The Behavior of Forecast Error Covariances for a Kalman Filter in Two Dimensions

STEPHEN E. COHN
Courant Institute of Mathematical Sciences, New York University, New York, New York

DAVID F. PARRISH
Development Division, National Meteorological Center, Washington, D.C.

(Manuscript received 2 July 1990, in final form 14 December 1990)

ABSTRACT

A Kalman filter algorithm is implemented for a linearized shallow-water model over the continental United States. It is used to assimilate simulated data from the existing radiosonde network, from the demonstration network of 31 Doppler wind profilers in the central United States, and from hypothetical radiometers located at five of the profiler sites. We provide some theoretical justification of Phillips' hypothesis, and we use the hypothesis, with some modification, to formulate the model error covariance matrix required by the Kalman filter.

Our results show that assimilating the profiler wind data leads to a large reduction of forecast/analysis error in heights as well as in winds, over the profiler region and also downstream, when compared with the results of assimilating the radiosonde data alone. The forecast error covariance matrices that the Kalman filter calculates to obtain this error reduction, however, differ considerably from those prescribed by the optimal interpolation schemes that are employed for data assimilation at operational centers. Height-height forecast error correlation functions spread out broadly over the profiler region. Height-wind correlation functions for a base point near the boundary of the profiler region are not antisymmetric with respect to the line of zero correlation, nor does the zero-line pass through the base point.

We explain why these effects on forecast error correlations are to be expected for wind profilers, which provide abundant wind information but no height information. Our explanation is supported by further experiments in which height observations assimilated from radiometers at just a few profiler sites reduce these effects.

1. Introduction

Optimal interpolation (OI) is currently the most widely used analysis procedure in meteorological data assimilation systems (Gandin 1963; Bergman 1979; McPherson et al. 1979; Lorenc 1981; Baker et al. 1987; Dimeto 1988). The success of OI depends on accurate knowledge of the error statistics of the observed data and of the forecast produced by the assimilating numerical weather prediction (NWP) model. Computational considerations, and also the lack of a complete theory, have led to the practice of specifying these statistics by way of simple models with little variation in space or time.

The Kalman, or Kalman–Bucy filter (Kalman 1960; Kalman and Bucy 1961; Jazwinski 1970; Gelb 1974; Anderson and Moore 1979) offers a systematic way to calculate the actual evolution of forecast error statistics in time and, in principle, provides the optimal estimate of the evolving atmospheric state. The distinguishing feature of the Kalman filter is not the actual analysis of data, as the same analysis equations appear in OI. Rather it is the evolution of the forecast/analysis error covariance matrix according to the dynamics of a forecast model.

Meteorological and oceanographic applications of the Kalman filter have been studied, so far using idealized one- and two-dimensional models (Cohn et al. 1981; Ghil et al. 1981; Dec et al. 1985a,b; Parrish and Cohn 1985; Miller 1986; Bennett and Budgell 1987, 1989; Dec 1990a,b). Extension to realistic NWP models is made difficult by the large computational burden imposed by the Kalman filter evolution equations. At the very least, however, the Kalman filter is a valuable pedagogical device. Two-dimensional studies can provide considerable insight into the nature of forecast error statistics and the impact of approximations used in the OI methodology.

The adjoint method is another approach that has been proposed recently for meteorological data assimilation (Lewis and Derber 1985; Le Dimet and Talagrand 1986; Courtier and Talagrand 1987; Talagrand and Courtier 1987; Thacker and Long 1988). This variational approach determines the initial conditions that would lead to the given data observed later, to the extent dictated by observation errors, and which would also be consistent with the forecast model equations.
The adjoint method appears to be considerably less expensive computationally than the Kalman filter, but it does not naturally account for the presence of model error. Derber (1989) has suggested, however, a straightforward modification of the adjoint method that estimates the first-order statistics of the model error, i.e., the model bias. Lorenc (1988b) has also taken some account of model error in the adjoint approach, to the extent that OI schemes do, by incorporating an OI estimate of the forecast error covariance matrix.

The relationship between Kalman filtering and the adjoint method has been studied by Lorenc (1986, 1988a) and by Bennett and Budgell (1987, 1989). In particular, the adjoint method, when its iterations are carried all the way to convergence, is known to be algebraically equivalent to the Kalman filter-smoother (cf., Jazwinski 1970, examples 7.2, 7.8) in the case of a linear perfect model.

The Kalman filter replaces knowledge of the forecast error covariance matrix, which is assumed in OI, by assumed knowledge of the model error covariance matrix, which is a much more basic quantity. The model error, or incremental error committed over one time step of the model forecast, is not influenced by past observations as the forecast error itself is. It arises solely from discretization error and error committed in physical parameterizations. Still, the model error covariance matrix is not actually known, so it must be estimated. This can be done adaptively, during the course of data assimilation (Dee et al. 1985a,b; Dee 1990a), or else offline, in some approximate fashion.

Either way, it would be useful to represent the model error covariance matrix in terms of a small number of parameters. Phillips (1986) has suggested what amounts to such a parameterization, by hypothesizing that model errors consist of slow modes only, whose amplitudes are uncorrelated with one another. This would leave about \( n/6 \) parameters to be estimated, \( n \) being the number of degrees of freedom of the forecast model, instead of all \( n(n + 1)/2 \) free parameters (elements) of the symmetric \( n \times n \) model error covariance matrix. He further hypothesizes an equipartition of energy among the slow modes. This would leave only one free parameter.

The Kalman filter is ideal for carrying out observing system simulation experiments (OSSE), for one can observe directly how different observing networks affect the forecast error covariance matrix, and in particular the forecast error variances. In this paper we simulate the North American radiosonde network first, to provide benchmark results, and then we simulate the additional observations provided by the demonstration network of 31 wind profilers (Beran 1986; Kuo and Guo 1989). Finally we simulate additional observations from a hypothetical network of five ground-based radiometers (Westwater et al. 1984).

Our forecast model is a linearized shallow-water model with 200-km resolution over a \( 3000 \times 5000 \) \( \text{km}^2 \) domain roughly circumscribing the continental United States. Several steps are taken to improve the computational efficiency of the Kalman filter, but no approximations are made. We adopt Phillips' (1986) hypothesis to the extent of assuming model errors to consist of uncorrelated slow modes. Instead of assuming an equipartition of energy, however, we employ an energy spectrum that results in forecast error correlations (for the radiosondes-only simulation) similar to those of realistic NWP models, and which respects the regularity criterion of Bennett and Budgell (1987, 1989).

Thus, the purpose of this paper is to demonstrate the behavior of forecast error covariances in a fairly realistic (albeit linear) setting. We find that the observations can have a profound effect upon forecast error variances and correlations, and upon their evolution in time. In particular, the introduction of observations from wind profilers causes the height–height forecast error correlations to become extremely broad, and causes associated changes to height–wind and wind–wind cross correlations. This behavior is not accounted for in the current OI methodology, and we explain why it occurs. We show also that height observations from radiometers collocated with just a few of the wind profilers lessens the broadening of the height–height correlations and the associated effects on the cross correlations.

The choice of the model error covariance matrix is also found to exert considerable influence upon the forecast error covariance matrix, as expected (cf., Phillips 1982). This influence is important, for it affects our quantitative, if not qualitative, results concerning the impact of the profiler observations. Tribbia and Baumhefner (1988) have also pointed out the importance of treating model error in the conduct of OSSE studies. For these reasons we also provide a rigorous theoretical justification of our adaptation of Phillips' (1986) hypothesis.

Specifically, we show that assuming model errors to consist of slow modes only is equivalent to asking the Kalman filter to perform initialization automatically at each analysis time. If one accepts both the desirability of initialization and the optimality of the Kalman filter, one must therefore accept this assumption also. Further, we show that assuming model errors to be uncorrelated in mode space is essentially equivalent to assuming that the contribution to the forecast error covariance matrix from accumulated model error covariances does not propagate in time. This is justified when most of the forecast error covariance propagation is due to assimilation of the observations themselves. We demonstrate theoretically and numerically that, for realistic observing patterns, substantial propagation due to data assimilation must in fact always occur.

We caution the reader that our results, both theoretical and experimental, apply strictly to linear dynamics. The advantage of our use of linear dynamics
is that numerous questions concerning data assimilation can then be answered quite precisely. The disadvantage, of course, is an attendant lack of realism.

The plan of the paper is as follows. In section 2 we review the Kalman filter approach to data assimilation. Section 3 describes the shallow-water model, the simulated observing systems and their assumed error characteristics, and our Kalman filter implementation. The model error covariance matrix is formulated in section 4, and the numerical experiments are reported in section 5. A summary and conclusions appear in section 6.

