Advances in Sequential Estimation for Atmospheric and Oceanic Flows

By Michael Ghil

Department of Atmospheric Sciences and Institute of Geophysics and Planetary Physics, University of California, Los Angeles, CA 90095-1565, USA

(Manuscript received 27 June 1995, in revised form 29 May 1996)

Abstract

What: Estimate the state of a fluid system — the atmosphere or oceans — from incomplete and inaccurate observations, with the help of dynamical models. When: After the observations have been made and before making a numerical forecast of the system. If the evolution of the system over some finite time is to be evaluated — i.e., if interested in climate rather than prediction — sequential estimation proceeds by scanning through the observations over the interval, forward and back. How: Admit that the dynamical model of the system isn't perfect either. Assign relative weights to the current observations and to the model forecast, based on past observations, that are inversely proportional to their respective error variances. Yes, but: To compute the forecast errors is computationally expensive. So what: Compromise! The thrust of this review is to illustrate some smart ways of (i) near-optimal, but computationally still feasible implementation of the extended Kalman filter (EKF), while using (ii) the EKF for observing system design, as well as for estimating (iii) the state and parameters of (iv) unstable and strongly nonlinear systems, including (v) the coupled ocean-atmosphere system.

1. Introduction

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 spatiotemporal 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 the emerging discipline of ocean and coupled ocean-atmosphere forecasting, the main purpose is to generate short “video loops,” between one initial state for subsequent prediction and the next; successive initial states are spaced one day — in the atmosphere — or one week to one month — in the oceans — apart (see, for instance, Fig. 1 in Ghil, 1989). 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 (Panel, 1991). The appropriate classes of problems are called filtering, prediction, and smoothing in estimation theory.

The fluid flow is assumed to be governed by a system of evolution equations in which the known, deterministic dynamics is perturbed by system noise, representing subgrid-scale phenomena and other model errors. These governing equations are, in principle, valid at all times, $-\infty < t < \infty$, while observations of the flow are available over a finite time interval, $t_0 < t < t_1$ [see, for instance, Section 7 and Eqs. (7.1)–(7.3) in Ghil and Malanotte-Rizzoli, 1991, GMR hereafter].

The filtering problem is that of determining the best estimate of the fluid’s state at the end of the time interval over which data are provided, $t = t_1$. The solution of this problem is provided, for a linear system with additive noise, by the Kalman (1960) filter. For a nonlinear system, no solution which is both computable in finite time and truly optimal exists. Various near-optimal, computable solutions do exist, and will be discussed in this review.

The prediction problem is that of determining the fluid’s state at times after the last available observation, $t > t_1$. Its solution for zero-mean additive noise in a linear system is simply to use the deterministic part of the governing equations for $t > t_1$. Estimating the initial state of a forecast from data up to initial time and paying no further attention to
these (past) data during the forecast itself is standard practice in NWP (Daley, 1991: D91 hereafter) and, as we see, makes perfect good sense.

The smoothing problem is that of estimating the fluid’s state optimally at times $t$ during the observing interval, $t_0 < t < t_1$. It is therewith the problem appropriate for climate-related feature movies (Bennett, 1992: B92 hereafter; Bennett et al., 1997; Fukumori et al., 1993; Gaspar and Wunsch, 1989; GMR, Sections 6.3.3 and 6.3.4). One of its solutions, the fixed-interval smoother (Caitlin, 1989), involves computing a forward Kalman-filter estimator for intervals $(t_0, t)$ with $t \leq t_1$, a backward estimator with $t_0 \leq t$, and finding the optimal linear combination between the two at each $t \in (t_0, t_1)$. Thus, the Kalman smoother and the “four-dimensional” variational method (Courtier, 1996; Le Dimet and Talagrand, 1986; Lewis and Derber, 1985) of deterministic control theory exhibit certain analogies.

The state approach to sequential estimation was introduced by Kalman (1960) to linear systems with a finite-dimensional state vector in discrete time, and extended by Kalman and Bucy (1961) to such systems in continuous time. At about the same epoch, NWP was becoming operational at a few major national weather services. It was using a procedure (Bergthorsson and Döös, 1955; Cressman, 1959) based on local polynomial interpolation to incorporate observational increments — against a background “first-guess” field — into the objectively analyzed initial state. This amounted to a suboptimal sequential scheme that implicitly linearizes the nonlinear, space- and-time discretized partial differential equations (PDEs) governing large-scale atmospheric flow about the instantaneous state. The problem before us, three and one-half decades later, is to approach convergence in the parallel development of the two strands of research started, on the one hand, with truly optimal estimation for linear systems of ordinary differential equations (ODEs) and, on the other, with highly suboptimal estimation for nonlinear systems of PDEs.

