_1 = \nabla^2(\Delta\psi_1) - \gamma_{1,2}^2[(\Delta\psi_1) - (\Delta\psi_2)] \quad (6.13d)$$

Haines uses this method for intermittent data assimilation in an identical-twin experiment and demonstrates rapid convergence to the control run in all layers. A more recent manuscript uses a similar method in a shallow-water model and compares the success with nudging (Haines *et al.*, 1991).

The attractiveness of the previously discussed methods lies in their simplicity, since their implementation involves only fairly straightforward modifications of existing dynamic models. A general disadvantage, on the other hand is that these empirical schemes are not well suited to address issues of consistency and errors in the estimated solution when applied to real situations, where the true reference ocean is unknown.

6.3.3. Kalman Filtering Applications

The theory of the K-filter was presented in Section 4.1, and its meteorological applications were presented in Section 5.3. The sequential nature of the state estimation provided by K-filtering makes it particularly well suited to the meteorological application of forecasting. In the oceanographic context, however, data at different times are stored and used simultaneously. Thus, time becomes a fourth coordinate, like space, and the K-filter can be used as a smoother (Bennett and Budgell, 1989; Gaspar and Wunsch, 1989), i.e., an optimal estimator that uses formally future data (see also Section 7 here).

The great advantage of sequential estimation methods is that they are capable of providing explicit error estimates, such as the error bars or the error covariance matrix of the obtained solution. More difficult, and this is true for any methodology discussed here, is the identification of systematic model errors as distinguished from forcing errors.

The K-filter has been applied to oceanographic problems by Budgell (1986a,b), Miller (1986, 1989), Webb and Moore (1986), Bennett and Budgell (1987, 1989), Carter (1989), Gaspar and Wunsch (1989), Miller and Cane (1989), and Miller and Ghil (1990). In most of these applications, relatively simple dynamical models were used, but Heemink and Kloosterhuis

(1990) have used a K-filter operationally and quite successfully in a nonlinear shallow-water model of the North Sea, for real-time prediction of water levels along the Dutch coast and of their all-important error bars during storm surges.

Webb and Moore (1986) used a projection method equivalent to a simplified K-filter to transfer the surface information provided by altimetry to the deeper layers. They used at update times a projection of the forecast field onto the hyperplane of sea surface elevations given by altimetric measurements [compare Daley, 1980, 1981, and discussion of Eq. (5.12) here] and studied the convergence of the assimilation process (cf. also Talagrand, 1981). They assumed that the measurements were error free and available everywhere at fixed time intervals. These authors approximated the oceanic fields by an error-free superposition of linear Rossby waves and showed that their method represents, in this case, a highly simplified K-filter [Eqs. (4.17) here with $\mathbf{R} = 0$, $\mathbf{Q} = 0$, and $\mathbf{P} = \mathbf{I}$, a unit matrix]. A result of this study was that the determination of the deeper structure of the ocean was limited by the phase separation that develops over each assimilation interval between modes of the ocean with the same horizontal wave number but different vertical structure, given fixed-length update intervals (cf. also Bube and Ghil, 1981).

Miller's (1986) work was motivated by data assimilation into an eddy-resolving open ocean model. He applied the full K-filter to a barotropic vorticity equation designed to capture some of the properties of open-ocean modelling. Miller showed that the filter can follow instabilities well, and its performance with open boundaries is as good as with the periodic conditions of Ghil *et al.* (1981).

Budgell (1986a) applied K-filtering to a one-dimensional linear shallow-water model, corresponding to the conservation equations for momentum and mass integrated across the section of an open channel. Nonlinear processes were also included by Budgell (1986b). Numerical applications to the Great Bay estuary in New Hampshire were successful in estimating the along-channel distributions and time evolutions of the surface elevation and total transport. Model errors were included as a stochastic forcing.

Bennett and Budgell (1987) showed that the time-continuous Kalman–Bucy (1961) filter with regular time and space sampling at a certain period and wavelength will not converge for waves of shorter periods and wavelengths. This result is intuitively obvious (Bube, 1981; Bube and Ghil, 1981) and merely indicates that the subgrid-scale problem in data assimilation, as in modeling, has no simple ready answers.

Kindle (1986) got similar results using an eddy-resolving numerical model. He found that the model integration would not converge given observed data unless the data had a time–space sampling rate equal to the time–space decorrelation scale of the model's eddy activity. Kindle used the direct insertion of observations into the numerical model discussed in Section 6.3.2 and got essentially the same results as Bennett and Budgell (1987), but slower convergence. It follows that the K-filter does not overcome the problem of resolution, but it does allow for a more rapid convergence of the assimilation for the periods and wavelengths that can be resolved by the model. The successful application of the K-filter by Carter (1989) for the assimilation of Lagrangian data from 39 isopycnal RAFOS floats in the Gulf Stream shows, on the other hand, that when the sampling is not on a regular time–space grid, some of these difficulties may be overcome (cf. also Bube, 1981; Bube and Ghil, 1981).

In the second part of their investigation, Bennett and Budgell (1989) examined methods for computing the Kalman smoother in an efficient way feasible for practical calculations. They show that the computation of the smoother may be completed in a well-conditioned way without having to store error covariance matrices throughout the integration time interval, thus reducing considerably the computational effort.

Miller (1989) showed that the K-filter, using minimal information from another source, can overcome the major problem of altimetric measurements, namely that of relative measurements only: since orbit determination is not sufficiently precise for an absolute measurement of sea level, differences in space and time only are provided. The K-filter converges to an absolute sea-surface height map from altimetric differences, provided absolute measurements are provided at one point in space only, e.g., from one tide gauge, at least in Miller's (1989) idealized setting.

Miller and Cane (1989) carried out the first application of the K-filter to a real oceanographic problem, with the scientific objective of producing monthly mean sea-level maps for the period 1978–1983 in the equatorial Pacific. As already remarked, a sophisticated assimilation technique is used in their application in conjunction with a simple dynamical model. This consists of the linearized momentum equations on an equatorial β-plane with the long-wave approximation (Cane, 1984). In the model, the motion is decomposed into vertical modes. The amplitude of each vertical mode is then expanded into the meridional normal modes of the equatorial wave guide, the Hermite functions (Cane, 1984). Thus, the solutions obtained by classical separation of variables take the form:

$$\begin{aligned}\begin{pmatrix} u_m \\ h_m \end{pmatrix} &= \frac{a_{Km}(x,t)}{2^{1/2}}\begin{pmatrix} \psi_0(y) \\ \psi_0(y) \end{pmatrix} \\ &+ \sum_n^N \frac{r_{n,m}(x,t)}{2 \cdot 2^{1/2}}\begin{pmatrix} (n+1)^{-1/2}\psi_{n+1} - n^{-1/2}\psi_{n-1} \\ (n+1)^{-1/2}\psi_{n+1} + n^{-1/2}\psi_{n-1} \end{pmatrix} \end{aligned} \tag{6.14}$$

here (u_m, h_m) is the amplitude of the mth baroclinic mode for the zonal velocity component and the sea-level height anomaly, respectively, $a_{Km}(x,t)$ is the amplitude of the Kelvin wave and $r_{n,m}(x,t)$ is the amplitude of the nth meridional mode Rossby wave and ψ_n is the nth meridional Hermite function. Thus, the equations governing the Kelvin and Rossby wave amplitudes are one-dimensional in space and simple enough for efficient K-filtering. Moreover, the model was used in a highly truncated version, with only two baroclinic modes and five meridional modes; experiments with up to nine meridional modes showed no significant changes in the results.

The model was run with monthly-average wind stress forcing for the six-year interval from January 1978 through December 1983, first in a predictive mode for simple comparisons between model results and tide-gauge data at island stations. Subsequently, the K-filter was applied using sea level data from six selected tide-gauge stations. The effectiveness of K-filtering was tested by comparing the raw model output and the filtered output with real observations at four tide-gauge stations not used in the assimilation. Figure 24a shows such a comparison. Clearly, the filtered model outputs (heavy lines) are, in general, in better agreement with the observations (dots) than the raw model output (light solid); the improvement, as discussed by Miller and Cane, is not always significant (see for instance, the results at Canton and at Kapingamarangi). The larger discrepancies, in both filtered and unfiltered results, towards the end of the time interval are probably due to the model's difficulty in simulating the anomalously large El Niño event of 1982–1983. The results of the assimilation will be greatly improved by adding a few observation from the planned TOGA Thermal Array for the ocean (TAO) (Miller, 1990).

As already remarked, the K-filter provides, as a very important byproduct, maps of the error estimate that quantify the goodness of the result. Figures 24b,c show contour maps of the rms error for the entire model domain; the expected rms error of the raw model output is shown in panel (b) and the rms error of the model updated at the six stations is shown in panel (c). The filter reduces the rms error by about 1 cm. The improvement is, not surprisingly, small since only 6 data points are used in the assimilation. Furthermore, both the wind forcing errors and model errors are probably important and are difficult to estimate (see, however, Dee *et al.*, 1985).

An application of K-filtering to a midlatitude oceanographic problem was made by Gaspar and Wunsch (1989). They use an even simpler model than Miller and Cane, i.e, the barotropic linear Rossby wave equation. This is surely rather less appropriate as a model for the Northwestern Atlantic, where the energetic Gulf Stream system produces major departures from linear, barotropic dynamics than Cane's (1984) model for the tropical Pacific. Gaspar and Wunsch's goal, however, is to determine the fraction of oceanic variability in the Northwest Atlantic, which is consistent with linear barotropic Rossby waves. By consistent, these authors mean that the observed

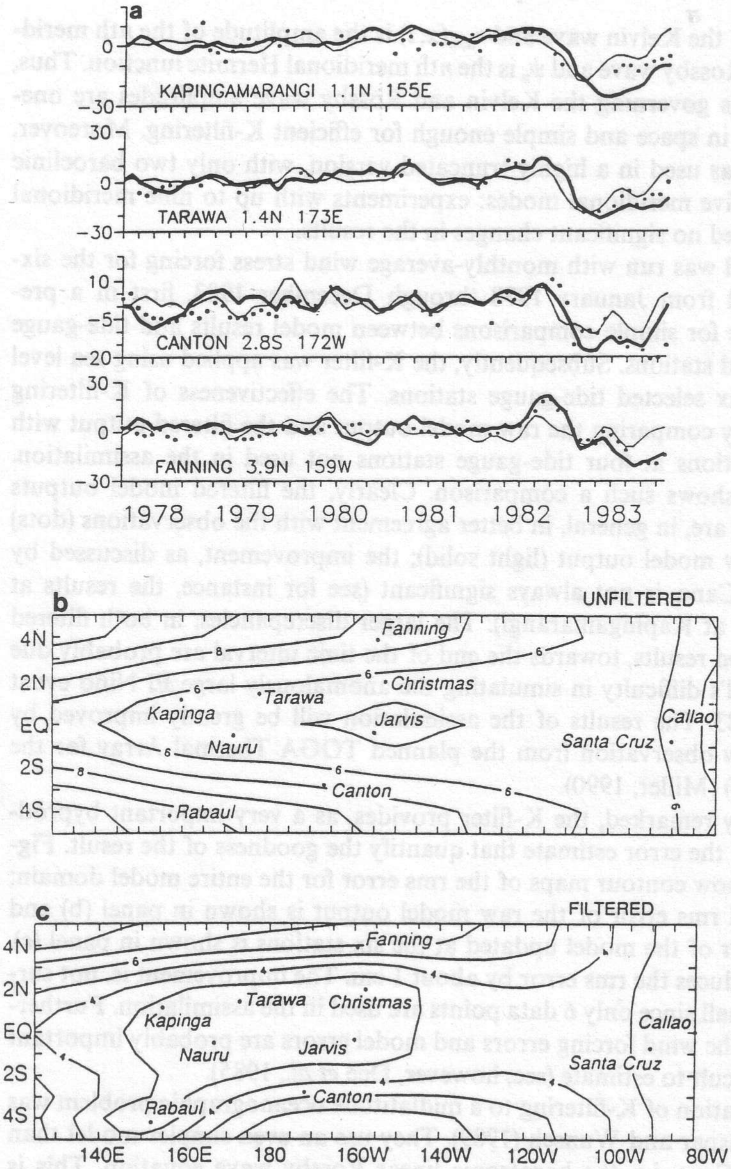


FIG. 24. (a) Comparison of observations (dots), raw model output (light solid) and filtered output (heavy solid) at four stations. Data from these stations are not used in the assimilation. Ordinate in cm. (b) Contour maps of expected rms error of raw model output in cm. (c) Contour map of expected rms error of filtered model output in cm (from Miller and Cane, 1989).

variability so described is indistinguishable from what the dynamical model demands.