Two appendices give the theoretical justification of Phillips' (1986) hypothesis. Appendix A describes the relationship between model error and initialization, and appendix B shows how covariance propagation is influenced by model error and by data assimilation.

2. The Kalman filter

We describe the linear case first. Suppose we are given a linearized numerical forecast model

\[
\mathbf{w}_k' = \mathbf{A}_{k-1} \mathbf{w}_{k-1}'.
\]  

(2.1)

Here \( \mathbf{w}_k' \) is the forecast vector at time \( t_k \), and \( \mathbf{w}_{k-1}' \) is the analysis vector at the previous time step; we simply define \( \mathbf{w}_k' = \mathbf{w}_{k-1}' \) if no observations were analyzed at time \( t_{k-1} \). Both the forecast and analysis vectors have length \( n \), the number of degrees of freedom of the model. In a finite-difference model, say, \( n \) is the number of grid points times the number of prognostic variables. The \( n \times n \) dynamics matrix \( \mathbf{A} \) may depend on time, \( \mathbf{A} = \mathbf{A}_{k-1} \), but does not depend on \( \mathbf{w}_{k-1}' \).

Since the objective of an analysis system is to estimate the true atmospheric state, whose evolution is modeled imperfectly by (2.1), we assume that the true state, denoted by \( \mathbf{w}_k' \), evolves instead according to

\[
\mathbf{w}_k = \mathbf{A}_{k-1} \mathbf{w}_{k-1} + \mathbf{b}_k' - 1,
\]  

(2.2a)

where the model error \( \mathbf{b}_k' \) is a random \( n \) vector that is assumed to be white in time, with mean zero and covariance matrix \( \mathbf{Q}_k \):

\[
\mathbb{E}\mathbf{b}_k' = 0,
\]  

(2.2b)

\[
\mathbb{E}(\mathbf{b}_k')(\mathbf{b}_l')^T = \mathbf{Q}_k \delta_{kl}.
\]  

(2.2c)

Here \( \mathbb{E} \) denotes the expectation, or ensemble-averaging operator, the superscript \( T \) denotes the transpose, and \( \delta_{kl} \) is the Kronecker delta, \( \delta_{kl} = 0 \) for \( k \neq l \) and \( \delta_{kk} = 1 \).

We assume further that meteorological observations \( \mathbf{w}_k^o \) are linear combinations of elements of the true state vector \( \mathbf{w}_k' \), contaminated by white noise:

\[
\mathbf{w}_k^o = \mathbf{H}_k \mathbf{w}_k' + \mathbf{b}_k^o.
\]  

(2.3a)

The length of the observation vector \( \mathbf{w}_k^o \) is the number of observations \( p \) available at time \( t_k \), so that \( p \) may depend on \( k \), \( p = p_k \). The \( p \times n \) observation matrix \( \mathbf{H}_k \) accounts for interpolation from model grid points to observation locations, and for any necessary conversion from state variables to observed variables, but is assumed not to depend on \( \mathbf{w}_k' \). The observation error \( \mathbf{b}_k^o \) is a random \( p \) vector, which is assumed to be white, with mean zero and covariance matrix \( \mathbf{R}_k \), and to be uncorrelated with the model error:

\[
\mathbb{E}\mathbf{b}_k^o = 0,
\]  

(2.3b)

\[
\mathbb{E}(\mathbf{b}_k^o)(\mathbf{b}_l^o)^T = \mathbf{R}_k \delta_{kl},
\]  

(2.3c)

\[
\mathbb{E}(\mathbf{b}_k^o)(\mathbf{b}_l')^T = 0.
\]  

(2.3d)

We will assume throughout that \( \mathbf{R}_k \) is nonsingular; i.e., that there are no perfect observations.

Under the stated assumptions, the Kalman, or Kalman–Bucy filter is the data assimilation system:

\[
\mathbf{w}_k' = \mathbf{A}_{k-1} \mathbf{w}_{k-1}',
\]  

(2.4a)

\[
\mathbf{P}_k' = \mathbf{A}_{k-1} \mathbf{P}_{k-1}' \mathbf{A}_{k-1}^T + \mathbf{Q}_{k-1},
\]  

(2.4b)

\[
\mathbf{K}_k = \mathbf{P}_k' \mathbf{H}_k' (\mathbf{H}_k \mathbf{P}_k' \mathbf{H}_k^T + \mathbf{R}_k)^{-1},
\]  

(2.4c)

\[
\mathbf{P}_k = (\mathbf{I} - \mathbf{K}_k \mathbf{H}_k) \mathbf{P}_k',
\]  

(2.4d)

\[
\mathbf{w}_k^o = \mathbf{w}_k' + \mathbf{K}_k (\mathbf{w}_k^o - \mathbf{H}_k \mathbf{w}_k'),
\]  

(2.4e)

for \( k = 1, 2, 3, \ldots ; \) cf., Gelb (1974), Ghil et al. (1981), or Jazwinski (1970). Equations (2.4a) and (2.4b) are the state forecast and estimation error covariance forecast, respectively. Equation (2.4c) gives the Kalman gain matrix \( \mathbf{K}_k \), which is the weight matrix applied to the observed-minus-forecast residual in (2.4e) to yield the analyzed field \( \mathbf{w}_k^o \). The matrix inverse in (2.4c) exists since we assumed \( \mathbf{R}_k \) to be nonsingular. Equation (2.4d) gives the matrix \( \mathbf{P}_k' \), which is needed at the next time step in (2.4b).

Because the dynamics and observations were assumed to be linear, one can show that the matrices \( \mathbf{P}_k' \) and \( \mathbf{P}_k^o \) in (2.4b)–(2.4d) are precisely the forecast and analysis error covariance matrices; i.e.,

\[
\mathbf{P}_k^{f,a} = \mathbb{E}(\mathbf{w}_k^{f,a} - \mathbf{w}_k') (\mathbf{w}_k^{f,a} - \mathbf{w}_k')^T.
\]  

(2.5a,b)

As a result, the Kalman filter represents a data assimilation scheme that is optimal in the sense of minimizing the analysis error variance and subsequent forecast error variance at every grid point, given all current and previous observational information. If in addition the model error \( \mathbf{b}_k' \) and observational error \( \mathbf{b}_k^o \) are Gaussian, then the Kalman filter is optimal in a much broader probabilistic sense (e.g., Jazwinski 1970, chapter 5).

In the nonlinear case, instead of (2.2a) one may assume that

\[
\mathbf{w}_k = \mathbf{a}_{k-1} (\mathbf{w}_{k-1}') + \mathbf{b}_{k-1}',
\]  

(2.6a)

where \( \mathbf{a}_k(\mathbf{w}) \) may depend nonlinearly on \( \mathbf{w} \), or even that
\[ w_k^f = a_{k-1}(w_{k-1}^f) + G_{k-1}(w_{k-1}^f) b_{k-1}^f, \]  
(2.6b)

for some nonlinear matrix function \( G_k = G_k(w) \). Further, instead of (2.3a), one may assume that

\[ w_k^o = h_k(w_{k}^f) + b_k^o, \]  
(2.7)

in which \( h_k(w) \) may depend nonlinearly on \( w \), as would be the case for satellite or ground-based radiometer measurements of temperature. The optimal (conditional mean and minimum variance) filter equations for the nonlinear case are known (cf., Jazwinski 1970, chapter 6), but their solution requires closure assumptions or other approximations, such as the Gaussian assumptions of Lorenz (1988a).

One of the simplest approximate nonlinear filters is the extended Kalman filter (EKF), which replaces (2.4a) and (2.4e) by

\[ w_k^f = a_{k-1}(w_{k-1}^f), \]  
(2.8a)

\[ w_k^o = w_k^f + K_k[w_k^o - h_k(w_k^f)], \]  
(2.8b)

and leaves (2.4b)–(2.4d) in place, defining

\[ A_k = \frac{\partial w_k(w)}{\partial w} \bigg|_{w=w_k^f}, \]  
(2.9a)

\[ H_k = \frac{\partial h_k(w)}{\partial w} \bigg|_{w=w_k^f}. \]  
(2.9b)

Many other approximate nonlinear filters are based on iterating the EKF in different ways.

Most of the statistical assumptions we have made, under which (2.4a)–(2.4e) is optimal for the linear case, have counterparts in the formulation of present-day optimal interpolation (OI) analysis schemes (e.g., Bergman 1979; Lorenz 1981). Many of these assumptions can be relaxed. For example, biases can be allowed in the model and observational errors [(2.2b), (2.3b)], correlations between these errors (2.3d) can be accounted for, and the errors need not even be white [(2.2c), (2.3c)]. Each of these generalizations is accounted for, provided of course that the relevant statistics are known, by additional terms in the corresponding optimal filter (Jazwinski 1970).