Since the First International Symposium on Assimilation of Observations in Meteorology and Oceanography, in July 1990, the intervening developments — between 1960 and 1990 — were covered by B92, D91, and GMR. The last chapter in D91 has one section on “four-dimensional variational analysis” and one on “the Kalman-Bucy filter.” GMR’s extensive review paper starts essentially where Daley’s book, on atmospheric data analysis, left off and covers these two advanced approaches in some detail, illustrating them with both atmospheric and oceanic applications. B92 reviews Kalman filtering and smoothing for the oceans, and connects them to the variational approach in a nonlinear context. In the present volume, Cohn (1997) provides a general introduction to sequential estimation theory and Ide et al. (1997) propose unified notation for both estimation-theoretic and control-theoretic (i.e., variational) methods and their application to practical data assimilation problems. The duality between sequential estimation and control theory, first shown by Kalman (1960), is discussed at some length by Gelb (1974), GMR (see Table IV and Section 5.4.1 there), and Wunsch (1988).

The purpose of this review is to cover some of the advances in sequential estimation for the atmosphere and oceans since 1990. Section 2 deals with the crucial observability issue, Section 3 with estimating unstable and strongly nonlinear systems, Section 4 with computational issues and filter stability, Section 5 with parameter estimation and coupled ocean-atmosphere models, and Section 6 mentions some open issues.

2. Observability

Cressman (1959) already pointed out the difficulties arising for operational NWP from the sparse and nonuniform distribution of atmospheric observations. The need to adapt objective-analysis methodology to these data-set shortcomings, on the one hand, and to those days’ computational constraints, on the other, led to the use of the successive-correction method (D91). Conventional oceanic observations used to be $10^6$ times fewer than the World Weather Watch (WWW) provides for the atmosphere, and the advent of satellite-borne observations still leaves oceanographers 10 times data-poorer, over all, than their NWP colleagues. Both estimates take into account the smaller space and longer time scales of the ocean (GMR), while oceanic observations are — moreover — confined, by-and-large, to the surface (Munk and Wunsch, 1982).

The question arises, therefore, for the atmosphere — and even more so for the oceans — whether their instantaneous state can be determined from the given data, in the presence of imperfect dynamical models of their large-scale evolution. Intuition suggests that at least as many observations are needed, over a certain time interval, as there are degrees of freedom required to describe the system’s evolution over that interval. A certain data redundancy is clearly necessary to filter out observational and model errors; furthermore, data clumped more closely than the shortest scale of motion we wish to resolve in a given application cannot be all given full weight in this accounting process, as a certain cancellation between their errors should be expected.

A striking result that confirms the sufficient, as well as necessary character of a minimal number of observations — at least under favorable circumstances — was obtained by Ide and Ghil (1997a, b); a preliminary presentation of their work appears in Ghil and Ide (1994), where other qualitative and quantitative observability results are briefly
Fig. 1. Large scale two-dimensional (2-D) flow field (streamlines: dashed) induced by vorticity concentrations (centroid positions: open circles) and associated observing system. Observations may be either Lagrangian (centroid positions: triangles) or Eulerian (velocities: vectors originating from observing stations marked by an ×). The two types of observations can complement each other; both contain errors (courtesy of K. Ide).

described as well. The work of Ide and Ghil (1995; 1997a, b) uses the fact that cyclonic and anticyclonic vorticity concentrations play an important role in organizing large-scale atmospheric flows (McWilliams, 1980). This role is even more important in the oceans, where the Rossby radius of deformation is smaller, and in certain meso-scale atmospheric motions, where diabatic phenomena can greatly intensify nascent vorticity concentrations.

The major role played by coherent vorticity concentrations in geophysical flows competes with and is maintained against planetary-wave dispersion (Holloway, 1983; McWilliams, 1984), as well as the effects of stratification (McWilliams et al., 1994; Rhines, 1979). Still, this role can be exploited by using a Lagrangian description of the flow, that combines prognostic ODEs for the vortex motion with a Poisson PDE for the rest of the flow field. This description goes back to the work of Helmholtz and Kirchhoff in the last century’s second half (e.g., Gröbli, 1877) and was first applied to geophysical, rotating flows by Stewart (1943) and by Morikawa and Swenson (1971). It reduces the number of degrees of freedom necessary to approximate fairly accurately the evolution of the entire flow field to the number of scalars used to follow the change in position and shape of the vorticity concentrations. Observations can be either Lagrangian, such as centroid positions of vortex blobs, or Eulerian, such as velocity measurements at fixed locations (see Fig. 1).