The model equation is

$$\frac{\partial}{\partial t}\nabla^2 h + \beta\frac{\partial h}{\partial x} = 0 \tag{6.15a}$$

where h, the surface elevation, is expanded into M horizontal Rossby modes:

$$h(x, y, t) = \sum_{m=1}^{M} \alpha_m \sin(\mathbf{K}_m \cdot \mathbf{X} - w_m t + \theta_m), \qquad \mathbf{X} = (x, y) \tag{6.15b}$$

Data from 10 successive 17-day repeat cycles of GEOSAT were used, covering the period 24 March to 9 September 1987, and K-filtering applied to the model. The authors kept 32 Rossby modes in Eq. (6.15b) and performed experiments without and with system noise, apart from a series of sensitivity tests. Five dominant Rossby modes were identified by using forward filtering only as well as fixed-interval smoothing over the entire time interval of 170 days. Figure 25a shows the time evolution of the estimated amplitudes α_m (left panel) and phases θ_m (right panel) of the five dominant modes in one of their experiments. The filtered (solid) and smoothed (dotted) estimates of large-amplitude changes are in good agreement when the changes are slow in time (see the curves for mode W2), but are less so when the changes are fast (see mode W3). In Fig. 25b, the characteristic wave numbers and periods are reported for the five barotropic Rossby modes found to carry significant energy as well as being simultaneously consistent with model and data. Only a very minor fraction, unfortunately, of total signal variance, 5–15%, is consistent with these five Rossby modes over several GEOSAT repeat cycles. This may not be too surprising considering the dynamical complexity of the region where the study is carried out.

6.3.4. *Applications of Variational Methods*

The connection between variational methods and sequential estimation was discussed in Section 4 and in Section 5.4.1, where the duality principle of Kalman (1960) was shown to be the basis of this connection. In this context, Kimeldorf and Wahba (1970) have shown that statistical interpolation produces fields that are the solution of a variational problem in which the function to be minimized is the sum of two terms, one representing the distance to the observations and the other some measure of smoothness of the fields.

The forward–backward data assimilation introduced by Morel *et al.* (1971) can also be related to variational assimilation. In their approach, the

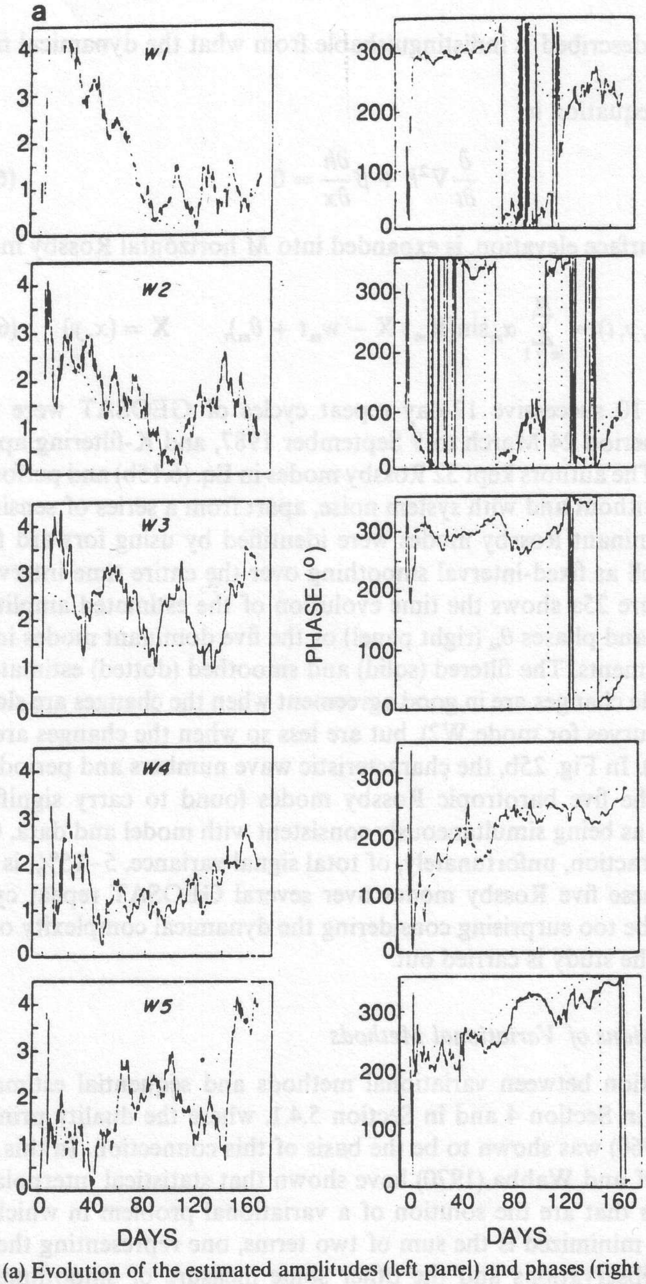


FIG. 25. (a) Evolution of the estimated amplitudes (left panel) and phases (right panel) of the five dominant modes in the experiment with system-noise variance $\sigma^2 = 10^{-6}$ m^2. (b) The five Rossby modes found to carry a significant amount of energy that are simultaneously consistent with both data and model (from Gaspar and Wunsch, 1989).

b

Wave	Wave vector (cycles/1000 km)	Period (days)
W1	(−3, 0)	77.1
W2	(−2, 3)	167.0
W3	(−2, −1)	64.2
W4	(−1, −1)	51.4
W5	(−1, −2)	128.4

FIG. 25. (*Continued*)

model is integrated forward and backward repeatedly over time to obtain an adjustment of the model to the observations. In the adjoint method, the model itself is integrated forward, but the adjoint of the model is used in the backward integration. Talagrand (1981) has shown that a sufficient condition for convergence of a forward–backward assimilation scheme, as described by Morel *et al.* (1971), is that the linearized perturbation equations be antisymmetric. The time-backward integration of the adjoint of an antisymmetric equation is identical to the equation itself, which explains the success reported by Morel *et al.* (1971) in the case of an antisymmetric equation.

Assimilation procedures based on the variational approach are not restricted, however, to antisymmetric equations. Thacker (1986) discusses the connection between K-filtering and the adjoint method for data assimilation using a linear model. Kalman filtering is simply a particularly efficient algorithm to minimize the distance between a model trajectory and given data, subject to certain assumptions (Sections 4.1 and 4.2; see also Paige and Saunders, 1977).

Bennett and McIntosh (1982) used a variational method in the investigation of tidal motion. Their results emphasize that the proper choice of data weights is of great importance. As already mentioned in Section 6.3.1, Provost (1983) and Provost and Salmon (1986) have used a variational technique to assimilate hydrographic station data to estimate the three-dimensional field of geostrophic velocities. They used the method of weak constraints (Sasaki, 1970), that is they found the smoothest velocity field consistent with the data and at the same time satisfied approximately the geostrophy constraints. A smooth solution was obtained by penalizing kinetic energy as well as enstrophy.

A direct minimization approach was also adopted by Legler *et al.* (1989) to develop an objective analysis technique for monthly average wind stress fields over the Indian Ocean. Their cost functional is composed of five quadratic terms, with a weight that determines the relative closeness of fit for each term. The first term is the square distance of the analysis to the first guess, the second to climatology, the third is a measure of smoothness, and the last two are kinematic constraints on the curl and divergence of the stress. The cost functional is minimized by using the conjugate-gradient method. Results for various weight combinations are presented and the optimal weight combination is found by comparison with a subjective analysis.

Applications of variational methods by the Mesoscale Air-Sea Interaction Group (MASIG) at Florida State University under the direction of J. J. O'Brien have addressed also parameter estimation in numerical modeling of hydraulic systems (Panchang and O'Brien, 1990). Usually these parameters are optimized by empirical tuning of the model to observations. Panchang and O'Brien (1990) used the adjoint method to determine the friction factor for tidal rivers.

In parallel work (Smedstad, 1989), adjoint equations were developed for a linear, reduced-gravity shallow-water model to assimilate island sea-level data in the equatorial Pacific. Due to the large latitudinal extent, spherical coordinates are used, with ϕ being the longitude and θ the latitude:

$$\frac{\partial U}{\partial t} - fV = -\frac{c^2}{a\cos\theta}\frac{\partial h}{\partial \phi} + \frac{\tau^\phi}{\rho} + A\nabla^2 U \quad (6.16a)$$

$$\frac{\partial V}{\partial t} + fU = -\frac{c^2}{a}\frac{\partial h}{\partial \theta} + \frac{\tau^\theta}{\rho} + A\nabla^2 V \quad (6.16b)$$

$$\frac{\partial h}{\partial t} + \frac{1}{a\cos\theta}\left[\frac{\partial U}{\partial \phi} + \frac{\partial}{\partial \theta}(V\cos\theta)\right] = 0 \quad (6.16c)$$

here (U, V) are the eastward and northward components of the transport, a is the earth's radius, (τ^ϕ, τ^θ) are the zonal and meridional components of the wind stress, and the wind data used are the pseudowind stress fields of Legler and O'Brien (1986). A is the horizontal eddy viscosity coefficient, $h(x, y, t)$ is the pycnocline interface, ∇^2 is the Laplacian operator in spherical coordinates, $c^2 = gH\frac{\Delta\rho}{\rho_0}$ is the reduced gravity wave speed, with $\Delta\rho$ the density difference between the two model layers. The parameter to be estimated by the adjoint method is c^2. The cost function to be minimized is

$$J(h, c^2) = \int_\Sigma \left[\frac{K_h}{2}(h - h')^2 + \frac{K_c}{2}(c^2 - c'^2)^2\right] d\Sigma \quad (6.17)$$

where h' represents an observation of the upper-layer thickness and c'^2 is an *a priori* best guess of the phase speed. K_h and K_c are validity coefficients and Σ represents the spatial and temporal domain over which the model is integrated.

First, simulation experiments were carried out using the model solution as observations. They showed that the assimilation algorithm is able to determine the spatial structure of c^2 with observations available only at three stations. The estimated c^2 is not sensitive to errors in the observations, assumed to be uncorrelated. Subsequently, the real sea-level observations from three Pacific island stations were used for the different periods. The year 1979 was chosen to represent a year without El Niño, while 1982–1983 was chosen to represent an El Niño year. The assimilations for the latter started in June 1982 and continued for one year. The initial guess for c^2 was 6.0 $\text{m}^2\,\text{sec}^{-2}$. The cost function decreased to almost 35% of its initial value after five iterations. The corresponding evolution of the spatial structure of c^2 is shown in Fig. 26.

After the first iteration, Fig. 26a, there is an adjustment in the western and central region of the basin, with c^2 values dropping near the western boundary and higher values in the central area. Figure 26b shows the zonal distribution of c^2 after the third iteration and Fig. 26c shows it after five iterations. In Fig. 26c, a steep slope has developed with lowest values close to the western boundary, $c^2 = 3.3\ \text{m}^2\,\text{sec}^{-2}$. The maximum value $c^2 = 7.0\ \text{m}^2\,\text{sec}^{-2}$ is reached near ~160°W, with a subsequent slow decrease eastward. The estimated spatial structure of c^2 shown in Fig. 26c is in good agreement with observations. On the other hand, assimilation for the year 1979 gave the opposite picture, with c^2 values higher in the west and lower in the east (not shown here).

A variational adjoint method has been used to assimilate real XBT data into a linear reduced-gravity model of the tropical Pacific by Sheinbaum and Anderson (1990a,b), who also carry out sensitivity studies to investigate some deficiencies of the results. They show that in assimilation experiments performed with simulated data, it is possible to distinguish between model errors and forcing (wind stress) errors.

It is well known that a serious problem of the adjoint method is its computational efficiency, which depends on the descent method used to minimize the cost function (e.g., Courtier and Talagrand, 1987). Several different conjugate-gradient algorithms exist, and some of them were tested by Navon and Legler (1987). They concluded that the subroutine CONMIN of Shanno and Phua (1980) gave the best convergence rates. Smedstad (1989), using this algorithm, achieved convergence in 10 iterations or less. Further work is in progress to reach the maximum efficiency in this area of optimization methods.