In fact the Kalman filter bears a close relationship with OI methods: the “analysis” stage (2.4c)–(2.4e) of the Kalman filter is merely a shorthand way of writing the usual OI analysis equations (cf., Cohn et al. 1981). The forecast error covariance matrix \( P_k^f \) appearing in (2.4c) and (2.4d), however, is replaced in the OI methodology by a fixed, readily computable approximation to \( P_k^f \) based on assumptions such as homogeneity, isotropy, and geostrophy. This approach can be thought of as one way of circumventing the considerable computational expense the Kalman filter incurs in actually evolving \( P_k^f \) according to (2.4b) and (2.4d). Additional computational savings are gained in OI schemes by employing local “data selection pro-
cedures,” which result in calculating in (2.4c) and using in (2.4d) and (2.4e) only selected elements of \( K_k \), in effect approximating \( K_k \) by a banded matrix of small bandwidth. Also, at most, only the diagonal elements of \( P_k^f \), that is, the analysis error variances, are calculated in OI schemes.

It is clear that the distinguishing feature of the Kalman filter is the presence of the forecast error covariance evolution equation (2.4b), which is coupled to the analysis error covariance equation (2.4d). The second term on the right side of (2.4b) is the model error covariance matrix \( Q_k \). This term accounts for the growth of forecast error covariance due to modeling error (2.2a), which consists of discretization and parameterization error. The first term governs the propagation of second moments of the forecast error. This propagation, which will be studied in some detail in appendix B, is especially important in light of the spatial variability of meteorological observing networks. For instance, this term gives the precise extent to which relatively large analysis error variances propagate from data-sparse regions into data-dense regions. The influence of variable data density on forecast error correlations is also given by this term.

Some ideas of both of these terms have been modeled, to some extent, in various OI implementations. For example, the prescribed forecast error variance growth rates of McPherson et al. (1979) essentially take the place of the main diagonal of \( Q_k \). The geographically dependent correlations employed by Baker et al. (1987) may take some account of the effect of variable data density on forecast error correlations.

As a practical consideration, it is worth emphasizing that the Kalman filter does require the covariance matrices \( Q_k \) and \( R_k \) as input, and that reasonably accurate knowledge of these matrices is important for filter performance (cf., Phillips 1982). The situation here, though, is simpler than that in OI, which instead requires specification of the complete forecast error covariance matrix \( P_k^f \) (as well as \( R_k \)): \( Q_k \) concerns only the model error, which is very basic information, whereas \( P_k^f \) contains accumulated and propagated contributions from all error sources. Still, little is known about the statistics of model errors, and it might ultimately be best to estimate them adaptively, during the assimilation process itself. Dee et al. (1985; a,b) have developed such an adaptive algorithm for estimating \( Q_k \) as well as \( R_k \), whose cost is the calculation of a number of additional Kalman filters. Recently a far less expensive approach to adaptive estimation has been proposed and successfully demonstrated for a simple model by Dee (1990a).

In addition to estimation of the required statistics, a matter of great practical concern is the enormous computational burden presented by the Kalman filter if implemented by brute force. For the current generation of numerical forecast models, the matrix \( P_k^f \) would be of size \( 10^6 \times 10^6 \) at least. Even if storage for
such a matrix were available, the calculation in (2.4b) would represent at least $10^6$ forecasts, since (2.4b) involves matrix–matrix operations whereas the forecast model (2.4a) itself is only a matrix–vector operation. Such brute force is not only infeasible, but it is probably not even warranted on the basis of cost/benefit considerations. Dee (1990a,b) shows that in fact much of the benefit of the Kalman filter may still be obtained when (2.4b) is replaced by a much simpler evolution equation.

Still there is intrinsic interest in observing how system (2.4) evolves, without approximation, for a sufficiently realistic weather model and observing network. That is the object of the present study. The two-dimensional model we use has 1200 variables, so even here the task is daunting. Our model matrix $A_k$ is independent of time, $A_k = A$, and it is strongly banded since we use an explicit finite-difference scheme with a compact, $3 \times 3$ stencil. The elements of $A$ are calculated only once and then stored by diagonals, saving both execution time and storage. Covariance matrices are also stored by diagonals, and the matrix–matrix operation indicated in (2.4b) is carried out by a diagonally matrix multiplication algorithm well suited to vector processors. The analysis equations (2.4c)–(2.4e) are implemented by a “serial observation processing algorithm” (Gelb 1974, p. 304) which, although it is algebraically equivalent to (2.4c)–(2.4e), requires no matrix inverses nor the solution of any linear algebraic equations when the observation error covariance matrix $R_k$ is diagonal.

These computational aspects of our Kalman filter implementation are described in detail by Parrish and Cohn (1985). The so-called “banded approximation” devised and implemented in that report is not made in the present study. We make no approximations whatsoever.

We do take the model error and observational error covariance matrices, $Q_k$ and $R_k$, to be known. Also we take the dynamics and observations to be linear, so that the filter is optimal and all its ingredients are known. Thus we have constructed an environment in which many questions regarding data assimilation can be answered in a precise way.

3. The numerical model and observing networks

a. The model equations

One of the simplest meaningful ways of formulating a Kalman-like filter for a nonlinear dynamical model is to start with a reference trajectory, such as the solution for prescribed complete initial data, then to obtain perturbation equations in the usual way by linearizing the model about this reference trajectory, and finally to employ the (linear) Kalman filter for the linearized dynamics (cf., Jazwinski 1970). Such a filter would be optimal if it happened that assumptions of the form (2.2) and (2.3) could describe the perturbed system. In practice it is usually better to use instead the EKF which, as discussed in the previous section, is based on continually updating the reference trajectory so that it coincides with the current estimate of the state (Jazwinski 1970).

To keep things simple we take the former approach, using a very simple reference trajectory, and we stick to the basic statistical assumptions (2.2) and (2.3). The nonlinear dynamics are the shallow-water equations for a midlatitude beta plane, and the reference trajectory is a constant zonal current $U$ that is in geostrophic balance with the reference geopotential $\Phi = \Phi(y)$. The linearized shallow-water system is then

\begin{align}
&u_t + U u_x + \phi_x - f v = 0, \\
v_t + U v_x + \phi_y + f u = 0, \\
&\phi_t + U \phi_x + \beta (u_y + v_x) + \phi_y v = 0.
\end{align}

The coordinates $x$ and $y$ point eastward and northward, respectively, $u$ and $v$ are the eastward and northward perturbation velocity components, and $\phi$ is the perturbation geopotential. We have

\begin{equation}
\frac{f}{U} + \phi_y = 0,
\end{equation}

and the Coriolis parameter is given by

\begin{equation}
f = f_0 + \beta y.
\end{equation}

The model domain is a rectangle from $x = 0$ to $x = 5000$ km and $y = 0$ to $y = 3000$ km, roughly circumscribing the continental United States (see Fig. 1), with periodic boundary conditions at the eastern and western boundaries and a rigid wall condition ($v = 0$) at the northern and southern boundaries. The circle of tangency of the beta plane is taken to be the center line ($y = y_1 = 1500$ km) of the domain, at latitude $\phi_1 = 37.5^\circ$, so that

\begin{equation}
\beta = \frac{2 \Omega}{a} \cos \phi_1 \approx 1.82 \times 10^{-11} \text{ s}^{-1} \text{ m}^{-1},
\end{equation}

and

\begin{equation}
f_0 = 2 \Omega \sin \phi_1 - \beta y_1 \approx 6.15 \times 10^{-5} \text{ s}^{-1}.
\end{equation}

Integrating (3.2) gives

\begin{equation}
\Phi = \Phi_0 - \left( f_0 y + \frac{1}{2} \beta y^2 \right) U.
\end{equation}

We take $\Phi_0 = 3 \times 10^4$ m$^2$ s$^{-2}$ and $U = 25$ m s$^{-1}$.