As a first step, Ide and Ghil (1997a) consider point-vortex systems in a barotropic flow. A point-vortex system, simple as it is, can experience chaotic motion when the number of vortices exceeds four (Aref, 1984). It turns out that this motion can be tracked by the extended Kalman filter (EKF) when a comparable number of Lagrangian observations, such as satellite-relayed positions of drifting buoys, are available. Technical difficulties arise, however — due to the very simplicity of this approximation — when using Eulerian observations, such as ship stations.

The next level of approximation of a Lagrangian model for isolated, coherent vortex structures is that of Rankine vortices (Ide and Ghil, 1997b), having a finite-radius core in solid-body rotation and a logarithmic velocity profile, like point vortices, outside this core. For such a system, regular as well as chaotic motion can be well tracked — when using the EKF — by an arbitrary combination of Eulerian and Lagrangian observations, provided their number is approximately equal to that of the vortices. The updating interval that is required to achieve good tracking with such a small number of observations has to be comparable with and slightly shorter than the time scale of vortex motion.

This rather neat result is illustrated for a system of four Rankine vortices, subject to stochastic perturbations that model subgrid-scale phenomena, in Fig. 2. It is clear from Fig. 2 that the number of velocity observations (e.g., ship stations or current meters) necessary to track the four vortices is typically four, and rarely larger than six.

The quantitative observability results above are quite encouraging, and might apply to the more common Eulerian models of atmospheric and oceanic flows. One rather straightforward way of applying the Lagrangian concepts on which these results are based to Eulerian oceanographic models is by the use of “feature models” (Robinson et al., 1989) or “contour analysis” (Mariano, 1990); it is these ideas, in fact, which inspired the work of Ide and Ghil (1997a, b) in the first place. The work is being extended to elliptic vortices (Ide and Wiggins, 1994) and to two-layer, baroclinic “hetons” (Hogg and Stommel, 1985) or “hermons” (Ide and Ghil, 1995).

Another result that indicates the need for only a minimal number of observations, this time applicable directly to Eulerian flow models, is due to Todling and Ghil (1994). They applied the Kalman filter to a linear, spatially two-dimensional (2-D) shallow-water (SW) model of the mid-latitude atmosphere in a zonally periodic β-channel. In the one-layer, barotropic version of the model, vigorous
instability is induced by a cosine-square shaped profile of the basic zonal flow. The rapid growth of
the barotropic instability (Fig. 3a) can be tracked by a single well-placed observation (Fig. 3b). This is due
to the fact that the flow is actually dominated by the spatial pattern of the instability, which is "known"
to the model dynamics; the observation is merely used, therefore, to determine this pattern's instantaneous amplitude at a given location. A similar result applies in the model's two-layer version, with respect to baroclinic instability (Todling, 1992).

Another interesting set of observability results concerns the relative usefulness of mass-field and velocity-field observations in determining the state of the flow system. This is a classical question, going back in meteorology to the planning stages of the Global Atmospheric Research Programme (Charney et al., 1969; Smagorinsky et al., 1970) and in tropical oceanography to Philander et al.'s (1987) work. Partial — and somewhat contradictory — answers were reviewed by Ghil (1989) and took into account various aspects of near-geostrophic dynamics and of relative error sizes in the mass and velocity data.

Jiang and Ghil (1993) built on these earlier results to provide a criterion for either type of data being more useful than the other, in terms of the product $\rho K_G K_E$ of three nondimensional numbers being larger than or less than one: (i) $\rho$ measures the degree to which the forecast errors in mass and velocity are geostrophically related, (ii) $K_G$ measures the scale of the motions observed with respect to the Rossby radius of deformation, and (iii) $K_E$ measures the relative size of observing errors in mass and velocity. For $\rho K_G K_E < 1$, mass data are more useful; velocity data are more valuable when $\rho K_G K_E > 1$. This analytic result, obtained for idealized single-wave flows, was verified numerically by Jiang and Ghil (1993) using the Kalman filter for a linear SW model in a rectangular mid-latitude ocean basin.

Based on knowledge about the underlying flow dynamics and various measurement techniques, an observing system can be designed to provide efficient and successful estimation. Observing-system optimization is critical for the oceans, where each data point counts, due to the limited number of subsurface observations. Barth and Wunsch (1990) and Barth (1992) considered specific optimization problems for acoustic-tomography arrays in linear ocean models, steady-state and time-dependent, respectively; these problems were solved by simulated annealing.

A fairly general solution to the observing-system,
3. Unstable and strongly nonlinear systems

Complete results on truly optimal estimation — whether on observability, controllability or filter stability and convergence (Gelb, 1974; Bucy and Joseph, 1987) — are only available for linear systems, mostly of ODEs. But the equations of motion for planetary flows are nonlinear and admit competing instabilities — barotropic, baroclinic, topographic and others — that make geophysical fluid dynamics such an interesting field of study and NWP, as well as its recently developing oceanic and coupled ocean-atmosphere counterparts, so difficult. Observations, moreover, can be nonlinear functionals of the states (Jazwinski, 1970; Gelb, 1974; Ide and Ghil, 1997a, b).