As already mentioned in Section 6.1.1, Wunsch (1988) applied a control-theoretical approach to a very simple advection-diffusion equation to study a tracer problem. The use of a transient tracer to invert for the flow and the mixing rates involves a two-step process. First, one starts with a forward

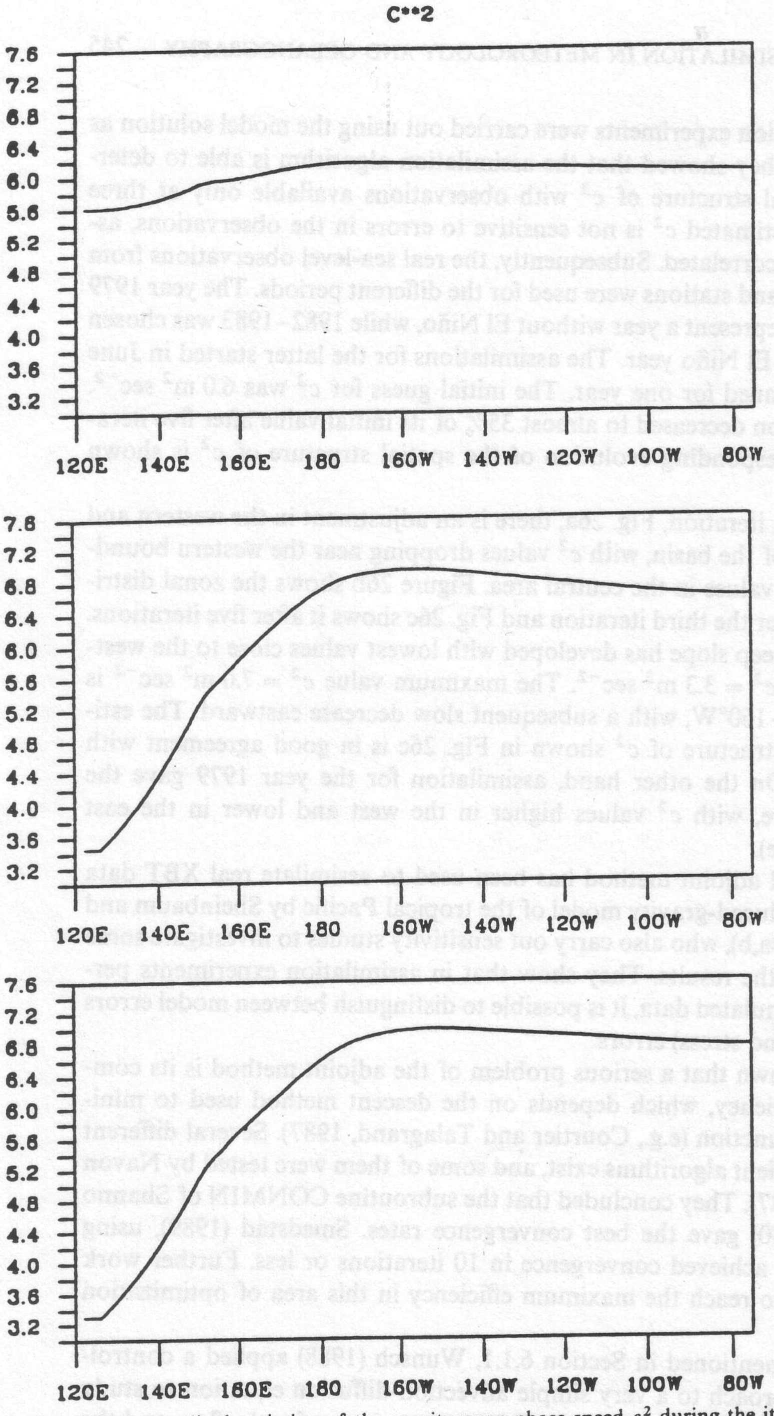


FIG. 26. The longitudinal variation of the gravity wave phase speed c^2 during the iterative process of the adjoint method when assimilating sea-level data from three equatorial Pacific island stations during the El Niño year 1982–1983. (a) Distribution after one iteration; (b) after three iterations; (c) after five iterations (from Smedstad, 1989).

model and determines if acceptable boundary conditions drive the model to reproduce the transient tracer distribution at the observation times. The second step is required only if the model does not reproduce the interior distribution and needs therefore to be improved. Constraints on the fluid flow and the mixing rates can then be used through parameter estimation techniques for boundary control problems.

Wunsch (1988) pointed out the connections between this approach and K-filtering and smoothing (cf. also Bennett and Budgell, 1989), on the one hand, and the adjoint variational method, on the other. Schröter (1989) and Tziperman and Thacker (1989) used simple, though nonlinear, models under rather specialized circumstances, such as one-dimensional fields (Schröter) or steady state (Tziperman and Thacker). These simplifications permit efficient implementations of the adjoint method.

Schröter (1989) uses a time-dependent one-dimensional nonlinear shallow-water model in a cyclic domain, coupled with an advection-diffusion equation for tracers containing a decay term, and applies the adjoint procedure. The most commonly used control variables in the adjoint approach are the initial data $s(t_1)$, t_1 being the initial time. Schröter demonstrates that in this case and for a discrete model, the gradient of the cost function J with respect to the initial data $s(t_1)$ is given by the corresponding Lagrangian parameter λ evaluated one time step backward from t_1 to t_0:

$$\nabla_s J|_{t=t_1} = \lambda(t_0) \tag{6.18}$$

Tziperman and Thacker (1989) use the nonlinear QG barotropic vorticity equation in a closed domain to calculate steady-state circulation from simulated vorticity and streamfunction observations for examples in which wind forcing and friction parameters are also unknown. In time-dependent form, their model is

$$\nabla^2\psi_t + \psi_x + R_0 J(\psi, \nabla^2\psi) = -\varepsilon_h \nabla^2\psi + \varepsilon_b \nabla^4\psi + \operatorname{curl}\vec{\tau}(x, y) \tag{6.19}$$

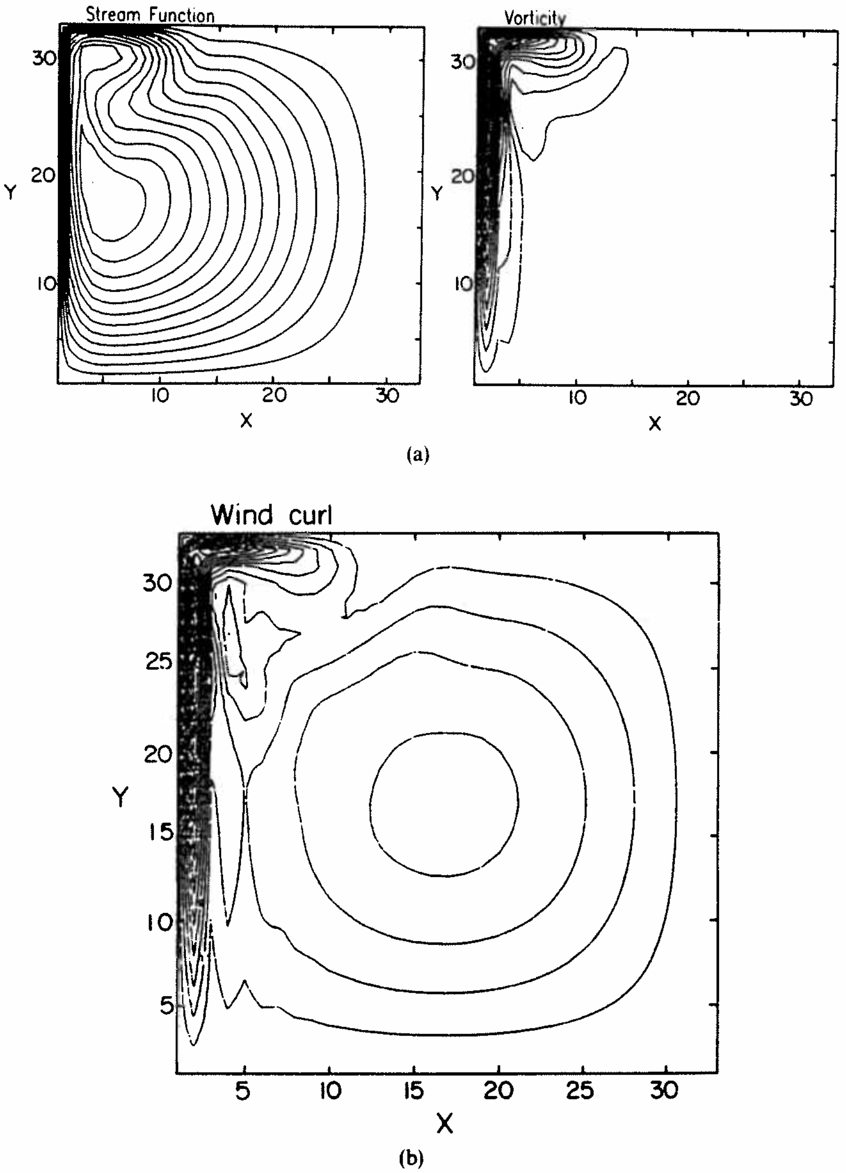
where ψ is the barotropic streamfunction, R_0 the Rossby number, ε_h the horizontal eddy viscosity coefficient, ε_b the bottom friction coefficient, $\vec{\tau}$ the wind-stress field, ∇^2 the Laplacian and J the Jacobian operators, respectively, and subscripts denote partial derivatives. Suppose that the model in Eq. (6.19) is given initial data for the vorticity, $\zeta|_{t=0} = \nabla^2\psi|_{t=0}$, which coincides with a steady-state solution. Then if the model is stepped forward to calculate the unknown fields $\psi|_{t=0}$, $\zeta|_{t=t_n}$, $\psi|_{t=t_n}$ at successive time levels t_n, the difference between the initial data and the solutions after one time step should vanish, as the initial state is the steady-state solution.

Thus, Tziperman and Thacker (1989) are able to carry out a series of sensitivity experiments, including error analysis, in a computationally efficient way. With a simple sinusoidal curl $\vec{\tau}$ field, the steady Stommel–Munk solution satisfies the nonlinear Eq. (6.19). This is shown in Fig. 27a. In their final experiment, friction parameters, wind forcing, and initial vorticity were all treated as unknowns and simultaneously calculated by the optimization. Figure 27b shows the curl $\vec{\tau}$ field thus calculated. Notice the very strong, small-scale forcing by wind stress curl in the western boundary current. This is necessary to balance the dissipation due to the values of the friction parameters found by the optimization, which were much too large. A typical number of iterations for the process to converge was on the order of 200, and a reasonable value for the friction parameter was still not obtained.

An adjoint method for the Harvard quasi-geostrophic model has been developed and applied to GULFCAST data by Moore (1991). The use of the adjoint approach for more highly resolved models is oceanography has been pioneered by Thacker and collaborators (Thacker, 1987, 1988, 1989; Thacker and Long, 1988; Long and Thacker, 1989a,b). Thacker and Long (1988) stress the advantage of deriving the adjoint equations in discretized form. In fact, the discretized form of a continuous adjoint model is not the adjoint of the forward model formulated in discretized form because of truncation errors inherent in the numerical discretization (see Courtier and Talagrand, 1987; Hall, 1986). Also, the duality between sequential estimation and variational estimation methods is very transparent when writing the adjoint in discretized form.

Long and Thacker (1989a) constructed the adjoint for a linearized equatorial ocean model. In a subsequent paper, Long and Thacker (1989b) assessed the performance of the adjoint data assimilation scheme when the different types of data sets are available, with particular emphasis on sea-level observations. In their approach, sea-level data alone are not sufficient and must be supplemented by subsurface information if more than a few baroclinic modes are allowed in the model ocean.

Thacker and Long have undertaken the onerous task of developing the adjoint code for the GFDL model documented by Cox (1984), the most complete GCM for the ocean and the one used by the largest number of oceanographic modelers. Using a model version with 20 points in latitude, 25 points in longitude, and 6 vertical layers, they spin up the model from rest by wind-stress driving. Two preliminary identical-twin experiments were carried out. The synthetic data set was the same in both cases: a full field of u-velocity observations at time step 0, temperature observations at time step 3, v-velocity observations at time step 6, and salinity observations at time step 9, extracted from the model control run. The difference between the two cases lies in the surface boundary conditions; in the first case, surface temperature and salinity are prescribed; in the second, surface fluxes of heat and moisture are specified.



(a)

(b)

FIG. 27. (a) The steady-state solution for the streamfunction ψ and vorticity ζ used as simulated initial state in Eq. (6.19). The parameters used to obtain this solution are $R_0 = 0.01$; $\varepsilon_b = 0.05$; $\varepsilon_h = 0.0001$; curl $\tau = -\sin(\pi x)\sin(\pi y)$. (b) Final solution for curl τ found by the optimization (from Tziperman and Thacker, 1989).