The total energy of the system, which is proportional to

\begin{equation}
E = \frac{1}{2} \int \int [\Phi(u^2 + v^2) + \phi^2] \, dx \, dy
\end{equation}

is conserved:

\begin{equation}
\frac{dE}{dt} = 0.
\end{equation}

There are no growing or decaying modes.
b. The discretization

Letting \( \mathbf{w} = (u, v, \phi)^T \), the shallow-water system (3.1) can be written as

\[
\frac{\partial}{\partial t} \mathbf{w} + \mathbf{L} \mathbf{w} = 0,
\]

where

\[
\mathbf{L} = \frac{\partial}{\partial x} \mathbf{A} + \frac{\partial}{\partial y} \mathbf{B} + \mathbf{C},
\]

and

\[
\mathbf{A} = \begin{pmatrix} U & 0 & 0 \\ 0 & U & 0 \\ \Phi & 0 & U \end{pmatrix},
\]

\[
\mathbf{B} = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & \Phi & 0 \end{pmatrix},
\]

\[
\mathbf{C} = \begin{pmatrix} 0 & -f & 0 \\ f & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}.
\]

We discretize (3.6) on a uniform grid of \( I \times J \) points and denote the elements of the discrete state vector by

\[
\mathbf{w}^k_{ij} \approx \mathbf{w}(i - 1) \Delta x, (j - 1) \Delta y, k \Delta t \]

for \( i = 1, 2, \ldots, I, j = 1, 2, \ldots, J, \) and \( k = 0, 1, 2, \ldots, \), where \( \Delta x = 5000 \text{ km}/I \) and \( \Delta y = 3000 \text{ km}/(J - 1) \). For all the experiments reported herein we take \( I = 25 \) and \( J = 16 \), so that \( \Delta x = \Delta y = 200 \text{ km} \) and there are \( 25 \times 16 \times 3 = 1200 \) state variables. Also we take nine time steps per hour, so \( \Delta t = 400 \text{ s} \), which is well within the CFL limit for our system.

For discretization we apply the Richtmyer two-step version of the Lax–Wendroff scheme (Richtmyer and Morton 1967), which is second-order accurate in time and space. This scheme has a compact \( 3 \times 3 \) stencil and uses only two time levels, thereby easing computer time and storage requirements, especially in the covariance evolution equation (2.4b). The first step calculates provisional values at the centers of grid boxes,

\[
\mathbf{w}^{k+1/2}_{i+1/2,j+1/2} = \mu_x \mu_y \mathbf{w}^k_{i+1/2,j+1/2} - \frac{1}{2} \mathbf{L}_{i+1/2,j+1/2} \mathbf{w}^k_{i+1/2,j+1/2},
\]

for \( i = 1, 2, \ldots, I \) and \( j = 1, 2, \ldots, J - 1 \), where the discrete operator \( \mathbf{L}_j \) is defined by

\[
\mathbf{L}_j = \lambda_x \mu_y \delta_x \mathbf{A}_j + \lambda_y \mu_x \delta_y \mathbf{B}_j + \Delta t \mu_x \mu_y \delta_x \mathbf{C}_j,
\]

with \( \lambda_x = \Delta t/\Delta x, \lambda_y = \Delta t/\Delta y \). The differencing and averaging operators \( \delta_x \) and \( \mu_x \) are defined by

\[
\delta_x \mathbf{w}_{ij} = \mathbf{w}_{i+1/2,j} - \mathbf{w}_{i-1/2,j},
\]

\[
\mu_x \mathbf{w}_{ij} = \frac{1}{2} (\mathbf{w}_{i+1/2,j} + \mathbf{w}_{i-1/2,j}),
\]

with analogous formulas for \( \delta_y \) and \( \mu_y \). The second step uses these provisional values to calculate

\[
\mathbf{w}^{k+1}_{ij} = \mathbf{w}^k_{ij} - \mathbf{L}_j \mathbf{w}^{k+1/2}_{ij}
\]

for \( i = 1, 2, \ldots, I \) and \( j = 2, 3, \ldots, J - 1 \); i.e., to
calculate the values at the new time step on the original grid, excluding the northern and southern boundaries.

To advance \( u_{ij}^k \) and \( \phi_{ij}^k \) at the boundaries \( j = 1 \) and \( j = J \), we apply the Lax–Friedrichs scheme to the first and third of Eqs. (3.6). The term \( (\Phi \psi) \), in the third equation is approximated by extending \( \Phi \psi \) across the boundaries as an odd function, which is a second-order accurate approximation since \( \Phi \psi = 0 \) at the boundaries. Thus, at the southern boundary \( j = 1 \) we have

\[
\begin{align*}
u_{ij}^{k+1} &= \frac{1}{2} \left( 2\mu_x^2 - 1 \right) u_{ij}^k + \frac{1}{2} \left( \lambda_x \mu_x \delta_x \left( U u_{ij}^k + \phi_{ij}^k \right) \right), \\
\phi_{ij}^{k+1} &= \frac{1}{2} \left( 2\mu_x^2 - 1 \right) \phi_{ij}^k + \frac{1}{2} \phi_{ij+1}^{k+1} - \lambda_x \mu_x \delta_x \left( U \phi_{ij}^k + \phi_{ij}^k \right) - \lambda_x \phi_{ij}^{k+1} \psi_{ij+1}^{k+1},
\end{align*}
\]

with similar formulas at the northern boundary \( j = J \). This boundary treatment is slightly different from the one described in Parrish and Cohn (1985).

Substituting (3.9) into (3.12) gives for the interior points, and (3.13) gives for the boundary points, explicit formulas for the elements of the state vector

\[
w_{k+1} = \{ w_{ij}^{k+1} ; i = 1, \cdots, I ; j = 1, \cdots, J \},
\]

(3.14)
in terms of the elements of \( w_k \) at neighboring grid points. The coefficients in these formulas are the elements of our \( 1200 \times 1200 \) dynamics matrix \( A \)—cf., Eq. (2.4a)—which are evaluated and stored by diagonals, once and for all time [cf., the closing paragraphs of section 2, and also Parrish and Cohn (1985), appendix C].

We note that \( A \) has 50 rows and columns of zeros, in accordance with the boundary condition (3.13b). Therefore, \( A \) has 50 null eigenvalues. These 50 rows and columns could be deleted from \( A \), along with 50 elements \( v_{11}, v_{12} \) from the state vector, but retaining them is more convenient for the diagonalwise storage and matrix multiplication scheme.

c. The observing network

Our observing network, depicted in Fig. 1, consists of 77 North American radiosonde stations enclosed by the computational domain, augmented by 31 Doppler wind profilers located in the central United States according to the site plan of Beran (1986). We also suppose radiometer observations of heights to be available at five of the profiler sites. The radiosondes observe \( \phi \), \( u \), and \( v \) every 12 h, the profilers observe \( u \) and \( v \) every hour, and the radiometers observe \( \phi \) every hour. The \( p_x \times n \) observation matrix \( H_k \) is determined through bilinear interpolation from the grid points to the observation locations.

We make the assumption that the observational errors are not correlated with each other, so the observation error covariance matrix \( R_k \) is diagonal. The diagonal elements are the squares (variances) of the standard errors for each type of observation. These are taken to be as follows:

\[
\sigma_u = 1.5 \text{ m s}^{-1} \quad \text{for profilers} \quad (3.15a)
\]

\[
\sigma_v = 2.8 \text{ m s}^{-1} \quad \text{for radiosondes} \quad (3.15b)
\]

\[
\sigma_a/g = 11 \text{ m} \quad \text{for radiosondes and for radiometers,} \quad (3.15c)
\]

where \( g = 9.8 \text{ m s}^{-2} \). The profiler errors are based on the observed performance of a network of profilers in Colorado (Collins 1987). The radiosonde errors are those currently assumed for the 500-mb level in the NMC Global Data Assimilation System (Deaven et al. 1988). The radiometer errors are based on the study of Westwater et al. (1984) and happen to be identical to the radiosonde height errors.

Having described the matrices \( A, H_k, \) and \( R_k \) for our experiments, it remains only to describe \( Q_k \) in the next section, and the initial conditions in the following one.

4. The model error covariance matrix

a. Motivation

In general the model error covariance matrix \( Q_k \) influences greatly the evolution of \( P_k \) and so, at least in principle, one should estimate \( Q_k \) directly from the data, perhaps adaptively, as discussed in section 2. The experiments reported in this paper use no real data, and therefore provide no real basis for estimating \( Q_k \). Instead, we attempt to determine what a “reasonable” choice of \( Q_k \) would be. Such an exercise is also intrinsically useful, because in so doing one can identify those parameters that are most relevant to a description of \( Q_k \), and direct estimation of \( Q_k \) could eventually proceed by estimating only those parameters.

Phillips (1986) hypothesized that “a data assimilation system with excellent forecast and analysis components and frequent access to good data should have first-guess errors similar to an ensemble of random slow modes with equipartition of energy.” As he explains further, by stipulating an assimilation system with excellent forecast and analysis components, the forecast error is to have a negligible first moment (i.e., expected value), while the stipulation that there be frequent access to good data means that the second moment (i.e., covariance matrix) of the forecast error is to be due primarily to errors in the forecast model itself, rather than in the observed data.