As mentioned already in Section 2, instability can actually help estimation in a 2-D linear SW model (Todling and Ghil, 1994). For the two-layer version of the model, baroclinic instability requires for its tracking a larger, but still very modest number of observations \( p \), compared to the total number of discrete model variables \( N \), \( p \ll N \) (Ghil and Todling, 1996). This result is consistent with the large effect of single observations found by Källén and Huang (1988) in limited-area forecast verifications. Rabier and Courtier (1992) have also shown that the adjoint algorithm for strong-constraint variational methods converges in many baroclinically unstable situations.

A classical way of dealing with nonlinearities (Gelb, 1974; Jazwinski, 1970) is the EKF, in which the equations of motion and the observing functions are linearized about the instantaneously estimated state at every time step: the estimate of the state itself evolves between updates according to the full, nonlinear governing equations, while the forecast-error covariance matrix evolves according to the linearized “tangent” system. The advective nonlinearity of the geophysical flow equations has typically small effects over the short time intervals between synoptic updates in NWP. Based on this fact and on the relatively good performance of the definitely suboptimal filters then in use at operational NWP centers (D91), Ghil et al. (1981) hypothesized that linearization in NWP would only be necessary at update times, every 6–12 hours, or even less frequently. Moreover — based on their results for a one-dimensional (1-D), linear SW model — they also suggested that, in between linearizations, an asymptotically valid, steady-state, Wiener

or “antenna,” design problem for a linear model is provided by B92. The measurements are defined by linear functionals on the solution space, called “individual antenna elements.” The independent observations are not, in general, associated with these, but with "principal antenna elements" that result from the analysis by diagonalizing appropriate matrices. A number of interesting applications are given in B92.

Ide and Ghil (1997a), in turn, explain analytically the contribution of Eulerian and Lagrangian observations to a vortex position at update time. Based on this description, the observing system can be optimized analytically (Ide and Ghil, 1997b), in terms of the distance between stations, their placement, and certain internal parameters of the assimilation method.
filter would provide an excellent approximation to the time-dependent EKF.

Lacarra and Talagrand (1988) showed that, in fact, for a 2-D nonlinear SW model on an f-plane, the tangent linear approximation captures most of the error growth for 24–36 hours. This is consistent with the numerical studies of Balgoyind et al. (1983) for a multi-layer, semi-operational primitive-equation model of the global atmosphere. A constant-coefficient version of the tangent linear approximation — which would be consistent with the Wiener filter mentioned above — still reproduces rather well the behavior of the fastest-growing, large-scale Rossby waves for up to 48 hours in Lacarra and Talagrand’s simple atmospheric model.

As NWP model resolution has increased and smaller-scale, convective phenomena (inter alia) are better resolved, the extended validity of such model linearizations has to be revisited, given the more rapid growth — and also rapid saturation and lower global energy levels — of the smaller-scale phenomena. Moreover, these promising concepts cannot be expected to hold, even for the largest scales, in the presence of filter divergence (see Section 4) or of strong nonlinearities arising from the presence of multiple attractors and associated with the boundaries between their basins. Miller et al. (1994) studied the performance of both variational methods and sequential estimators for a stochastically perturbed double-well potential. The EKF can track the scalar trajectory from one well into the other, provided the observations are accurate or frequent enough. A strong-constraint variational method (Sasaki, 1970) cannot produce a transition, since the “perfect,” purely deterministic version of the model that such a method assumes exhibits no transition between the two wells. A weak-constraint, four-dimensional (4-D) method can improve slightly upon an EKF-provided first guess, due to the greater smoothness of a trajectory based on later, as well as earlier observations.

The double-well problem can give some insight into the interaction of stochastic perturbations — representing subgrid-scale phenomena in low- and even high-resolution models — with nonlinear, but stable, dynamics. The Lorenz (1963) model allowed Miller et al. (1994) to inspect the performance of advanced data assimilation methods for strong deterministic nonlinearities in the presence of vigorous instabilities. The standard EKF loses track of the trajectory in the neighborhood of the model’s Z-axis, where the solution has to “decide” between circling one unstable convective state or the other (Fig. 4a) and where the estimated error variance increases sharply (Fig. 4b). The variational strong-constraint method also misses transitions between one “wing” of the Lorenz “butterfly” and the other, or crosses over when it should not (Fig. 4c), due to the presence of multiple minima in the cost functional that measures the distance between the trajectory being estimated and the observations (Fig. 5; see also Gauthier, 1992).