Figure 28a shows the behavior of the cost function for the first case, normalized by the cost at first guess. Convergence is very rapid; the cost is reduced by four orders of magnitude in one iteration and seven orders of magnitude in 15 iterations. In the second case, convergence stalled after a cost reduction of little more than one order of magnitude (not shown). To examine the reason for this, a cross-section of the cost function was evaluated running through the point at which convergence stalled and through the unknown model state being sought. The results are shown in Fig. 28b. The true minimum being sought is at iteration 13. The descent method had converged on a secondary minimum in the cost, a consequence of the nonlinearity of the optimization problem. In fact, Miller and Ghil (1990) have shown that the number of secondary minima increases with the length of the time interval

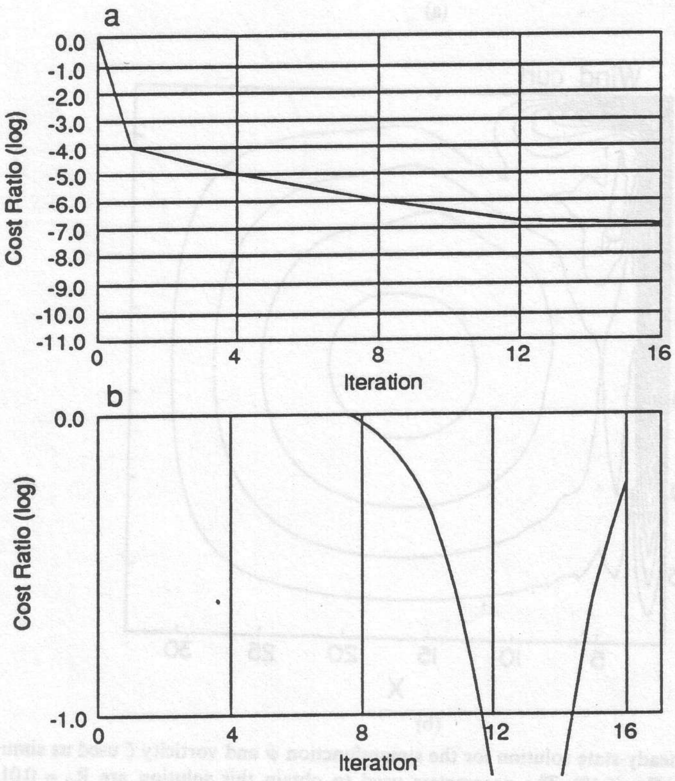


FIG. 28. (a) Cost vs. iteration for the specified temperature and salinity case. (b) Cost section between local and global minima in the case of specified surface heat and water flux. Values are scaled by the cost at first guess and are plotted on a logarithmic scale (courtesy of R. Long and W. C. Thacker).

over which the minimization is carried out (cf. also F. Gauthiez, personal communication, 1990; R. Ziegler, personal communication, 1990).

A first application of the GFDL numerical model with its adjoint code has been made for the North Atlantic using a simplified version of the model and of the adjoint in which the horizontal momentum equations are steady, linearized, and include only the wind stress and a linear bottom stress term (Tziperman *et al.*, 1991a,b). The fully nonlinear, time-dependent GFDL model with the related adjoint are presently being applied to and tested for an idealized Northern Atlantic basin and a realistic configuration of the Eastern Mediterranean Sea. The latter application is part of the modelling effort ongoing on in the Physical Oceanography of the Eastern Mediterranean Sea (POEM) program (Malanotte-Rizzoli and Robinson, 1988).

7. Concluding Remarks

The ambitious and elusive goal of data assimilation is to provide a dynamically consistent "motion picture" of the atmosphere and oceans, in three space dimensions, with known error bars. The ingredients for generating this four-dimensional space-time movie are a large number of observations with different spatio-temporal distributions and error characteristics, on the one hand, and an imperfect knowledge of and ability to solve the equations of fluid motion, on the other.

The purposes of generating this movie can differ: in numerical weather prediction (NWP) and in the emerging discipline of ocean forecasting, the main emphasis is on short loops between successive initial states for subsequent prediction, one day (in the atmosphere) or one week to one month (in the oceans) apart. In climate-related problems, whether atmospheric or oceanic, the emphasis is on full-length "feature movies", based on all the information available for long time intervals, e.g., for the entire duration of a field experiment or of even longer historic data records. The appropriate classes of problems are called prediction, filtering, and smoothing in estimation theory.

Consider a system

$$\dot{\mathbf{w}} = \mathbf{N}(\mathbf{w}) + \mathbf{u}, \qquad -\infty < t < \infty \tag{7.1}$$

where $\mathbf{w}(t)$ is a state vector of grid-point values or spectral coefficients, $\mathbf{N}$ is the known part of the nonlinear dynamics at a given resolution, and $\mathbf{u}(t)$ is system noise, representing subgrid-scale phenomena and other model errors.

Observations $\mathbf{z}(t)$ are given as

$$\mathbf{z} = \mathbf{H}\mathbf{w} + \mathbf{v}, \qquad t_0 < t < t_f \tag{7.2}$$

where $\mathbf{H}$ represents the fact that the observations are partial and indirect and that interpolation between the regular model grid and the irregular observational grid has to occur; $\mathbf{v}(t)$ is observational noise, representing both instrumental and sampling error. Typically $\dim \mathbf{z} \equiv p \ll n \equiv \dim \mathbf{w}$.

The *filtering* problem is that of determining the best estimate $\hat{\mathbf{w}}(t)$ at the end of the time interval over which data are provided, $t = t_f$. The solution of this problem is provided, for a linear system, by the K-filter (Kalman, 1960; Sections 4.1, 5.3 and 6.3.3 here). For a nonlinear system [Eq. (7.1)], no solution which is both computable and truly optimal exists. Various near-optimal, computable solutions do exist (Sections 4.1, 4.2, 5.3.2, 5.4.3 and 6.3).

The *prediction* problem is that of determining $\hat{\mathbf{w}}(t)$ at times after the last available observation, $t > t_f$. Its solution for zero-mean system noise, $E\mathbf{u}(t) \equiv 0$, is simply

$$\hat{\mathbf{w}}^{\cdot} = \mathbf{N}(\hat{\mathbf{w}}), \qquad t > t_f \tag{7.3a}$$

$$\hat{\mathbf{w}}(t_f) = \hat{\mathbf{w}}_f \tag{7.3b}$$

here E is the expectation operator (the ensemble mean), $(\ )^{\cdot}$ denotes time derivatives, and $\hat{\mathbf{w}}_f$ is the solution of the filtering problem for Eqs. (7.1) and (7.2). Estimating the initial state of a forecast from data up to initial time and paying no further attention to the data during the forecast itself is standard practice in NWP and, as we see, makes perfectly good sense.

The *smoothing* problem is that of estimating $\hat{\mathbf{w}}(t)$ optimally at interior points, $t_0 < t < t_f$. It is therewith the problem appropriate for climate-related feature movies (Bennett and Budgell, 1989; Gaspar and Wunsch, 1989; Sections 6.3.3 and 6.3.4 here). One of its solutions involves computing a forward K-filter estimator $\hat{\mathbf{w}}_1(t)$ for intervals (t_0, t) with $t \leq t_f$, a backward estimator $\hat{\mathbf{w}}_2(t)$ for the adjoint of Eq. (7.1) linearized about $\hat{\mathbf{w}}_1(t)$ for intervals (t, t_f) with $t_0 \leq t$, and finding the optimal linear combination between $\hat{\mathbf{w}}_1(t)$ and $\hat{\mathbf{w}}_2(t)$ at each $t \in (t_0, t_f)$. Thus, the Kalman smoother and the adjoint method (Penenko and Obraztsov, 1976; Le Dimet and Talagrand, 1986; Sections 4.2, 5.4.3 and 6.3.4 here) of deterministic optimization theory exhibit certain analogies. The difference is that the smoother of stochastic estimation theory also provides automatically the requisite error bars on the estimated states, whereas the adjoint is easier to formulate.

While this chapter is already rather long, certainly longer than the authors originally planned or expected, it is far from exhaustive. It is by-and-large restricted to the problem of state estimation, having touched only occasionally upon the important problems of parameter estimation (in Section 6.3.4) or noise estimation (in Section 4.1; see also Dee *et al.*, 1985, and Ghil, 1990).

Still, the main points should be clear:

1. The use of dynamic models in a data-assimilation mode is essential to

compensate for the incompleteness, irregular distribution in time and space, and varying error properties of observations in both meteorology and oceanography. The production of both short loops and feature-length movies, for prediction and climate studies, respectively, depends on the ideas and methods of 4-D data assimilation (see also Panel, 1991).

2. For the same model, better assimilation methods extract more information from the same data (see Section 5.1, especially Table II, and Section 6.3.2).

3. Active research on data assimilation is burgeoning rapidly in both meteorology and oceanography. Operational NWP requirements have produced a mature data-assimilation technology in meteorology, from which climatic research has benefitted as well. For the atmosphere, research is concentrating on the implementation of advanced methods from sequential estimation and optimization theory, on the one hand, and on the merging of short NWP loops into climatic 40-year movies, on the other.

4. In oceanography, interest in data assimilation is much more recent, but experience has been gathering rapidly. The advantages of more sophisticated methods are becoming clear (Section 6.3.2). The variety of problems, models, and data sets in oceanography will require a great deal of additional research on methodology before the optimal combination of dynamic model, data assimilation scheme, and sampling strategy for each ocean domain and spatiotemporal scale is decided upon.

An illustration of the latter point is given by the question of what determines the degree of success or lack thereof in an oceanographic assimilation. If the goal is to provide only improved estimates of the larger space scales and longer time scales of motion, i.e., of the quasi-steady component of the circulation, then the main requirement for the success of the assimilation is to obtain a reasonably good estimate of the statistics of the mesoscale eddy field, and not to map it. In this case, steady models with simplified dynamics like those used by the inverse methods discussed in Section 6.3.1 may be adequate. In these latter approaches, mesoscale activity is not explicitly resolved but parameterized, for instance, through eddy-viscosity coefficients.

On the other hand, the ocean is characterized by transient, energetic motions with a broad spectrum in frequency and wave number. A steady component of the circulation may not even exist and be only a model resulting from the analysis of data sets sparse in space and time, like hydrographic data sets, for which steadiness is assumed *a priori*. Hence, eddy-resolving general circulation models (EGCMs) are necessary to study the richness of transient oceanic motions and the variety of their interactions. This is a major goal of a substantial part of the oceanographic community and is especially important for the understanding of energetic systems such as the Western boundary currents, where process studies on mesoscale variability are crucial.

These mesoscale processes also have profound influences in far-away regions of the gyre, determining for instance the penetration scale of the Western boundary jet into the gyre interior (Holland and Schmitz, 1985) or the baroclinic instabilities of the Sverdrup return flow (Holland, 1986). A simple knowledge of the eddy statistics is not sufficient to address these issues, but phase information must also be provided, i.e., visualization and mapping of single realizations is important. Thus, in oceanic data-assimilation problems, the choice of a model and related data assimilation scheme and the definition of success of the assimilation process itself depend crucially on the scientific issue of interest as the starting point

Computational constraints impose, at present, a trade-off between the physical complexity and spatial resolution of the model, on the one hand, and the sophistication of the data assimilation method used in any given study, on the other, for both meteorology and oceanography. As raw computational speed increases, and parallel architectures, coarse- and fine-grained, evolve, we should be able to combine both realistic models and advanced assimilation methods into powerful 4-D data assimilation cycles for the coupled ocean–atmosphere system.

The key issues for advanced data-assimilation methods, whether based on sequential estimation or control theory, are (a) to reduce the computational complexity of implementation algorithms; (b) to provide reliable information on the errors of the estimated fields; and (c) to deal adequately with strongly nonlinear situations. It will be an exciting decade for data assimilation and for the improvement of our ability to describe and understand atmospheric and oceanic flows on global and local scales.