That first-order moments are to be negligible (in fact zero) is already tacit assumption in our Kalman filter framework, since (2.2b) implies with (2.1) and (2.2a) that

\[
E(w_i^f - w_k^f) = 0 \quad (4.1)
\]
for linear models $A$. The stipulation on second moments means, in our framework, that $Q_{k-1}$ is to be the dominant of the two terms on the right-hand side of the forecast error covariance evolution equation (2.4b). Thus, for the Kalman filter, Phillips’ hypothesis is that whenever $Q_{k-1}$ dominates in (2.4b), then $P_k$ should be similar to the covariance matrix of an ensemble of random slow modes with equipartition of energy. That is to say, $Q_{k-1}$ itself should be (similar to) the covariance matrix of such an ensemble. The extent to which $Q_{k-1}$ dominates in (2.4b) will be studied in our numerical experiments.

Phillips argued his hypothesis on physical and intuitive grounds and showed further that the covariance statistics of his hypothetical ensemble do agree in many respects remarkably well with those of the extensive diagnostic study due to Hollingsworth and Lönnberg (1986) and Lönnberg and Hollingsworth (1986). One exception was that the hypothetical ensemble yields too small a horizontal scale length for geopotential error correlations, and this discrepancy was alleviated by partitioning energy differently in the ensemble (Phillips 1986). Our model error covariance matrix $Q_k$ will therefore be specified according to Phillips’ hypothesis, i.e., in that it shall be the covariance matrix of an ensemble of random slow modes, but with a modified partitioning of energy. This partitioning will be described, along with implementation details, towards the end of this section. First we will describe the general structure of $Q_k$ resulting from Phillips’ hypothesis, and in so doing we shall supplement his reasoning by providing some theoretical justification for this structure.

b. Structure of $Q$

Rather than confining the discussion at this point to our specific dynamics matrix $A$ of section 3, let us take $A$ to be an arbitrary $n \times n$, time-independent matrix. We assume that $A$ has $n$ linearly independent eigenvectors, and therefore an eigendecomposition

$$A = \mathbf{V}\hat{\Lambda}\mathbf{V}^{-1},$$

where

$$\hat{\Lambda} = \text{diag}(\hat{\lambda}_1, \hat{\lambda}_2, \ldots, \hat{\lambda}_n)$$

is the diagonal matrix of eigenvalues $\hat{\lambda}_j$ of $A$, and where the columns $v_j$ of

$$\mathbf{V} = [v_1, v_2, \ldots, v_n]$$

are the corresponding eigenvectors, or normal modes, of $A$. The eigenvalues need not all be simple ($\hat{\lambda}_i \neq \hat{\lambda}_j$ if $i \neq j$), but if they are, then the decomposition (4.2) is guaranteed to exist.

The eigenvalues can be written (uniquely) in polar form

$$\hat{\lambda}_j = \rho_j \exp(i\nu_j \Delta t),$$

with $\nu_j$ and $\rho_j$ real. This $\nu_j$ is, by definition, the propagation frequency of mode $v_j$ for the dynamics (2.4a), while $\rho_j > 0$ is the amplification factor of mode $v_j$. Thus, mode $v_j$ propagates with frequency $\nu_j$ and grows, decays, or is neutrally stable according to whether $\rho_j > 1$, $\rho_j < 1$, or $\rho_j = 1$. Mode $v_j$ does not propagate if either $\nu_j = 0$ or $\rho_j = 0$. The eigenvectors (modes) themselves are uniquely defined, up to a normalization factor of course, with the exception of those belonging to any repeated eigenvalues. Such eigenvectors are not unique since, for example, if $\hat{\lambda}_1 = \hat{\lambda}_2$, then clearly any linear combination of $v_1$ and $v_2$ is also an eigenvector for eigenvalue $\hat{\lambda}_1$. For now we make no particular choice of the eigenvectors belonging to repeated eigenvalues, nor any particular choice of normalization.

Now, since the columns of $\mathbf{V}$ are the normal modes, the elements of the vector $\mathbf{b}_k$ defined by

$$\mathbf{b}_k = \mathbf{V}\mathbf{b}_k'$$

are the (generally complex) modal components of the model error $\mathbf{b}_k'$ defined in (2.2). The covariance matrix $\hat{Q}_k$ of $\mathbf{b}_k'$ is the positive semidefinite matrix defined by

$$\hat{Q}_k = E(\hat{b}_k')(\hat{b}_k')^*,$$

the asterisk denoting the complex conjugate transpose. In particular, the diagonal elements of $\hat{Q}_k$ are real and nonnegative. Premultiplying (4.7) by $\mathbf{V}$ and postmultiplying by $\mathbf{V}^*$, and using (4.6) and (2.2e), gives

$$Q_k = \mathbf{V}\hat{Q}_k\mathbf{V}^*$$

as the relationship between $Q_k$ and its modal transform $\hat{Q}_k$.

By random modes, Phillips (1986) means vectors whose modal components are uncorrelated; i.e., whose covariance matrix is diagonal in mode space. Applying Phillips’ hypothesis to the model error therefore gives us the following assumption.

1) ASSUMPTION 1—$Q_k$ HAS THE FORM (4.8), WITH $\hat{Q}_k$ DIAGONAL

We remark that the structure of $Q_k$ given by this assumption does not depend on the particular normalization employed to define the modes, but it does depend on how modes belonging to repeated eigenvalues are defined. To see this, simply observe that if $\mathbf{V}$ is one matrix of $n$ linearly independent eigenvectors, then every other such matrix must have the form

$$\mathbf{V} = \mathbf{V}\mathbf{D},$$

where $\mathbf{D}$ is nonsingular and $D_{ij} = 0$ if $\hat{\lambda}_i \neq \hat{\lambda}_j$; so $\mathbf{D}$ is necessarily diagonal if the eigenvalues $\hat{\lambda}_i$ are all simple. Substituting (4.9) into (4.8) gives a redefined $Q_k$,

$$Q_k = \mathbf{V}\hat{Q}_k\mathbf{V}^*,$$

where

$$\hat{Q}_k = \mathbf{D}\hat{Q}_k\mathbf{D}^*,$$
which under assumption 1 is still necessarily diagonal only if \( D \) is also.

Put another way, one may always take submatrices of \( Q_k \) corresponding to repeated eigenvalues to be diagonal, by appropriate choice (4.9) of eigenvector basis for those eigenvalues (at least if this basis can be taken to be independent of time, in particular if \( Q_k \) does not depend on time). For those submatrices, therefore, assumption 1 just amounts to making a specific choice of such basis. The dynamics matrix \( A \) of section 3 does turn out to have two repeated eigenvalues. Our choice of corresponding eigenvectors, as well as normalization, is described later in this section, with other implementation matters.

According to (2.4b)–(2.4d), the sequence \( P_k', k = 1, 2, 3, \ldots \), is forced by the sequence of model error covariance matrices \( Q_k \) and by the sequence of observations. The accumulation of both these inputs up to each time \( t_k \) is acted upon at time \( t_k \) by the dynamics matrix \( A \) in (2.4b). In appendix B it is shown that assumption 1 is essentially that the input due to the model error covariance matrices is not propagated by this action of the dynamics matrix. This is not a result, per se, of the usual decoupling one obtains in mode space. Rather, it is a result of the fact that, as shown in appendix B, the propagation frequency of the mode corresponding to the \((i, j)\)th element of \( Q_k \) is \( v_i - v_j \) [cf., (4.5)]: the diagonal elements of \( Q_k \) never propagate since \( v_i - v_j = 0 \) when \( i = j \). Each \( j \)th diagonal element can grow, remain constant, or dissipate, according to whether \( \rho_j > 1 \), \( \rho_j = 1 \), or \( \rho_j < 1 \), but cannot propagate.

It is also shown in appendix B that, for all “realistic” observing patterns, it is impossible for the input due to the observations not to propagate. Thus, the propagating part of \( P_k' \) always arises partly from the observations, and solely from them under assumption 1.

Put in another way, we have the following simple and intuitive grounds for justification of assumption 1: it is justified whenever most of the propagating part of \( P_k'\) is due to the observations, for in this case one may as well simply take \( Q_k \) to be nonpropagating. The extent to which it might be the case that most of the propagation is due to observations has never been examined in an operational setting, and to do so one would have to quantify just how much of the propagation is due to observations and how much is due to model error. The arguments in appendix B do indicate, however, that the amount of propagation due to observations increases as data density becomes more variable, and as the observation error variances in the mass and wind fields become less “commensurate” in a sense that will be defined there.

As a final remark concerning assumption 1, it is shown in appendix B that an assumption that \( Q_k \) be nonpropagating is actually slightly more general than assumption 1, and has the technical advantage of being independent of how modes belonging to repeated eigenvalues are defined.