It is clear from Fig. 5 that the number of minima increases as the time interval $T = t_1 - t_0$ over which the minimization is carried out increases. We strongly suspect that, in the limit of increasing $T$, the graph of the cost functional becomes fractal for the Lorenz model, its distinct minima being associated with distinct sheets of the attractor. Multiple minima have been documented for fair-sized ocean-o-
graphic models (e.g., GMR, Fig. 28, and discussion thereof). Li (1991) has studied a resonant triad of Rossby waves and shown that the complexity of the cost functional’s “landscape,” when plotted against the initial amplitude of the Rossby waves, increases as T increases. White (1993) has shown rigorously for Burgers’ equation that a unique minimum obtains only for sufficiently short time intervals.

Two modifications of the EKF can be implemented for large-scale models: the “sanity checker” used by engineers to reinitialize the EKF when the observational residuals — between the actual observation and the model’s prediction of that observation — give clear indications of divergence; and an empirical statistical error model based on Monte-Carlo simulation. Both permitted very satisfactory solution tracking for indefinite time in the Lorenz model. Monte-Carlo simulation can also serve to reduce the computational burden of the standard EKF (see Section 4). No similar remedy for the malfunction of the 4-D variational (4DVar) method (e.g., Courtier, 1997) was found in this instance. In practice, 4DVar seems to work best when using continuous cycles based on relatively short intervals, and is provided a priori with good guesses of a background field and its error-covariance matrix.

Increasingly complex solution behavior arises in a nonlinear system as the forcing increases. The typical succession is from steady states, through periodic and low-order chaotic solutions (Lorenz, 1963; Ghil and Childress, 1987), to fully turbulent ones. Actual planetary flows can be described — according to the resolution afforded by the observations, the model assimilating them, and the data-assimilation method being used — as being in either in the chaotic or the fully turbulent regime. Multiple equilibria and limit cycles serve as roadposts on our way to understanding the observed behavior. They can do so in helping us explore in turn, and hence improve, the performance of data-assimilation methods.

Jiang (1994) studied the performance of improved optimal interpolation (IOI: Daley, 1992b; Todling and Cohn, 1994) in a nonlinear version of Jiang and Ghil’s (1993) reduced-gravity SW model, subject to random perturbations. It was shown that, in a forecast mode, the model loses track of transitions between one possible equilibrium and the other, or of the phase of the periodic solution that prevails for stronger wind stress. The equilibria are distinguished by the subtropical or subpolar gyre of the model’s double-gyre circulation being stronger, and the periodic solution by interannual shifts in the position and intensity of the eastward jet that separates the two gyres (Jiang et al., 1995; Speich et al., 1995).

Jiang and Ghil (1995, 1997) showed that the IOI could recapture, within a few weeks of altimetric data use, the correct steady state or phase of the periodic solution (Fig. 6). The deviations of the instantaneous values of the thickness of the model’s upper, active layer from its average value at the same location over the solution’s period are called “anomalies” (following meteorological and climatological usage) in the figure’s caption and are strongest inside and near the changing triangle formed by the two separation points of the north- and southward western boundary currents and their confluence into the eastward jet. The evolution in time of these anomalies is shown along a line parallel to the rectangular basin’s western boundary, crossing the zonal symmetry axis of the wind stress near the confluence point.

Strong positive and negative anomalies alternate along this symmetry axis (y = 1000 km in the figure). The shift in meridional position of anomalies of one sign is either sudden (at t \( \geq 12 \) months) or gradual (t \( \geq 30 \) months), indicating that these are relaxation oscillations (Jiang et al., 1995). The phase of these relaxation oscillations (Fig. 6b) is
completely lost after 15 months of stochastic perturbations, with negative anomalies replacing positive ones along the axis of zero wind-stress curl (Fig. 6a). Assimilation of altimetric data with a 12-day repeat period starting at $t = 18$ months restores the correct phase and general shape of the periodic solution (Fig. 6c). Similar results were obtained by Jiang and Ghil (1997) for aperiodic solutions (not shown here).

4. Computational implementation and filter stability

A major obstacle to more extensive EKF implementation is the computational burden of advancing in time the forecast-error covariance matrix $P^f$ (notation here follows Ide et al., 1997). For a full $N \times N$ matrix $P^f$, this requires $O(N^2)$ operations (Boggs et al., 1995). As forecast and simulation models tend to keep pace with computational developments in increasing their spatial resolution or physical complexity, and hence $N$, it will be necessary to devise $O(N \log N)$ implementations, independently of current and prospective computer architectures.