Acknowledgments

It is a pleasure to thank many of our colleagues active in data assimilation and related areas for the insights they provided, for sending reprints and preprints, and for sharing their thoughts. C. Wunsch read an early version of the manuscript and made valuable comments. D. Anderson, M. Cane, S. Cohn, P. Courtier, J. Derber, K. Haines, H. Hurlburt, A. Ingersoll, J. Lewis, R. Miller, R. McPherson, A. Moore, J. O'Brien, O. Talagrand, R. Todling, J. Tribbia, and D. Thompson also read the manuscript at various stages and provided useful comments. Two anonymous reviewers added their input to this already diverse range of comments. All remaining errors and any residual lack of clarity are the authors' fault alone. We are indebted to Ms. Dorothy Frank at M.I.T. and to B Gola, C. Monroe, and C. Wong at UCLA for typing many revisions of the manuscript. S. Cohn, P. Courtier, J. Derber, H. Hurlburt, J. Marshall, R. Miller, A. Moore, O. Smedstad, C. Thacker, R. Todling, E. Tziperman, and C. Wunsch also provided publication-quality figures of their work for inclusion in the paper. This work was supported by NASA Grant NAG-5713, ONR Grant N00014-89-J-1845, and NSF Grant OCE-8919439 (M.G.) and by ONR Grant N00014-90-J-1481 and NASA Project UPN 161-25-63-03 (P.M.R.).

References

Agnon, Y., Malanotte-Rizzoli, P., Cornuelle, B. D., Spiesberger, J. L., and Spindel, R. L. (1989). The 1984 bottom-mounted Gulf Stream tomographic experiment. *J. Acoust. Soc. Am.* **85**, 1958-1966.

Anderson, D. L. T., and Moore, A. M. (1989). Initialization of equatorial waves in ocean models. *J. Phys. Oceanogr.* **19**, 116-121.

Anthes, R. A. (1974). Data assimilation and initialization of hurricane prediction models. *J. Atmos. Sci.* **31**, 701-719.

Atlas, R. A., Ghil, M., and Halem, M. (1982). The effect of model resolution and satellite sounding data on GLAS model forecasts. *Mon. Weather Rev.* **110**, 662-682.

Atlas, R. A., Busalacchi, A. J., Ghil, M., Bloom, S., and Kalnay, E. (1987). Global surface wind and flux fields from model assimilation of Seasat data. *J. Geophys. Res.* **92**, 6477-6487.

Backus, G. E., and Gilbert, J. F. (1967). Numerical applications of a formalism for geophysical inverse theory. *Geophys. J. Roy. Astron. Soc.* **13**, 247-276.

Baer, F. (1977). Adjustment of initial conditions required to suppress gravity oscillations in nonlinear flows. *Beitr. Phys. Atmos.* **50**, 350-366.

Baer, F., and Tribbia, J. J. (1977). On complete filtering of gravity modes through nonlinear initialization. *Mon. Weather Res.* **105**, 1536-1539.

Baker, W. E., Bloom, S. C., Woollen, J. S., Nestler, M. S., Brin, E., Schlatter, T. W., and Branstator, G. W. (1987). Experiments with a three-dimensional statistical objective analysis scheme using FGGE data. *Mon. Weather Rev.* **115**, 273-296.

Balgovind, R., Dalcher, A., Ghil, M., and Kalnay, E. (1983). A stochastic-dynamic model for the spatial structure of forecast error statistics. *Mon. Weather Rev.* **111**, 701-722.

Bellman, R. E., Kagiwada, H. H., Kalaba, R. E., and Sridhar, R. (1966). Invariant imbedding and nonlinear filtering theory. *J. Astronaut. Sci.* **13**, 110-115.

Bengtsson, L. (1975). "Four-Dimensional Data Assimilation of Meteorological Observations," GARP Publ. Ser. No. 15. WMO/ICSU, Geneva.

Bengtsson, L., Ghil, M., and Källén, E., eds. (1981). "Dynamic Meteorology: Data Assimilation Methods." Springer-Verlag, New York.

Bennett, A. F., and Budgell, W. P. (1987). Ocean data assimilation and the Kalman filter, spatial regularity. *J. Phys. Oceanogr.* **17**, 1583-1601.

Bennett, A. F., and Budgell, W. P. (1989). The Kalman smoother for a linear quasi-geostrophic model of ocean circulation. *Dyn. Atmos. Oceans* **13**(3-4), 219-268.

Bennett, A. F., and McIntosh, P. C. (1982). Open ocean modelling as an inverse problem: Tidal theory. *J. Phys. Oceanogr.* **12**, 1004-1018.

Bergthorsson, P., and Döös, B. R. (1955). Numerical weather map analysis. *Tellus* **7**, 329-340.

Berkovitz, L. D. (1974). "Optimal Control Theory." Springer-Verlag, New York.

Berry, P., and Marshall, J. (1989). Ocean modelling studies in support of altimetry. *Dyn. Atmos. Oceans* **13**(3-4), 269-300.

Bertsekas, D. P. (1982). "Constrained Optimization and Lagrange Multiplier Methods." Academic Press, London.

Bierman, G. J. (1977). "Factorization Methods for Discete Sequential Estimation." Academic Press, New York.

Bleck, R., and Boudra, D. B. (1986). Wind-driven spin-up in eddy-resolving ocean models formulated in isopycnic and isobaric coordinates. *J. Geophys. Res.* **91**, 7611-7621.

Boer, G. J., and Shepherd, T. G. (1983). Large-scale two-dimensional turbulence in the atmosphere. *J. Atmos. Sci.* **40**, 164-184.

Bratseth, A. M. (1986). Statistical interpolation by means of successive corrections. *Tellus* **38A**, 439-447.

Bretherton, F. P., Davis, R. E., and Fandry, C. B. (1976). A technique for objective analysis and design of oceanographic experiments applied to MODE-73. *Deep-Sea Res.* **23,** 559–581.

Bryan, K. (1969). A numerical method for the study of the circulation of the World Ocean. *J. Comput. Phys.* **3,** 347–376.

Bryan, K., and Sarmiento, J. L. (1985). Modeling ocean circulation. *Adv. Geophys.* **28A,** 433–459.

Bube, K. P. (1981). Determining solutions of hyperbolic systems from incomplete data. *Commun. Pure Appl. Math.* **34,** 799–830.

Bube, K. P., and Ghil, M. (1981). Assimilation of asynoptic data and the initialization. *In* "Dynamic Meteorology: Data Assimilation Methods" (L. Bengtsson, M. Ghil, and E. Källén, eds.), pp. 111–138. Springer-Verlag, New York.

Bucy, R. S., and Joseph, P. D. (1987). "Filtering for Stochastic Processes with Applications to Guidance," 2nd ed. Chelsea, New York.

Budgell, N. P. (1986a). Stochastic filtering of linear shallow water wave processes. *SIAM, J. Sci. Stat. Comput.*, 34–42.

Budgell, N. P. (1986b). Nonlinear data assimilation for shallow water equations in branched channels. *J. Geophys. Res.* **91,** 10633–10644.

Bunker, A. F., Charnock, H., and Goldsmith, R. A. (1982). A note on the heat balance of the Mediterranean and Red Seas. *J. Mar. Res.* **40,** Suppl., 73–84.

Busalacchi, A. J., and Picaut, J. (1983). Seasonal variability from a model of the tropical Atlantic ocean. *J. Phys. Oceanogr.* **13,** 1564–1588.

Cacuci, D. G. (1981). Sensitivity theory for nonlinear systems. *J. Math. Phys.* **22,** 2794–2812.

Cane, M. A. (1984). Modeling sea level during El Niño. *J. Phys. Oceanogr.* **14,** 1864–1879.

Cane, M. A., and Paden, R. J. (1984). A numerical model for the low frequency equatorial dynamics. *J. Phys. Oceanogr.* **14,** 1853–1863.

Carter, E. F. (1989). Assimilation of Lagrangian data into a numerical model. *Dyn. Atmos. Oceans* **13**(3–4), 355–348.

Carton, J. A., and Hackert, E. C. (1989). Application of multi-variate statistical objective analysis to the circulation in the tropical Atlantic Ocean. *Dyn. Atmos. Oceans* **13** (3–4), 491–515.

Chao, S.-Y. (1984). Bimodality of the Kuroshio. *J. Phys. Oceanogr.* **14,** 92–103.

Charney, J. G., and DeVore, J. G. (1979). Multiple flow equilibria in the atmosphere and blocking. *J. Atmos. Sci.* **36,** 1205–1216.

Charney, J. G., Fjørtoft, R., and von Neumann, J. (1950). Numerical integration of the barotropic vorticity equation. *Tellus* **2,** 237–257.

Charney, J. G., Halem, M., and Jastrow, R. (1969). Use of incomplete historical data to infer the present state of the atmosphere. *J. Atmos. Sci.* **26,** 1160–1163.

Chester, D. B., and Malanotte-Rizzoli, P. (1991). Acoustic tomography in the Straits of Florida. *J. Geophys. Res.* **96,** 7023–7048.

Clancy, R. M. (1987). Real-time applied oceanography at the Navy's Global Center. *Mar. Technol. Soc. J.* **21,** 33–46.

Clancy, R. M., Pollock, K. D., Cummings, J. A., and Phoebus, P. A. (1988). "Technical Description of the Optimal Interpolation System (OTIS) Version 1: A Model for Oceanographic Data Assimilation," FNOC Tech. Note 422-86-02, 422 Branch. Fleet Numerical Oceanography Center, Monterey, California.

Cohn, S. E. (1982). Methods of sequential estimation for determining initial data in numerical weather prediction. Ph.D. Thesis, Courant Institute of Mathematical Sciences, New York University, New York.

Cohn, S. E., and Dee, D. P. (1988). Observability of discretized partial differential equations. *SIAM J. Numer. Anal.* **25,** 586–617.

Cohn, S. E., and Morone, L. L. (1984). "The Effect of Horizontal Gradients of Height-Field Forecast Error Variances upon OI Forecast Error Statistics," Office Note 296. Natl. Meteorol. Cent., Washington, D.C.

Cohn, S. E., and Parrish, D. F. (1991). The behavior of forecast error covariances for a Kalman filter in two dimensions. *Mon. Weather Rev.* (in press).

Cohn, S. E., Ghil, M., and Isaacson, E. (1981). Optimal interpolation and the Kalman filter. *In* "Proceedings of the 5th Conference on Numerical Weather Prediction," pp. 36–42. Am. Meteorol. Soc., Boston, Massachusetts.

Cooper, N. S. (1988). The effect of salinity on tropical ocean models. *J. Phys. Oceanogr.* **18,** 697–707.

Cornuelle, B., Wunsch, C., Behringer, D., Birdsall, T., Brown, M., Heinmiller, R., Knox, R., Metzger, K., Munk, W., Spiesberger, J., Spindel, R., Webb, D., and Worcester, P. (1985). Tomographic maps of the ocean mesoscale. Part I. Pure acoustics. *J. Phys. Oceanogr.* **15,** 133–152.

Cornuelle, B., Munk, W., and Worcester, P. (1988). Ocean acoustic tomography from ships. *J. Geophys. Res.* **94,** 6232–6250.

Courtier, P., and Talagrand, O. (1987). Variational assimilation of meteorological observations with the adjoint vorticity equation. II. Numerical results. *Q. J. R. Meteorol Soc.* **113,** 1329–1347.

Courtier, P., and Talagrand, O. (1990). Variational assimilation of meteorological observations with the direct and adjoint shallow-water equations. *Tellus* **42A,** 531–549.

Cox, M. D. (1984). "A Primitive Equation Three-Dimensional Model of the Ocean," GFDL Ocean Group Tech. Rep. No. 1. GFDL/NOAA, Princeton University, Princeton, New Jersey.

Cressman, G. (1959). An operational objective analysis system. *Mon. Weather Rev.* **87,** 367–374.

Daley, R. (1980). On the optimal specification of the initial state for deterministic forecasting. *Mon. Weather Rev.* **108,** 1719–1735.

Daley, R. (1981). Normal mode initialization. *In* "Dynamic Meteorology: Data Assimilation Methods" (L. Bengtsson, M. Ghil, and E. Källén, eds.), pp. 77–109. Springer-Verlag, New York.

Daley, R. (1991). "Atmospheric Data Analysis." Cambridge Univ. Press, Cambridge, U.K.

Davis, R. E. (1978). On estimating velocity from hydrographic data. *J. Geophys. Res.* **83,** 5507–5509.

Dee, D. P. (1991). Simplification of the Kalman filter for meteorological data assimilation. *Q. J. R. Meteorol. Soc.* **117,** 365–384.

Dee, D. P., Cohn, S. E., Dalcher, A., and Ghil, M. (1985). An efficient algorithm for estimating noise covariances in distributed systems. *IEEE Trans. Autom. Control* **AC-30,** 1057–1065.

DeMey, P., and Robinson, A. R. (1987). Assimilation of altimeter eddy fields in a limited-area quasi-geostrophic model. *J. Phys. Oceanogr.* **17,** 2280–2293.