Suppose now that we define a cutoff frequency \( \nu_c \), modes belonging to eigenvalues having \(|v_j| < \nu_c \) being called slow modes and the remainder being called fast modes. For a suitable ordering of the eigenvalues along the diagonal of \( A \), we may then partition \( V \) as

\[
V = [V_s, V_f].
\]

(4.12)

the columns of \( V_s \) and \( V_f \) being the slow and fast modes, respectively. Disregarding assumption 1 for the moment, suppose we partition \( Q_k \) accordingly, as

\[
\tilde{Q}_k = \begin{bmatrix}
\tilde{C}_k & \tilde{F}_k \\
\tilde{C}_k^* & \tilde{F}_k
\end{bmatrix},
\]

(4.13)

so that \( \tilde{C}_k \) and \( \tilde{F}_k \) are the covariance matrices of the slow and fast mode components of \( b_i \), respectively, while \( \tilde{C}_k \) is the cross-covariance matrix between them.

That Phillips’ ensemble is to consist of slow modes only means that \( \tilde{F}_k = 0 \) and \( \tilde{C}_k = 0 \), in which case it follows from (4.8), (4.12), and (4.13) that

\[
Q_k = \bar{V}_s \tilde{C}_k \bar{V}_s^*.
\]

(4.14)

Accordingly, we have assumption 2.

2) ASSUMPTION 2—\( Q_k \) HAS THE FORM (4.14), WITH \( \tilde{C}_k \) THE COVARIANCE MATRIX OF THE SLOW-MODE COMPONENTS OF THE MODEL ERROR \( b_i \).

An argument similar to that which follows the statement of assumption 1 shows that the structure imposed on \( Q_k \) by assumption 2 is invariant under both the particular normalization of the modes and the particular definition of modes belonging to repeated eigenvalues.

In appendix A we show that if assumption 2 is made, then the Kalman filter carries out initialization automatically at each analysis time (provided appropriate conditions on \( w_0 \) and \( P_0 \) are also met): the state estimate \( w_k'\) evolves slowly, no matter what observations are taken. Actually, this result is by no means new (cf., Petersen 1973), but it appears not to be widely known. Conversely, we point out that if \( w_k' \) is to evolve slowly, regardless of the observing pattern, then assumption 2 must be satisfied. Thus it is clear that assumption 2 is appropriate precisely to the extent that one considers a slowly evolving state estimate to be desirable.

We shall make both assumptions 1 and 2—which means that \( Q_k \) is to have the form (4.14)—with \( \tilde{C}_k \) diagonal and having real nonnegative diagonal elements. This reduces drastically the number of parameters defining \( Q_k \) from \( n(n + 1)/2 \) (since \( Q_k \) is a symmetric \( n \times n \) matrix) to about half the number of slow modes, or about \( n/6 \) (since \( Q_k \) is also real).

Finally, were we also to assume an equipartition of...
energy among the random slow mode components, the diagonal elements of $\mathbf{S}_k$ would then all be equal, if the modes are assumed to have been normalized by energy. Thus from (4.14) we would have

$$\mathbf{Q}_k = \alpha_k \mathbf{V}_s^* \mathbf{V}_s^*$$

(4.15)

$\alpha_k$ being a positive scalar, the only parameter remaining, depending possibly on time. In particular the model error correlations (in physical space) would be fixed, and the variances would be determined by the sole parameter $\alpha_k$. We shall not make this assumption for it yields unrealistically short correlation lengths, as noted by Phillips (1986) and as we shall also see. Instead $\mathbf{S}_k$ is chosen as described below.

c. Implementation

The diagonalization (4.2) of our $1200 \times 1200$ dynamics matrix $\mathbf{A}$ was carried out by performing, first, a discrete Fourier transform in the zonal direction—since the coefficients of the dynamics are independent of the zonal coordinate—followed by numerically solving a set of small ($48 \times 48$) sparse eigenproblems to obtain the meridional structure of each normal mode.

Examination of the resulting eigenvalues $\lambda_j$, revealed unambiguously two repeated eigenvalues: $\lambda_j = 1$ (for which $\rho_j = 1$ and $\nu_j = 0$) with multiplicity 13, and $\lambda_j = 0$ (for which $\rho_j = 0$ and $\nu_j$ is undefined) with multiplicity 50. The 13 modes with eigenvalue $\lambda_j = 1$ are stationary ($\nu_j = 0$) and are, therefore, trivially slow modes. They correspond to the geostrophic stationary solutions

$$u = -\frac{\phi_y}{f}, \quad v = 0, \quad \phi = \phi(y) \text{ arbitrary},$$

(4.16)

of the continuous dynamics (3.1). The 50 modes for $\lambda_j = 0$ are nullvectors of $\mathbf{A}$ corresponding to the 50 null rows and columns of $\mathbf{A}$ described in section 3. These modes were classified as fast modes, so that $\mathbf{Q}_k$ will have no component along them according to assumption 2. Equivalently, this means that we have required the (variance of the) $v$ component of the model error $b_k$ to vanish at the northern and southern boundaries: the boundaries are assumed to be deterministically rigid walls.

The remaining 1137 eigenvalues exhibited a clear separation, but not always by frequency. Discretization error complicated the frequency distribution for zonal wavenumbers from 8–12, but the ratio of divergence to vorticity was always a very strong indicator, with a large jump occurring between slow and fast modes. Thus there are 325 slow modes altogether, including the stationary ones, so that $\mathbf{S}_k$ is a $325 \times 325$ diagonal matrix.

We remark that all 1137 simple eigenvalues had $0 < \rho_j < 1$, reflecting the dissipative nature of the differ-
Most importantly, it can be shown that all eigenfunctions belonging to distinct eigenvalues are automatically energy-orthogonal; i.e.,

\[(w_i, w_j) = 0 \quad \text{when} \quad \omega_i \neq \omega_j, \tag{4.26}\]

where the inner product is defined by

\[(w_i, w_j) = \frac{1}{2} \int \int w_i^* M w_j dx dy, \tag{4.27}\]

and \(M\) was defined in (4.18). Those corresponding to repeated eigenvalues may also be taken to be energy-orthogonal:

\[(w_i, w_j) = 0 \quad \text{when} \quad \omega_i = \omega_j \quad \text{and} \quad i \neq j; \tag{4.28}\]

normalizing the eigenfunctions to have equal energy \(e\) then gives simply

\[(w_i, w_j) = \delta_{ij} e \quad \text{for all} \quad i, j. \tag{4.29}\]

Our normalization (4.17) is a discrete analog of (4.29) for \(i = j\), and our choice of orthogonalization (4.19) for the stationary modes corresponds to (4.28). It is not true that (4.19) holds automatically for all \(i \neq j\), at least not for the matrix \(N\) we have defined, although this would be true for some very carefully constructed difference schemes.

d. Choice of model error spectrum

We take the model error covariance matrix to be independent of time, \(Q_k = Q\). Since \(V_\tau\) is now a known 1200 \(\times\) 325 matrix, in light of (4.14) it remains only to describe the diagonal matrix \(S_{kk} = \hat{S}\).

For each slow mode \(v_j\), we define a planetary wavenumber \(n_j\) according to

\[n_j = \frac{2 \pi a}{L} \left[ m_j^2 + \left( \frac{z_j + 1}{2} \right)^2 \right]^{1/2}, \tag{4.30}\]

where \(a\) is the earth radius, \(L = 5000\) km is the length of the periodic channel, \(m_j\) is the zonal wavenumber on the periodic domain, and \(z_j\) is the number of zero-crossings in the meridional direction. We take the diagonal elements \(S_{jj}\) of \(\hat{S}\) to be functions of this planetary wavenumber alone,

\[S_{jj} = \gamma g(n_j), \tag{4.31}\]

\(g(n)\) being a prescribed function of \(n\). The function \(g(n)\) determines the shape of the model error correlations, and the parameter \(\gamma\) determines the model error variances. The value of \(\gamma\) was chosen to yield realistic forecast error variance growth rates for the radiosondes-only experiment, as described in the next section.

For \(g(n) = 1\) (equipartition), the resulting \(Q\) has \(hh\) and \(hv\) correlations as shown in Fig. 2. While the \(hv\) correlation (Fig. 2b) has the expected structure for geostrophically related variables, the horizontal length scale is very short compared to the model grid increment. This is due to the fact that the \(hh\) correlation of Fig. 2a is rather sharply peaked near the base point.

Bennett and Budgell (1987) argue that the tail of the spectrum \(g(n)\) is constrained by considerations of regularity of model solutions; their prescription is that \(g(n) \propto n^{-p}\) with \(p > 3\). We selected \(g(n)\), for all \(n\), in such a way that the \(hh\) correlation would be close to that used in the NMC regional analysis system (DiMego 1988). The resolution of that system is comparable to that of our model.