GMR outlined a number of ideas for doing so (Section 5.3.1 and references there; see also Todling, 1997). These include: 1) various forms of covariance modeling that improve upon the currently operational “optimal interpolation” (OI: D91), such as spectral statistical interpolation (Parrish and Derber, 1992), three-dimensional variational analysis (Courtier, 1997) or physical-space statistical analysis (Guo and Da Silva, 1997); 2) using sim-
plified dynamics to advance the forecast-error covariances, for instance restricting $P^f$ to the slow, Rossby modes only (Cohn and Parrish, 1991) or advecting the covariance matrix of height-field errors only, with height-wind and wind-wind errors computed diagnostically by the geostrophic relation, as in OI (Dee, 1991); 3) implementing the EKF at lower resolution than the dynamics (LeMoyne and Alvarez, 1991); 4) using a localized approximation for $P^f$ (Riedel, 1993); and 5) using an asymptotic, steady-state filter (Heemink, 1988).

A particularly interesting form of local approximation is based on the observation that forecast-error correlations decay over a horizontal distance comparable to the Rossby radius of deformation (Balgovind et al., 1983); this in turn is comparable to the distance covered by a finite-difference stencil used in discretizing the flow PDEs to yield the dynamics, or “model” matrix $M$. Thus $M$ is exactly (block-) banded and $P^f$ approximately so. Parrish and Cohn (1985) used diagonalization by diagonals to achieve an $O(N \log N)$ algorithm for advancing the banded matrix $P^d$ obtained by neglecting the small correlations beyond a prescribed distance $d$ altogether. This algorithm performed well in approximating the full Kalman filter for a 2-D linear SW model in a mid-latitude $\beta$-channel, as long as the positive definiteness of $P^d_k(k\Delta t) \equiv P^d_k$ could be maintained; the time step is $\Delta t$ and $k$ is the number of steps since the beginning of the assimilation cycle. Computational performance on a vector machine improved as $N$ increased (see also Table III in GMR).

The loss of positive definiteness of $P^d_k$ with increasing $k$ is not a matter of concern for the banded approximation only. Many applications of the classical Kalman-filter implementation have been shown to diverge, due typically to the appearance of negative eigenvalues of the full $P^f_k$ (Schlee et al., 1967). Evensen (1992) documented a similar occurrence of EKF divergence for a multi-layer quasi-geostrophic (QG) ocean model, while Gauthier et al. (1993) showed in an atmospheric nondivergent barotropic model the occurrence of unbounded growth of errors in the tangent linear model over data-void areas.

Stabler implementations of both the linear Kalman filter and the EKF result from square-root algorithms (Bierman, 1977), which conserve the positivity of the error-covariance matrices involved by construction, in the same way that the square of a real scalar is necessarily a positive scalar. Boggs et al. (1995) implemented such a square-root filter — in which the matrix $P^d_k$ or $P^d_k$ is factorized at $k = 0$ and then it is the factors that are advanced in time — for a nonlinear 2-D, two-layer SW model, with realistic forcing and topography. This provided stable and accurate performance of the banded approximation (Fig. 7). A square-root filter with a different idea for rank reduction was implemented in a 1-D linear SW model by Verlaan and Heemink (1995).

The numerical experiment illustrated in Fig. 7 applies to the model’s single-layer, barotropic version, in a rectangular domain centered at 45$^\circ$N, on a $17 \times 24$ grid yielding 1224 state variables. Model errors in the zonal and meridional velocities $u$ and $v$ and in free-surface height $h$ are assumed to have standard deviations of $10$ ms$^{-1}$, $5$ ms$^{-1}$ and 200 m per half day, respectively, while the corresponding observational errors are 2 ms$^{-1}$, 0.75 ms$^{-1}$ and 75 m. Updates occur every 3 hours and the figure shows clear convergence of the estimates to the true state.

Very similar results obtain for a bandwidth $b$ — i.e., the number of grid points, away from a base point in each coordinate direction, over which nonzero covariances are calculated — equal to 2, 3 and 4; the number of nonzero diagonals $2b + 1$ of $P^d$ is given by $b = 3n_x(2b + 1)$, where is the number of longitudinal grid points, with $n_x = 17$ in the figure. The resulting state and error-covariance estimates did not differ in any significant way from those obtained using the full square-root filter, while the classical, nonfactorized implementation of the EKF led invariably to filter divergence. Boutilier (1994) prevented suboptimal-filter divergence in a simplified version of an operational NWP model by intro-
ducing a relaxation of the forecast-error covariances to climatological values, i.e., by assuming error saturation at or near the level of natural variability.