Derber, J. C. (1989). A variational continuous assimilation technique. *Mon. Weather Rev.* **117,** 2437–2446.

Derber, J. C., and Rosati, A. (1989). A global ocean data assimilation system. *J. Phys. Oceanogr.* **19,** 1333–1347.

De Swart, H. E., and Grasman, J. (1987). Effect of stochastic perturbations on a low-order spectral model of the atmospheric circulation. *Tellus* **39A,** 10–24.

Donner, L. J. (1988). An initialization for cumulus convection in numerical weather prediction models. *Mon. Weather Rev.* **116,** 377–385.

Efron, B. (1982). "The Jacknife, the Bootstrap and Other Resampling Plans." Society of Industrial and Applied Mathematics, Philadelphia, Pennsylvania.

Eliassen, A. (1954). "Provisional Report on Calculation of Spatial Covariance and Autocorrelation of the Pressure Field," Rep. No. 5. Inst. Weather Clim. Res., Academy of Science, Oslo (reprinted in Bengtsson *et al.*, 1981, pp. 319–330).

Errico, R. M. (1982). The strong effects of non-quasigeostrophic dynamic processes on atmospheric energy spectra. *J. Atmos. Sci.* **39,** 961–968.

Errico, R. M. (1989). "Theory and Application of Nonlinear Normal Mode Initialization," NCAR/TN-344 + IA. National Center for Atmospheric Research, Boulder, Colorado.

Fiadeiro, M. E., and Veronis, G. (1984). Obtaining velocities from tracer distributions. *J. Phys. Oceanogr.* **14**, 1734–1746.

Fletcher, R. (1987). "Practical Methods of Optimization," 2nd ed. Wiley, New York.

Freeland, H. J., and Gould, W. J. (1976). Objective analysis of meso-scale ocean circulation features. *Deep-Sea Res.* **23**, 915–923.

Gandin, L. S. (1963). "Objective Analysis of Meteorological Fields." Gidrometeorol. Izd. Leningrad (in Russian), (English Translation by Israel Program for Scientific Translations, Jerusalem, 1965).

Gandin, L. S. (1988). Complex quality control of meteorological observations. *Mon. Weather Rev.* **116**, 1137–1156.

Gaspar, P., and Wunsch, C. I. (1989). Estimates from altimeter data of barotropic Rossby waves in the northwestern Atlantic ocean. *J. Phys. Oceanogr.* **19**, 1821–1844.

Gelb, A., (1974). "Applied Optimal Estimation." MIT Press, Cambridge, Massachusetts.

Gent, P. R., and McWilliams, J. C. (1984). Balanced models in isentropic coordinates and the shallow water equations. *Tellus* **36A**, 166–171.

Ghil, M. (1980). The compatible balancing approach to initialization and four-dimensional data assimilation. *Tellus* **32**, 198–206.

Ghil, M. (1986). Sequential estimation and satellite data assimilation in meteorology and oceanography. *In* "Variational Methods in Geosciences" (Y. K. Sasaki, T. Gat-Chen, L. White, M. M. Zaman, C. Ziegler, L. P. Chang, and D. J. Rusk, eds.), pp. 91–100. Elsevier, Amsterdam.

Ghil, M. (1989). Meteorological data assimilation for oceanographers. Part I. Description and theoretical framework. *Dyn. Atmos. Oceans* **13**(3–4), 171–218.

Ghil, M. (1990). Sequential estimation in meteorology and oceanography: Theory and numerics. *Proc. Int. Symp. Assimil. Obs. Meteorol. Oceanogr.* 85–90.

Ghil, M., and Childress, S. (1987). "Topics in Geophysical Fluid Dynamics, Atmospheric Dynamics, Dynamo Theory and Climate Dynamics." Springer-Verlag, New York.

Ghil, M., and Mosebach, R. (1978). Asynoptic variational method for satellite data assimilation. *In* "The GISS Sounding Temperature Impact Test" (M. Halem *et al.*, eds.), NASA Tech. Memo. 78063, pp. 3.32–3.49. U.S. Govt. Printing Office, Washington, D.C.

Ghil, M., Shkoller, B., and Yangarber, V. (1977). A balanced diagnostic system compatible with a barotropic prognostic model. *Mon. Weather Rev.* **105**, 1223–1238.

Ghil, M., Halem, M., and Atlas, R. (1979). Time-continuous assimilation of remote-sounding data and its effect on weather forecasting. *Mon. Weather Rev.* **107**, 140–171.

Ghil, M., Cohn, S., Tavantzis, J., Bube, K., and Isaacson, E. (1981). Applications of estimation theory to numerical weather prediction. *In* "Dynamic Meteorology: Data Assimilation Methods" (L. Bengtsson, M. Ghil, and E. Källén, eds), pp. 139–224. Springer-Verlag, New York.

Ghil, M., Cohn, S. E., and Dalcher, A. (1982). Sequential estimation, data assimilation, and initialization. *In* "The Interaction Between Objective Analysis and Initialization" (D. Williamson, ed.), Publ. Meteorol. 127 (Proc. 14th Stanstead Seminar), pp. 83–97. McGill University, Montreal.

Ghil, M., Mullhaupt, A., and Pestiaux, P. (1987). Deep water formation and Quaternary glaciations. *Clim. Dyn.* **2**, 1–10.

Gill, A. E. (1982). "Atmosphere-Ocean Dynamics." Academic Press, New York.

Gill, P. E., Murray, W., and Wright, M. H. (1982). "Practical Optimization." Academic Press, London.

Gustafsson, N. (1981). A review of methods for objective analysis. *In* "Dynamic Meteorology:

Data Assimilation Methods" (L. Bengtsson, M. Ghil, and E. Kallen, eds.), pp. 17–76. Springer-Verlag, New York.

Haidvogel, D. B., Wilkin, J., and Young, R. E. (1991). A semi-spectral primitive equation coastal ocean circulation model using vertical sigma and horizontal orthogonal curvilinear coordinates. *J. Comput. Phys.* **94,** 151–185.

Haines, K. (1991). A direct method of assimilating sea surface height data into ocean models with adjustments to the deep circulation. *J. Phys. Oceanogr.* **21,** 843–868.

Haines, K., Malanotte-Rizzoli, P., Holland, W. R., and Young, R. E. (1991). Inter-comparison of methods for the assimilation of altimetry data into a shallow water model. *Dyn. Atmos. Oceans* (submitted for publication).

Hall, M. C. G. (1986). Application of adjoint sensitivity theory to an atmospheric general circulation model. *J. Atmos. Sci.* **43,** 2644–2651.

Halpern, D. (1987). Data assimilation and ocean general circulation models. *EOS, Trans. Am. Geophys. Union* **68,** 731–733.

Harlan, J., and O'Brien, J. J. (1986). Assimilation of scatterometer winds into surface pressure fields using a variational method. *J. Geophys. Res.* **91,** 7816–7836.

Heemink, A. W., and Kloosterhuis, H. (1990). Data assimilation for non-linear tidal models. *Int. J. Numer. Methods Fluids* **11,** 1097–1112.

Ho, Y. C. (1963). On the stochastic approximation method and optimal filtering theory. *J. Math. Anal. Appl.* **6,** 152–154.

Hoffman, R. N. (1982). SASS wind ambiguity removal by direct minimization. *Mon. Weather Rev.* **110,** 434–445.

Hoffman, R. N. (1984). SASS wind ambiguity removal by direct minimization. Part II. Use of smoothness and dynamical constraints. *Mon. Weather Rev.* **112,** 1829–1852.

Hoffman, R. N. (1986). A four-dimensional analysis exactly satisfying equations of motion. *Mon. Weather Rev.* **114,** 388–397.

Hoke, J. E., and Anthes, R. A. (1976). The initialization of numerical models by a dynamic initialization technique. *Mon. Weather Rev.* **104,** 1551–1556.

Holland, W. R. (1978). The role of mesoscale eddies in the general circulation of the ocean, Numerical experiments using a wind-driven quasi-geostrophic model. *J. Phys. Oceanogr.* **8,** 363–392.

Holland, W. R. (1986). Quasigeostrophic modelling of eddy-resolved ocean circulation. *In* "Advanced Physical Oceanographic Numerical Modeling" (J. J. O'Brien, ed.), pp. 203–232. Reidel, Dordrecht.

Holland, W. R. (1989). Altimeter-data assimilation into ocean circulation models—some prelimary results in oceanic circulation models. *In* "Combining Data and Dynamics" (D. L. T. Anderson and J. Willebrand, eds.), pp. 203–230. Kluwer Academic Publ., Amsterdam.

Holland, W. R., and Lin, L. B. (1975). On the origin of mesoscale eddies, and their contribution to the general circulation of the ocean. Part I. A preliminary numerical experiment. *J. Phys. Oceanogr.* **5,** 642–657.

Holland, W. R., and Malanotte-Rizzoli, P. (1989). Along-track assimilation of altimeter data into an ocean circulation model: Space versus time resolution studies. *J. Phys. Oceanogr.* **19,** 1507–1534.

Holland, W. R., and Rhines, P. B. (1980). An example of eddy-induced ocean circulation. *J. Phys. Oceanogr.* **10,** 1010–1031.

Holland, W. R., and Schmitz, W. F. (1985). On the zonal penetration scale of model midlatitude jets. *J. Phys. Oceanogr.* **15,** 1859–1875.

Hollingsworth, A. (1989). The role of real-time four-dimensional data assimilation in the quality control, interpretation, and synthesis of climate data. *In* "Oceanic Circulation Models:

Combining Data and Dynamics" (D. L. T. Anderson and J. Willebrand, eds.), pp. 303–342. Kluwer Academic Publishers, Amsterdam.

Hollingsworth, A., Lorenc, A., Tracton, S., Arpé, K., Cats, G., Uppsala, S., and Kallberg, P. (1985). The response of numerical weather prediction systems to FGGE level IIb data. Part I. Analyses. *Q. J. R. Meteorol. Soc.* **111,** 67–101.

Howe, B. M., Worcester, P. F., and Spindel, R. C. (1987). Ocean Acoustic Tomography: Mesoscale velocity. *J. Geophys. Res.* **92,** 3785–3805.

Hurlburt, H. E. (1986). Dynamic transfer of simulated altimeter data into subsurface information by a numerical ocean model. *J. Geophys. Res.* **91,** 2372–2400.

Hurlburt, H. E., Fox, D. N., and Metzger, E. J. (1990). Statistical inference of weakly correlated subthermocline fields from satellite altimeter data. *J. Geophys. Res.* **95,** 11375–11411.

Jazwinski, A. H. (1970). "Stochastic Processes and Filtering Theory." Academic Press, New York.

Kalman, R. E. (1960). A new approach to linear filtering and prediction problems. *J. Basic Eng.* **82D,** 35–45.

Kalman, R. E., and Bucy, R. S. (1961). New results in linear filtering and prediction theory. *ASME J. Basic Eng.* **83D,** 95–108.

Kalman, R. E., and Koepcke, R. W. (1958). Optimal synthesis of linear sampling control systems using generalized performance indexes. *Trans. ASME* **80,** 1820–1826.

Kanamitsu, M. (1981). Some climatological and energy budget calculations using the FGGE III-b analyses during January 1979. *In* "Dynamic Meteorology: Data Assimilation Methods" (L. Bengtsson, M. Ghil, and E. Källén, eds.), pp. 263–318. Springer-Verlag, New York.

Keppenne, C. (1989). Bifurcations, strange attractors and low-frequency atmospheric dynamics. Ph.D. Thesis, Université Catholique de Louvain, Belgium.

Kerr, T. H. (1990). Fallacies in computational testing of matrix positive definiteness/semi-definiteness. *IEEE Trans. Aerosp. Electron. Syst.* **AES-26,** 415–420.

Kimeldorf, G., and Wahba, G. (1970). A correspondence between Bayesian estimates on stochastic processes and smoothing by splines. *Ann. Math. Stat.* **41,** 495–502.

Kindle, J. G. (1986). Sampling strategies and model assimilation of altimetric data for ocean monitoring and prediction. *J. Geophys. Res.* **91,** 2418–2432.

Kitade, T. (1983). Nonlinear normal mode initialization with physics. *Mon. Weather Rev.* **111,** 2194–2213.

Ko, D. S., DeFerrari, H. A., and Malanotte-Rizzoli, P. (1989). Acoustic tomography in the Florida Strait: Temperature current and vorticity measurements. *J. Geophys. Res.* **94,** 6197–6211.