Thus, we set

\[g(n) = \left[ 1 + \frac{\Phi_0 n^2}{f_0^2 a^2} \right] e^{-\alpha n^2}. \tag{4.32}\]

The second factor here dominates for large \(n\), with the result that the Bennett and Budgell criterion is indeed satisfied. The first factor is included to approximately offset the energy normalization (4.17) of our modes. As a result, by taking a Fourier transform it follows that (4.32) leads to roughly Gaussian \(hh\) correlations in physical space,

\[\text{corr}(r) \approx \exp(-k_R n^2), \tag{4.33}\]

away from the boundaries at least, where \(r\) is the distance between the two points being correlated, and

\[\alpha = (4a^2 n_R)^{-1}. \tag{4.34}\]

For the regional analysis system at NMC, \(n_R\) lies in the range \(1 \times 10^{-6} < n_R < 4.1 \times 10^{-6} \text{ km}^{-2}\), depending on the vertical level (DiMego 1988). Since we are simulating flow at 500 mb, we used the midtroposphere value of \(n_R = 3 \times 10^{-6} \text{ km}^{-2}\).

The resulting \(hh\) and \(hv\) correlations are shown in Fig. 3. Comparing Fig. 3a with Fig. 2a, one sees that the \(hh\) correlation is now not as sharply peaked. In fact Fig. 3a compares quite well with the \(hh\) correlation used in the NMC regional analysis system (DiMego 1988). The \(hv\) correlation in Fig. 3b now has a reasonable length scale also.

5. Numerical experiments

Three experiments were carried out, each corresponding to a different observing pattern. Experiment A, in which only the simulated radiosonde observations described in section 3 and Fig. 1 are assimilated, was conducted to calibrate the growth parameter \(\gamma\) of Eq. (4.31), and to provide a benchmark against which to compare the results of the other two experiments. In experiment B, observations are assimilated from both the radiosonde and wind profiler networks. Experiment C incorporates simulated observations from the five ground-based radiometers, in addition to the radiosonde and profiler observations. The observation errors were given in (3.15).
In all three experiments we calculated only Eqs. (2.4b)–(2.4d) of the Kalman filter, as our focus is on the evolution of the covariance matrix $P_{k/z}$ rather than on that of the state estimate $w_{k/z}$. All three experiments were based on the choice of model error covariance matrix $Q$ described in section 4. The initial analysis error covariance matrix $P_0$ was set equal to a constant times $Q$, where the constant was determined so that the initial height variance averaged over the full domain was approximately equal to the asymptotic maximum of this quantity for experiment A, as shown in Fig. 4a. All experiments were run out to 2.5 days and used about 30 min of CPU time on the Cyber 205 at NMC, not including the time required to set up the matrix $Q$.

Figure 4 shows, as a function of the time $t_k$, the domain-averaged forecast/analysis error standard deviations in $h = \phi/g$ [panel (a)] and $u$ [panel (b)] for
experiment A; i.e., the square root of the average of elements along the diagonal of $P_d^{-1}$. Corresponding results for $v$ (not shown) are similar to those for $u$, but are somewhat smaller due to boundary effects. The periodic patterns seen in Fig. 4, due to observing pattern periodicity and model linearity, reflect error reduction at each observation time and subsequent growth due to model error. A periodic steady state is reached quickly, after the first observation time, although a very slight growing trend is visible in Fig. 4b. Existence of this trend implies that our model is not completely observable as defined in appendix A. The 12-h forecast error variance growth rates (differences between consecutive lower and upper extrema in Figs. 4a,b) of about $(9.3 \text{ m})^2$ for $h$ and $(2.5 \text{ m s}^{-1})^2$ for $u$, resulting from our chosen value of the parameter $\gamma$, are comparable to values prescribed in the NMC regional data assimilation system (DiMego 1988).

The average height analysis error standard deviation—about 6 m in Fig. 4a—is considerably smaller than the observation error standard deviation of 11 m given in (3.15c). In fact, if for the moment we denote by $\sigma_o^2$, $\sigma_f^2$, and $\sigma_a^2$ the observation, forecast (at analysis time), and analysis error standard deviations for height
at some location, then for the Kalman filter one would have

\[ \frac{1}{(\sigma^a)^2} = \frac{1}{(\sigma^f)^2} + \frac{1}{(\sigma^o)^2} \]  (5.1)

if the forecast errors (as well as observation errors) in all fields were uncorrelated in space. Taking the average value \( \sigma^f = 15.3 \text{ m} \) from Fig. 4a, and \( \sigma^o = 11 \text{ m} \), one finds that \( \sigma^a = 6 \text{ m} \) in Fig. 4a is much smaller than the value \( \sigma^o \approx 8.9 \text{ m} \) given by formula (5.1). This little calculation shows that forecast errors are substantially correlated in space (as will be studied directly later) and, more importantly, that the Kalman filter is able to take much advantage of this correlated information, especially in view of the fact that there are only 77 height observations, while all 400 grid points have been considered in the average.

On the other hand, the average \( u \) analysis error, shown in Fig. 4b to be about 3.75 m s\(^{-1}\), is larger than the observation error of 2.8 m s\(^{-1}\). Wind analysis errors near the radiosonde stations are smaller than the observation errors of course, but evidently the wind corrections are sufficiently local as to not survive averaging over the domain. Also, the local wind analysis errors (not shown) are not much smaller than the observation errors, because wind forecast errors tend to be considerably larger than the observation errors, as shown in Fig. 4b.

Figure 5 shows the results for experiment B corresponding to those of Fig. 4 for experiment A, now averaged over the three subregions depicted in Fig. 1, as well as over the full domain. Comparing Fig. 5a with Fig. 4a shows that the hourly profiler observations of wind alone also reduce height errors substantially. For example, the average height analysis error at synoptic times has been reduced from about 6 to less than 3 m over the profiler region. The largest height error reduction occurs over the profiler region of course, but a large reduction is also seen downstream of the profilers, and some reduction upstream is also noticeable.

The wind errors themselves (Fig. 5b) are quite small over the profiler region. Upstream they are comparable to those of the radiosondes-only experiment (Fig. 4b), while the errors downstream are clearly much smaller than the errors upstream. This demonstrates the downstream propagation of wind profiler information.

Downstream propagation is further visualized in Fig. 6, which shows for experiment B the spatial distribution of forecast error standard deviations in \( h \) [panel (a)] and \( u \) [panel (b)] after 60 h elapsed time. In both panels the contours are spaced less tightly downstream of the profilers than upstream, indicating downstream propagation. Also the large bullseye pattern of Fig. 6a shows quite graphically how beneficial the wind profiler data have been in reducing height field forecast errors.

We turn now to forecast error correlation maps, produced from \( P_{k'} \) at time \( t_k = 60 \text{ h} \). For the radiosondes-only experiment A, all height–height, height–wind, and wind–wind correlation functions, for various selected base points, turned out to be very similar to the corresponding correlation functions from \( Q \) (cf., Fig. 3), and therefore are not shown. In other words, the forecast error correlation structure for experiment A is dominated by that of the model error. In particular the radiosonde observations alone, being rather plentiful and fairly uniformly distributed over the whole domain, result in very little covariance propagation (cf., the discussion in appendix B).

The introduction of wind profiler data changes this situation radically. Figures 7a,b show for two different base points the \( hh \) forecast error correlation at 60 h for experiment B. These two correlation functions are quite different from one another after lateral translation, indicating that spatial homogeneity has been lost entirely. There is also some loss of isotropy. The correlation function with base point centered in the profiler region (Fig. 7b) is extremely broad; for example the 0.7 contour lies roughly where the 0.3 contour for \( Q \) lies (Fig. 3a).
Figures 8a,b show the corresponding $hv$ correlation functions for experiment B. Observe the marked asymmetry in Fig. 8a: the negative lobe to the west of the base point has a strong minimum of $-0.81$ (compared with $-0.56$ in Fig. 3b), while the positive lobe peaks at just $0.37$ (compared with $0.57$ in Fig. 3b).

Also, the correlation at zero separation is $-0.44$, instead of the value 0.0 that arises in existing OI schemes because of their assumptions of spatially constant height forecast error variances and geostrophic cross-correlations [cf., Bergman (1979)]. The $hv$ correlation function centered in the profiler region (Fig. 8b) looks much more like that for $Q$ (Fig. 3b), but with a weaker minimum ($-0.33$) and maximum ($0.31$).

Figure 9 shows the $vh$ correlation function corresponding to the $hv$ correlation function of Fig. 8a. Clearly the change is not simply the sign reversal that arises in the usual OI methodology (Bergman 1979). Actually Fig. 9 is comparable to Fig. 3b, after a sign change and translation, with the notable exception that again the point of symmetry in Fig. 9 is shifted away from the base point.