Fukumori et al. (1993) implemented a steady-state, fixed-interval Kalman smoother (see Section 1 here, B92 and GMR) — related to the Wiener-filter idea mentioned early in Section 3 — to assimilate data into a low-resolution version of Haidvogel et al.'s (1991) semi-spectral primitive-equation model, linearized about a state of rest. Both synthetic and Geosat altimeter data were used in this efficient implementation, but the data were too noisy and the model too crude to account for a large fraction of the Geosat residual variance.

Fukumori and Malanotte-Rizzoli (1994) explored further — for an idealized Gulf-stream model with $N = O(10^3)$ — the ideas of linearization about a fixed nonzero state and of rank reduction of $P^f$ by retaining only large-scale modes. They applied this suboptimal, approximate filter to observing-system simulation experiments (OSSE) using simplified altimetric and tomographic observing patterns (Fig. 8). For their Gulf-stream model and sea-level vs. mean-velocity measurements, this OSSE shows lower estimation errors for the altimetric (heavy solid) than tomographic (light solid) observing system.

Yet another class of simplified, approximate implementations of the EKF proceed from Monte Carlo ideas. Evensen (1994) proposed such a method to approximate $P^f$ in a QG ocean model. Miller et al. (1994) used Monte-Carlo simulation to approximate higher-moment evolution in the Lorenz model. It is clear that in any such suboptimal-filter implementation, like in many other Monte-Carlo applications (e.g., multi-dimensional integration), the use of particularly skillful sampling methods — such as stratified sampling (Balgovind et al., 1983), importance sampling (B92) or a determination of the most rapid error-growth directions between updates (Buizza et al., 1993; Lacarra and Talagrand, 1988; Toth and Kalnay, 1993) — is essential to ensure the adequacy of a small sample.

Daley (1992a) proposed a method for propagating variances only in a 1-D linear QG model, using fixed correlations, as in OI. Cohn (1993) extended this approach to multiple space dimensions, showing how to generalize it to allow for partial changes in the correlation structure as well. Todling and Cohn (1994) formulated a framework for the systematic comparison of various suboptimal data-assimilation schemes, and gave relative “ratings,” between OI and the full Kalman filter, for the performance of five such schemes in a 2-D linear, $\beta$-plane SW model.

The importance of “feature movies,” rather than “video loops,” for climate studies (Bengtsson and Shukla, 1988; Panel, 1991) has led to a number of so-called reanalysis efforts of past meteorological data, aimed at obtaining consistent long-term data sets with known error bars (Kalnay and Jenne, 1991; Schubert et al., 1993). In this context, Cohn et al. (1994) have pointed out that fixed-lag Kalman smoothing (FLKS: Anderson and Moore, 1979) is intrinsically more efficient than fixed-interval smoothing (Gaspar and Wunsch, 1989; Fukumori et al., 1993), while Ménard and Daley (1996) have shown that the use of 4DVar methods based on the adjoint approach is suboptimal with respect to fixed-interval smoothing. Cohn et al. (1994) proposed, in fact, a particularly efficient implementation of the FLKS and applied it to the 2-D linear SW model of Parrish and Cohn (1985).

5. Parameter estimation and coupled models

In the discussion so far, both the dynamical and the statistical models were assumed to be perfect, i.e., all the coefficients in the dynamics and parameters in the error models were assumed to be known. This, of course, is an idealization and various details of atmospheric and oceanic dynamics, as well as of the observing-system errors and subgrid-scale noise are not known. Adaptive filters which determine certain parameters on line have a long tradition in the engineering literature (Gelb, 1974; Ljung, 1987).

Dee et al. (1985) provided a computationally efficient implementation of an adaptive filter that estimates the model-error and observational-error covariances under reasonable assumptions on the structure of these matrices. Dee (1995) modified and extended this work in a number of ways that allow for the estimation of a fairly large number of statistical-model parameters in suboptimal data-assimilation schemes, including OI and modifications thereof. In Section 4, covariance modeling was discussed from the point of view of efficient numerical filter implementation. The lack of sufficient statistical information on flow-field and observing-system error structure is an equally strong incentive for ingenious covariance modeling (Dee, 1991, 1995).

A particularly interesting class of poorly known dynamic parameters are those associated with the coupling of oceanic and atmospheric models. Coupled tropical-ocean/global-atmosphere models are being used for the experimental prediction of seasonal-to-interannual climate variability (Cane et al., 1986). The qualitative behavior of such models, periodic or chaotic, depends crucially on the strength of the coupling between the atmosphere and ocean, expressed as a constant of proportionality $\mu$ between the wind stress $\tau$ and the sea-surface temperature $T_s$ (Jin et al., 1994, 1996). Hao and Ghil (1994) showed that the nature of the wind-stress error, systematic or random, can strongly affect the effectiveness of ocean-data assimilation in a reduced-gravity SW model on an equatorial $\beta$-plane.