Krishnamurti, T. N., Bedi, H. S., Weckley, W., and Ingles, K. (1988). Reduction of the spinup time for evaporation and precipitation in the spectral model. *Mon. Weather Rev.* **116,** 907–920.

Lacarra, J. F., and Talagrand, O. (1988). Short-range evolution of small perturbations in a barotropic model. *Tellus* **40A,** 81–95.

Lanczos, C. (1961). "Linear Differential Operators." Van Nostrand, Princeton, New Jersey.

Le Dimet, F.-X., and Talagrand, O. (1986). Variational algorithms for analysis and assimilation of meteorological observations. *Tellus* **38A,** 97–110.

Leetmaa, A., and Ji, M. (1989). Operational hindcasting of the tropical Pacific. *Dyn. Atmos. Oceans* **13**(3–4), 465–490.

Legler, D. M., and O'Brien, J. J. (1986). Analysis of near real-time wind stress data from the tropical Pacific Ocean. *Atmos. Newsl.*, February.

Legler, D. M., Navon, I. M., and O'Brien, J. J. (1989). Objective analysis of pseudostress over the Indian Ocean using a direct-minimization approach. *Mon. Weather. Rev.* **117,** 709–720.

Leith, C. (1980). Nonlinear normal mode initialization and quasi-geostrophic theory. *J. Atmos. Sci.* **37,** 958–968.

Levitus, S. (1982). "Climatological Atlas of the World Ocean." NOAA Prof. Pap. 13. Natl. Oceanic Atmos. Admin., Rockville, Maryland.

Levitus, S. (1989). Interpentadal variability of temperature and salinity at intermediate depths of the North Atlantic ocean, 1970–1974 versus 1955–1959. *J. Geophys. Res.* **94C,** 6091–6131.

Lewis, J. M., and Derber, J. C. (1985). The use of adjoint equations to solve a variational adjustment problem with advective constraints. *Tellus* **37A,** 309–322.

Lions, J.-L. (1971). "Optimal Control of Systems Governed by Partial Differential Equations." Springer-Verlag, Berlin.

Long, R. B., and Thacker, W. C. (1989a). Data assimilation into a numerical equatorial ocean model, Part I. The model and the assimilation algorithm. *Dyn. Atmos. Oceans* **13**(3–4), 379–412.

Long, R. B., and Thacker, W. C. (1989b). Data assimilation into a numerical equatorial ocean model, Part II. Assimilation experiments. *Dyn. Atmos. Oceans* **13**(3–4), 413–440.

Lönnberg, P., and Hollingsworth, A. (1986). The statistical structure of short range forecast errors as determined from radiosonde data. Part II. Covariance of height and wind errors. *Tellus* **38A,** 137–161.

Lorenc, A. (1981). A global three-dimensional multivariate statistical interpolation scheme. *Mon. Weather Rev.* **109,** 701–721.

Lorenc, A. (1986). Analysis methods for numerical weather prediction. *Q. J. R. Meteorol. Soc.* **112,** 1177–1194.

Lorenc, A. C., and Hammon, O. (1988). Objective quality control of observations using Bayesian methods. Theory, and a practical implementation. *Q. J. R. Meteorol. Soc.* **114,** 515–543.

Lorenz, E. N. (1963). Deterministic nonperiodic flow. *J. Atmos. Sci.* **20,** 130–141.

Lorenz, E. N. (1980). Attractor sets and quasi-geostrophic equilibrium. *J. Atmos. Sci.* **37,** 1685–1699.

Lorenz, E. N. (1986). On the existence of a slow manifold. *J. Atmos. Sci.* **43,** 1547–1557.

Luther, M. E., and O'Brien, J. J. (1985). A model of the seasonal circulation in the Arabian Sea forced by observed winds. *Prog. Oceanogr.* **14,** 353–385.

Machenhauer, B. (1977). On the dynamics of gravity oscillations in a shallow water model, with applications to normal mode initialization. *Beitr. Phys. Atmos.* **50,** 253–271.

Madsen, N. K., Rodrique, G. H., and Karush, J. I. (1976). Matrix multiplication by diagonals on a vector/parallel processor. *Inf. Process. Lett.*, **5,** 41–45.

Malanotte-Rizzoli, P., and Holland, W. R. (1986). Data constraints applied to models of the ocean general circulation. Part I. The steady case. *J. Phys. Oceanogr.* **16,** 1665–1687.

Malanotte-Rizzoli, P., and Holland, W. R. (1988). Data constraints applied to models of the ocean general circulation. Part II. The transient, eddy resolving case. *J. Phys. Oceanogr.* **18,** 1093–1107.

Malanotte-Rizzoli, P., and Robinson, A. R. (1988). POEM: Physical Oceanography of the Eastern Mediterranean. *Oceanogr. Rep. EOS* **69.**

Malanotte-Rizzoli, P., and Young, R. E. (1991). Can localized clusters of velocity data be useful for data assimilation? *Dyn. Atmos. Oceans* (submitted for publication).

Malanotte-Rizzoli, P., Young, R. E., and Haidvogel, D. B. (1989). Initialization and data assimilation experiments with a primitive equation model. *Dyn. Atmos. Oceans* **13**(3–4), 349–378.

Marchuk, G. I. (1975). Formulation of the theory of perturbations for complicated models. *Appl. Math. Optimi.* **2,** 1–33.

Mariano, A. J. (1990). Contour analysis: A new approach for melding geophysical fields. *J. Atmos. Oceanic Technol.* **7,** 285–295.

Marshall, J. L. (1985a). Determining the ocean circulation and improving the geoid from satellite altimetry. *J. Phys. Oceanogr.* **15**, 330–349.

Marshall, J. L. (1985b). Altimetric observing simulation studies with an eddy resolving circulation model. *In* "The Use of Satellite Data in Climate Models" (compiled by J. J. Hunt), pp. 73–76. ESA Scientific and Technical Publication Branch, Noordwijk, The Netherlands.

McPherson, R. D., Bergman, K. H., Kistler, R. E., Rasch, G. E., and Gordon, D. S. (1979). The NMC operational global data assimilation system. *Mon. Weather Rev.* **107**, 1445–1461.

McWilliams, J. C., Owens, W. B., and Hua, B. L. (1986). An objective analysis of POLYMODE local dynamics experiment. Part I. General formalism and statistical model selection. *J. Phys. Oceanogr.* **16**, 483–504.

Mercier, H. (1986). Determining the general circulation of the ocean: a non-linear inverse problem. *J. Geophys. Res.* **91**, 5103–5909.

Miller, R. N. (1986). Toward the application of the Kalman filter to regional open ocean modelling. *J. Phys. Oceanogr.* **16**, 72–86.

Miller, R. N. (1987). "Theory and Practice of Data Assimilation for Oceanography," Rep. Meteorol. Oceanogr., No. 26. Harvard University, Cambridge, Massachusetts.

Miller, R. N. (1989). Direct assimilation of altimetric differences using the Kalman filter. *Dyn. Atmos. Oceans* **13**(3–4), 317–334.

Miller, R. N. (1990). Tropical data assimilation experiments with simulated data: The impact of tropical ocean and global atmosphere thermal array for the ocean. *J. Geophys. Res.* **95**, 11461–11483.

Miller, R. N., and Cane, M. A. (1989). A Kalman filter analysis of sea level height in the tropical Pacific. *J. Phys. Oceanogr.* **19**, 773–790.

Miller, R. N., and Ghil, M. (1990). Data assimilation in strongly nonlinear current systems. *Proc. Int. Symp. Assimil. Obs. Meteorol. Oceanogr.* 93–98.

Miyakoda, K. (1956). On a method of solving the balance equation. *J. Meteorol. Soc. Jpn.* **34**, 68–71.

Miyakoda, K., and Talagrand, O. (1971). The assimilation of past data in dynamical analysis. *Tellus* **23**, 310–317.

Mooers, C. M. R., Robinson, A. R., and Thompson, J. D. (1987). "Ocean Prediction Workshop 1986: A Status and Prospect Report on the Scientific Basis and the Navy's Needs." Institute for Naval Oceanography, NSTL, Mississippi.

Moore, A. M. (1990). Linear equatorial wave mode initialization in a model of the tropical Pacific Ocean: An initialization scheme for tropical ocean models, *J. Phys. Oceanogr.* **20**, 423–445.

Moore, A. M. (1991). Data assimilation in a quasigeostrophic open ocean model of the Gulf Stream region using the adjoint method. *J. Phys. Oceanogr.* **21**, 398–427.

Moore, A. M., and Anderson, D. L. T. (1989). The assimilation of XBT data into a layer model of the tropical Pacific Ocean. *Dyn. Atmos. Oceans* **13**(3–4), 441–464.

Moore, A. M., Cooper, N. S., and Anderson, D. L. T. (1987). Initialization and data assimilation in model of the Indian Ocean. *J. Phys. Oceanogr.* **17**, 1965–1977.

Morel, P., Lefevre, G., and Rabreau, G. (1971). On initialization and non-synoptic data assimilation. *Tellus* **23**, 197–206.

Munk, W., and Wunsch, C. (1979). Ocean acoustic tomography: A scheme for large-scale monitoring. *Deep-Sea Res.* **26A**, 123–161.

Munk, W., and Wunsch, C. (1982). Observing the ocean in the 1990s. *Philos. Trans. R. Soc. London, Ser. A* **307**, 439–464.

Navon, I. M., and De Villiers, R. (1983). Combined penalty multiplier optimization methods to enforce integral invariants conservation. *Mon. Weather Rev.* **III,** 1228–1243.

Navon, I. M., and Legler, D. (1987). Conjugate gradient methods for large-scale minimization in meteorology. *Mon. Weather Rev.* **15,** 1479–1502.

Olbers, D. (1989). A geometrical interpretation of inverse problems. *In* "Oceanic Circulation Models: Combining Data and Dynamics" (D. L. T. Anderson and J. Willebrand, eds.), pp. 79–94. Kluwer Academic Publ., Netherlands.

Paige, C. C., and Saunders, M. A. (1977). Least squares estimation of discrete linear dynamical systems using orthogonal transformations. *SIAM J. Numer. Anal.* **14,** 180–193.

Panchang, V. G., and O'Brien, J. J. (1990). On the determination of hydraulic model parameters using the adjoint state formulation. *In* "Modeling Marine Systems" (A. M. Davies, ed.), pp. 5–18. CRC Press, Boca Raton, Florida.

Panel on Model-Assimilated Data Sets for Atmospheric and Oceanic Research (1991). "Four-Dimensional Model Assimilation of Data: A Strategy for the Earth System Sciences" National Academy Press, Washington, D.C.

Panofsky, H. (1949). Objective weather map analysis. *J. Meteorol.* **6,** 386–392.

Parker, R. L. (1972). Inverse theory with grossly inadequate data. *Geophys. J. Ry. Astron. Soc.* **29,** 123–138.

Parrish, D. F., and Cohn, S. E. (1985). A Kalman filter for a two-dimensional shallow-water model. *In* "Proceedings of the 7th Conference on Numerical Weather Prediction," pp. 1–8. Am. Meteorol. Soc., Boston, Massachusetts.

Pedlosky, J. (1987). "Geophysical Fluid Dynamics," 2nd ed. Springer-Verlag, New York.

Penenko, V., and Obraztsov, N. N. (1976). A variational initialization method for the fields of the meteorological elements. *Meteorol. Hydrol. (Eng. Transl.)* **11,** 1–11.

Philander, S. G. H., Hurlin, W. J., and Pacanowsky, R. C. (1987). Initial conditions for a general circulation model of the tropical ocean. *J. Phys. Oceanogr.* **17,** 147–157.

Phillips, N. A. (1971). Ability of the Tadjbakhsh method to assimilate temperature data in a meteorological system. *J. Atmos. Sci.* **28,** 1325–1328.

Phillips, N. A. (1976). The impact of synoptic observing and analysis systems on flow pattern forecasts. *Bull. Am. Meteorol. Soc.* **57,** 1225–1250.

Phillips, N. A. (1982). On the completeness of multi-variate optimum interpolation for large-scale meteorological analysis. *Mon. Weather Rev.* **110,** 1329–1334.

Phillips, N. A. (1986). The spatial statistics of random geostrophic modes and first-guess errors. *Tellus* **38A,** 314–332.

Provost, C. (1983). A variational method for estimating the general circulation of the ocean. Ph.D. Thesis, University of California, San Diego.

Provost, C., and Salmon, R. (1986). A variational method for inverting hydrographic data. *J. Mar. Res.* **44,** 1–34.