Smedstad and O'Brien (1991) had estimated the reduced gravity, and hence the speed of internal
waves, in a noise-free SW model by a variational method. Hao and Ghil (1995) applied the EKF to the Jin et al. (1994, 1996) intermediate coupled model in the presence of observational and model error to estimate the ocean-atmosphere state, as well as the coupling parameter \( \mu \) and an upwelling parameter \( \delta_s \), from simulated TOGA-TAO (Hayes et al., 1991) array data. Estimating the model parameters helped considerably the state estimation (Fig. 9); convergence to the correct parameter value (Fig.

Fig. 8. Relative RMS errors, normalized by the errors of the perturbed run with no updates, when assimilating altimetric (heavy solid) or reciprocal tomography (light solid) data: a) barotropic stream function, b) surface zonal velocity, and c) mid-depth density (from Fukumori and Malanotte-Rizzoli, 1994).

Fig. 9. Time-longitude plots of sea-surface temperature (SST) anomalies along the Equator for: a) forecast using the model with incorrect \( \mu \) and \( \delta_s \), b) reference solution with no wind-stress error, and c) assimilation results using the simulated TOGA-TAO array data to correct both \( \mu \) and \( \delta_s \), as well as the model state. The contour interval is 0.5 degree (from Hao and Ghil, 1995).
10) depended on the number of data available and the sensitivity of the state-estimation error to the parameter value.

In the examples given here, dynamical parameters were assumed known when estimating statistical parameters, and vice-versa. In strongly unstable and nonlinear systems (see Section 3 and references there), the simultaneous estimation of both dynamical and statistical parameters can be rendered more difficult by the neglect of third- and higher-order moments in EKF-related schemes. This might be an additional reason for considering Monte-Carlo-based methodology in the context of parameter estimation as well.

6. Open questions

Instead of summarizing this short paper, which is itself but a summary of research presented in greater detail elsewhere, it seems appropriate to extract from each section of its main text a set of two or three questions that can help focus future research in the corresponding area.

i) What is an optimal observing system for the ocean-atmosphere system's evolution on time scales from hours to decades? Compared to the enormous cost of deploying and maintaining instruments, is sufficient information being returned by existing and currently planned systems to improve significantly our understanding and prediction of the planet's fluid envelope? Can advanced observing-system design increase the benefit-cost ratio?

ii) Can the advanced estimation methods applied successfully to strong instabilities and nonlinearities in low-order or simple models be extended to full-blown simulation and prediction models? Given the cost to life and property of a single wrong forecast, can the computational cost of advanced data assimilation be justified by a predictable reduction in the number of such "misses"?

iii) What is the optimal trade-off between the computational cost of simulation or forecasting, on the one hand, and that of data assimilation, on the other? Should the design and implementation of future atmospheric, oceanic and coupled models include that of appropriate data-assimilation methods associated with each model? How important is an estimate of simulation or forecasting error?

iv) Can parameter estimation become a systematic tool for improving geophysical flow models? Can it help solve, in particular, the crucial coupling problem for tropical and global ocean-atmosphere and other coupled models?

Acknowledgments

It is a pleasure to acknowledge my recent learning from D. Boggs, F. Gauthiez, Z. Hao, K. Ide, S. Jiang, C. Keppenne, R.N. Miller, and R. Todling, much more than the references to our joint papers do. The stimulation from the participants of the WMO International Symposium in Tokyo, March 1995, was crucial in producing this review, as was the support over the years from NASA's Grant NAG-5713 and ONR's Grant N00014-93-1-0673 (and its predecessors). D. Boggs, I. Fukushima, Z. Hao, K. Ide, S. Jiang, R.N. Miller, and R. Todling kindly provided original figures. K.E. Hartman typed successive versions and organized the graphics most efficiently, helped in the latter by W. Weibel. This review was completed while holding the Condorcet Visiting Chair of the Ecole Normale Supérieure, Paris, and enjoying the hospitality of its Département Terre-Atmosphère-Océan and of the Laboratoire de Météorologie Dynamique du CNRS at the Ecole. Conversations with and re- or preprints from S.E. Cohn, D.P. Dee, K. Ide, B. Legras, and R. Todling helped put the finishing touches. Perceptive comments from D.P. Dee,
K. Ide, J.C. McWilliams, R. Todling and two anonymous reviewers were useful at the revision stage. This is publication no. 4859 of UCLA’s Institute of Geophysics and Planetary Physics.

References

Ghil, M., 1989: Meteorological data assimilation for oceanographers. Part I: Description and theoretical


McWilliams, J.C., 1984: The emergence of isolated coherent vortices in turbulent flow. J. Fluid Mech., 140, 21–43.