Rasch, P. J. (1985). Developments in normal mode initialization. Part II. A new method and its comparison with currently used schemes. *Mon. Weather Rev.* **113,** 1753–1770.

Rienecker, M., Mooers, C. N. R., and Robinson, A. R. (1987). Dynamical interpolation and forecast of the evolution of mesoscale features off Northern California. *J. Phys. Oceanogr.* **17,** 1189–1213.

Robinson, A. R. (1987). Predicting open ocean currents, fronts and eddies. *In* "Three Dimensional Ocean Models of Marine and Estuarine Dynamics" (J. C. F. Nihoul and B. M. Jamart, eds.), pp. 89–112. Elsevier, Amsterdam.

Robinson, A. R., and Leslie, W. B. (1985). Estimation and prediction of oceanic eddy fields. *Prog. Oceanogr.* **14,** 485–510.

Robinson, A. R., and Walstad, L. J. (1987). The Harvard open ocean model: Calibration and application to dynamical process forecasting and data assimilation studies. *J. Appl. Numer. Math.* **3,** 89–121.

Robinson, A. R., Carton, J. A., Pinardi, N., and Mooers, C. M. R. (1986). Dynamical forecasting and dynamical interpolation: An experiment in the California current. *J. Phys. Oceanogr.* **16,** 1561–1579.

Robinson, A. R., Spall, M. A., Leslie, W. G., Walstad, L. J., and McGillicuddy, D. J. (1987). "Gulfcasting: Dynamical Forecast Experiments for Gulf Stream Rings and Meanders, November 1985–June 1986," Harvard University Reports in Meteorology and Oceanography, No. 22. Harvard University, Cambridge, Massachusetts.

Robinson, A. R., Spall, M. A., and Pinardi, N. (1988). Gulf Stream simulations and the dynamics of ring and meander processes. *J. Phys. Oceanogr.* **18,** 1320–1353.

Robinson, A. R., Spall, M. A., Walstad, L. J., and Leslie, W. G. (1989). Data assimilation and dynamical interpolation in gulfcast experiments. *Dyn. Atmos. Oceans* **13**(3–4), 269–300.

Rodgers, C. D. (1977). Statistical principles of inversion theory. *In* "Inversion Methods in Atmospheric Remote Sounding" (A. Deepak, ed.), pp. 117–138. Academic Press, New York.

Rossby, T. R., ed. (1990). "The Synoptician," Newsletter No. 1, January. University of Rhode Island, Providence.

Rutherford, I. D. (1972). Data assimilation by statistical interpolation of forecast error fields. *J. Atmos. Sci.* **29,** 809–815.

Sasaki, Y. (1958). An objective analysis based on the variational method. *J. Meteorol. Soc. Jpn.* **36,** 77–88.

Sasaki, Y. (1970). Some basic formalisms in numerical variational analysis. *Mon. Weather Rev.* **98,** 875–883.

Schlatter, T. W., Branstator, G. W., and Thiel, L. G. (1976). Testing a global multivariate statistical objective analysis scheme with observed data. *Mon. Weather Rev.* **104,** 765–783.

Schröter, J., and Wunsch, C. (1986). Solution of nonlinear finite difference ocean models by optimization methods with sensitivity and observational strategy analysis. *J. Phys. Oceanogr.* **16,** 1855–1874.

Schröter, J. (1989). Driving of non-linear time dependent ocean models by observation of transient tracers—a problem of constrained optimization. *In* "Oceanic Circulation Models: Combining Data and Dynamics" (D. L. T. Anderson and J. Willebrand, eds.), pp. 257–286. Kluwer Academic Publ., Amsterdam.

Seaman, R. S. (1988). Some real data tests of the interpolation accuracy of Bratseth's successive correction method. *Tellus* **40A,** 173–176.

Shanno, D. F., and Phua, K. H. (1980). Remark on algorithm 500—a variable method for unconstrained nonlinear minimization. *ACM Trans. Math. Software* **6,** 618–622.

Sheinbaum, J., and Anderson, D. L. T. (1990a). Variational assimilation of XBT data. Part I. *J. Phys. Oceanogr.* **20,** 672–688.

Sheinbaum, J., and Anderson, D. L. T. (1990b). Variational assimilation of XBT data. Part II. Sensitivity studies and use of smoothing constraints. *J. Phys. Oceanogr.* **20,** 689–704.

Smagorinsky, J., Miyakoda, K., and Strickler, R. (1970). The relative importance of variables in initial conditions for dynamical weather prediction. *Tellus* **122,** 141–157.

Smedstad, O. M. (1989). Data assimilation and parameter estimation in oceanographic models. Ph.D. Thesis, Florida State University, Gainesville.

Stephens, J. J. (1970). Variational initialization with the balance equation. *J. Appl. Meteorol.* **9,** 732–739.

Stommel, H., and Schött, F. (1977). The beta-spiral and the determination of the absolute velocity field from hydrographic station data. *Deep-Sea Res.* **24,** 325–329.

Sutera, A. (1981). On stochastic perturbations and long-term climate behavior. *Q. J. R. Meteorol. Soc.* **107,** 137–152.

Taft, B. A. (1978). Structure of the Kuroshio south of Japan. *J. Mar. Res.* **36,** 77–117.

Talagrand, O. (1981). A study of the dynamics of four dimensional data assimilation. *Tellus* **33,** 43–60.

Talagrand, O., and Courtier, P. (1987). Variational assimilation of meteorological observations with the adjoint vorticity equation. I. Theory. *Q. J. R. Meteorol. Soc.* **113,** 1311–1328.

Tarantola, A. (1987). "Inverse Problem Theory. Methods for Data Fitting and Model Parameter Estimation." Elsevier, Amsterdam.

Thacker, W. C. (1986). Relationships between statistical and deterministic methods in data assimilation. *In* "Variational Methods in the Geosciences" (Y. Sasaki *et al.*, eds.), pp. 173–179. Elsevier, New York.

Thacker, W. C. (1987). A cost-function approach to the assimilation of asynoptic data. *J. Sci. Comput.* **2,** 137–158.

Thacker, W. C. (1988). Fitting models to data by enforcing spatial and temporal smoothness. *J. Geophys. Res.* **93,** 10,655–10,665.

Thacker, W. C. (1989). Large least-square problems and the need for automating the generation of adjoint codes. *In* "Computational Solution of Nonlinear Systems of Equations" Lect. Appl. Math. Am. Math. Soc. Providence, Rhode Island.

Thacker, W. C., and Long, R. B. (1988). Fitting dynamics to data. *J. Geophys. Res.* **93,** 1227–1240.

Thiébaux, H. J., and Pedder, M. A. (1987). "Spatial Objective Analysis." Academic Press, London.

Thompson, J. D. (1986). Altimeter data and geoid error in mesoscale ocean prediction: some results from a primitive equation model. *J. Geophys. Res.* **91,** 2401–2417.

Thompson, J. D., and Schmitz, W. J. (1989). A limited area model of the Gulf Stream: Design and initial experiments. *J. Phys. Oceanogr.* **19,** 791–814.

Thompson, P. D. (1961). A dynamical method of analyzing meteorological data. *Tellus* **13,** 334–349.

Todling, R., and Ghil, M. (1990). Kalman filtering for a two-layer, two-dimensional shallow-water model. *Proc. Int. Symp. Assimil. Obs. Meteorol. Oceanogr.* 454–459.

Tracton, M. S., Desmarais, A. J., van Maeren, R. J., and McPherson, R. D. (1981). On the system dependency of satellite sounding impact—Comments on recent impact test results. *Mon. Weather Rev.* **109,** 197–200.

Tribbia, J. J. (1979). Nonlinear initialization on an equatorial beta-plane. *Mon. Weather Rev.* **107,** 704–713.

Tribbia, J. J. (1982). On variational normal mode initialization. *Mon. Weather Rev.* **110,** 455–470.

Tziperman, E., and Thacker, W. C. (1989). An optimal control/adjoint equations approach to studying the oceanic general circulation. *J. Phys. Oceanogr.* **19,** 1471–1485.

Tziperman, E., Thacker, W. C., Long, R. B., and Hwang, S. M. (1991a). Oceanic data analysis using a general circulation model: Simulations. In preparation.

Tziperman, E., Thacker, W. C., Long, R. B., Hwang, S. M., and Rintoul, S. R. (1991b). A North Atlantic inverse model using an oceanic general circulation model. In preparation.

Vautard, R., and Legras, B. (1986). Invariant manifolds, quasi-geostrophy and initialization. *J. Atmos. Sci.* **43,** 565–584.

Verron, J. (1990). Altimeter data assimilation into an ocean circulation model: Sensitivity to orbital parameters. *J. Geophys. Res.* **95,** 11443–11459.

Verron, J., and Holland, W. R. (1989). Impacts des données d'altimètrie satellitaire sur les simulations numériques des circulations génerales océaniques aux latitudes moyennes. *Ann. Geophys.* **7,** 31–46.

Wahba, G. (1978). Improper priors, spline smoothing, and the problem of guarding against model errors in regression. *J. R. Stat. Soc., Ser. B* **40,** 364–372.

Wahba, G. (1982). Variational methods in simultaneous optimum interpolation and initialization. *In* "The Interactions Between Objective Analysis and Initialization" (D. Williamson, ed.), Publ. Meteorol. 127 (Proc. 14th Stanstead Seminar), pp. 178–185. McGill University, Montreal.

Webb, D. J. (1989). Assimilation of data into ocean models. *In* "Oceanic Circulation Models: Combining Data and Dynamics" (D. L. T. Anderson and J. Willebrand, eds.), pp. 233–256. Kluwer Academic Publ., Amsterdam.

Webb, D. J., and Moore, A. (1986). Assimilation of altimeter data into ocean models. *J. Phys. Oceanogr.* **16**, 1901–1913.

White, W. B., Tai, C. K., and Holland, W. R. (1990a). Continuous assimilation of simulated GEOSAT altimetric sea level into an eddy resolving numerical ocean model. Part I. Sea level differences. *J. Geophys. Res.* **95**, 3219–3236.

White, W. B., Tai, C. K., and Holland, W. R. (1990b). Continuous assimilation of simulated GEOSAT altimetric sea level into an eddy resolving numerical ocean. Part II. Referenced sea level differences, *J. Geophys. Res.* **95**, 3236–3251.

White, W. B., Tai, C.-K., and Holland, W. R. (1990c). Continuous assimilation of GEOSAT altimetric sea level observations into a numerical synoptic ocean model of the California current. *J. Geophys. Res.* (in press).

Wiener, N. (1949). "Extrapolation, Interpolation and Smoothing of Stationary Time Series with Engineering Applications." MIT Press, Cambridge, Massachusetts.

Wiener, N. (1956). Nonlinear prediction and dynamics. *Proc. Berkeley Symp. Math. Stat. Probab., 3rd* Vol. 3, pp. 247–252.

Williamson, D., ed. (1982). "The Interaction Between Objective Analysis and Initialization," Publ. Meteorol. 127 (Proc. 14th Stanstead Seminar). McGill University, Montreal.

World Ocean Circulation Experiment (WOCE) (1989). "The U.S. Contribution to WOCE Numerical Modelling," Planning Report No. 14. WOCE.

Wunsch, C. I. (1977). Determining the general circulation of the oceans: A preliminary discussion. *Science* **196**, 871–875.

Wunsch, C. (1978). The North Atlantic general circulation west of 50°W determined by inverse methods. *Rev. Geophys. Space Phys.* **16**, 583–620.

Wunsch, C. (1981). Low-frequency variability of the sea. *In* "Evolution of Physical Oceanography" (B. A. Warren and C. Wunsch, eds.), pp. 342–374. MIT Press, Cambridge, Massachusetts.

Wunsch, C. (1988). Transient tracers as a problem in control theory. *J. Geophys. Res.* **93**, 8099–8110.

Wunsch, C. I. (1989a). Using data with models, ill-posed and time-dependent ill-posed problems. *In* "Geophysical Tomography" (Y. Desaubies, A. Tarantola, and J. Zinn-Justin, eds.), pp. 3–41. Elsevier, Amsterdam.

Wunsch, C. I. (1989b). Tracer inverse problems. *In* "Oceanic Circulation Models: Combining Data and Dynamics" (D. L. T. Anderson and J. Willebrand, eds.), pp. 1–78. Kluwer Academic Publ. Amsterdam.

Wunsch, C. I., and Grant, B. (1982). Towards the general circulation of the North Atlantic ocean. *Prog. Oceanogr.* **11**, 1–59.