The results shown in Figs. 7–9 can be understood in the following way. First, it is shown in appendix A
that since our $\mathbf{Q}$ is based on slow modes alone, the results produced by the Kalman filter are exactly the same as what would be produced had the measuring instruments been capable of observing only the slow-mode components of the true state. We know that for our simple model the slow subspace coincides roughly with the space of geostrophically balanced fields. Thus the wind profiler observations are roughly equivalent to observations of the gradient of the height field: from the profiler data we have a rather massive amount of information concerning height gradients over the profiler region, but none concerning the (average) height itself. Thus for the height field, the forecast error in the few largest spatial scales must overwhelm that in the smaller scales. Fourier transform theory shows that such an error spectrum produces broad correlations, as seen in Fig. 7b.

The appearance of the height–wind and wind–wind correlations for experiment B, such as those shown in Figs. 8 and 9, can also be readily understood. Since we know $\mathbf{P}_{h'}$ consists of slow modes only, Eq. (A.21a), and again since the slow subspace is roughly geo-
strophic, it must be the case that the eight covariance functions from $P_k$ involving winds can be approximated by geostrophically differentiating the $hh$ covariance function from $P_k$—as much is done in the OI methodology (Bergman 1979, appendix B)—but taking into account the spatial variability of the height variances (Fig. 6a) as well as that of the correlations (Fig. 7) themselves. We tried this, using a fourth-order accurate numerical differentiation scheme. The results are not shown since they were nearly identical to Figs. 6b, 8, and 9, and to other cross correlations we have not shown. Incidentally, this supports the idea of using simplified dynamics to evolve covariances, perhaps even as simple as the dynamics suggested by Dee (1990b), at least under assumption 2 of section 4.

One might be surprised that the rather lopsided $uv$ correlation field of Fig. 8a, for example, is obtained geostrophically from the fairly isotropic $hh$ correlation field of Fig. 7a. The explanation here is that the asymmetry of Fig. 8a is due to the presence of a very steep gradient in the height error standard deviation field (Fig. 6a) near the base point. Cohn and Morone (1984)
and Morone and Cohn (1985) have studied the effect of such gradients upon geostrophically derived correlation functions. They showed in particular that such gradients tend to strengthen one lobe at the expense of weakening the other, while shifting the point of zero correlation away from the base point.

To test our claim that the broad $hh$ correlation function depicted in Fig. 7b is due to a relative paucity of baseline height information over the profiler region, we conducted experiment C in which height observations were made available every hour from ground-based radiometers collocated with just five of the profiler sites; experiment C is otherwise identical to experiment B. The resulting $hh$ forecast error correlation field after 60 h, for the base point located in the middle of the profiler region, is shown in Fig. 10. This field is indeed much tighter than the corresponding field for experiment B, shown in Fig. 7b. It is still somewhat broader than the $hh$ correlation field for $Q$ shown in Fig. 3a. The other correlation fields (not shown), for various base points, were also similar to those for $Q$.

The broadness of the $hh$ correlation function of Fig.
7b for experiment B leads one to expect the additional height observations of experiment C to have a fairly large impact on forecast/analysis height errors. This is borne out by Fig. 11, which for experiment C is analogous to Fig. 5a for experiment B. In Fig. 11 one sees that height forecast errors over the profiler region no longer increase from hour to hour as in Fig. 5a. At synoptic times, the height forecast error standard deviation averaged over the profiler region is now 5.5 m, whereas in Fig. 5a it was 7.7 m. Even averaged over the entire domain, this error has decreased from about 12 to about 11 m. Additional results (not shown) indicated, however, that the additional height information has almost no impact on wind errors.

The lack of homogeneity and isotropy witnessed in the correlation structures for experiment B does show up in the actual analysis weights (columns of the gain matrix $K_a$) themselves. Figure 12 shows for experiment
B at time \( t_k = 60 \) h the height correction at every point that would arise from an observed-minus-forecast residual of \( 1 \) m s\(^{-1}\) in \( v \) at the indicated radiosonde station (which happens to coincide with the western base point used in panel a of the correlation figures). The corresponding \( h \nu \) correlation function was shown in Fig. 8a. The maximum in Fig. 12 is 2.44 m and the minimum is \(-6.27\) m. The overall pattern is quite similar to that of Fig. 8a.

6. Summary and conclusions

We have implemented a Kalman filter algorithm for a linearized shallow-water model over the continental United States, and we have used it to assimilate simulated data from the existing radiosonde network, from a demonstration network of 31 wind profilers over the central United States, and from radiometers located at five of the profiler sites. We have adopted Phillips'
(1986) hypothesis, except for energy equipartitioning, to determine a reasonable choice for the model error covariance matrix. The model error energy spectrum was chosen in such a way as to yield, for the radiosonde-only experiment, forecast error correlations resembling those of operational models. Our main results can be summarized as follows:

(i) When only the radiosonde data are assimilated, the forecast error covariance matrix is so much more strongly forced by the model error covariance matrix than by the effect of data assimilation itself that forecast error correlations at synoptic times are nearly indistinguishable from model error correlations. This confirms experimentally Phillips' (1986) reasoning, briefly put, that over regions of relatively dense and uniform data coverage, the correlation structure of forecast error is determined mostly by that of the model error. This also justifies our application of Phillips' subsequent hypothesis directly to the model error covariance matrix itself, rather than to the forecast error covariance matrix.

(ii) Phillips' hypothesis that model error should consist of slow modes only is both necessary and sufficient for the Kalman filter to carry out initialization automatically, at each analysis time, for any observations having any error characteristics.

(iii) Phillips' hypothesis that the modal components of model error should be uncorrelated with one another amounts to an assumption that the effect of assimilating the data dominates the effect of model error in the propagation of forecast analysis error covariances. The theoretical results of appendix B show that data assimilation always gives rise to such propagation, unless there are sufficiently many wind and height observations, and unless certain relationships are satisfied between wind and height observation error variances.

(iv) Wind profiler observations can markedly reduce forecast analysis errors in heights as well as in winds. At the synoptic times during assimilation, the height forecast and analysis error standard deviations averaged over the profiler region were 7.7 and 2.5 m, respectively, while for the radiosonde-only experiment these figures were 15.3 and 6.0 m, averaged over the whole domain. For the zonal wind component, the corresponding figures were 2.7 and 1.7 m s\(^{-1}\) with the profilers, and 6.25 and 3.75 m s\(^{-1}\) without them.

(v) There is substantial downstream propagation of the profiler information, as predicted by theory, since the presence of the profilers juxtaposes a region of high wind data density with a surrounding area of relatively lower wind and height data density. Upstream the average height forecast and analysis error standard deviations at synoptic times were 12.7 and 5.5 m, respectively, but only 9.8 and 4.0 m downstream. For winds they were 5.8 and 3.5 m s\(^{-1}\) upstream, but only 4.5 and 2.8 m s\(^{-1}\) downstream.

(vi) Introduction of wind profiler data also leads to marked inhomogeneity and anisotropy of the height-height forecast error correlations; that is, to significant forecast error correlations in mode space, as expected from theory. In particular these correlations become extremely broad over the profiler region, because while the plentiful and accurate wind data determine height gradients well, they cannot determine the baseline height. The resulting height-wind and wind-wind cross correlations can be explained by geostrophic differentiation of the height-height covariance field. The effects of correlation function inhomogeneity and anisotropy are also visible in the analysis weights.

(vii) Introduction of height data from radiometers collocated with just five profiler stations tightens the height-height correlations considerably. The additional height data have little impact on wind forecast errors, but they reduce height forecast errors over the profiler region, from 7.7 to 5.5 m.

Our experimental results, particularly (6) above, indicate some potentially serious limitations of the current OI methodology, especially for assimilating profiler data: OI schemes do not mimic the behavior of forecast error correlation functions seen here. In general, therefore, we expect the Kalman filter to yield substantially more accurate analyses and forecasts than OI methods, as previously demonstrated in direct, quantitative comparisons between OI and the Kalman filter (Cohn et al. 1981; Dee 1990b). The challenge, of course, is still to find an approximation to the Kalman filter that would achieve most of the filter's benefits at a fraction of the cost, especially in a nonlinear setting. We hope that our results help spur progress toward this so-far elusive goal.

Acknowledgments. We wish to thank R. McPherson and M. Ghil, who encouraged us to pursue this work, and J. Derber, P. Long, and two anonymous reviewers for their very thoughtful reviews of the manuscript. This paper was completed while one of us (S.E.C.) visited NMC as an UCAR Senior Fellow. He is grateful to E. Kalnay, G. W. Curtis and M. Austin, and to the staff and steering committee of the UCAR Visiting Scientist Program for helping make this possible. S.E.C. also acknowledges the generous support of NASA, which funded this effort through Grant NAG-5-820, and of NOAA, through Grant NA84AA-D-00018.

APPENDIX A

Model Error and Initialization

The Kalman filter state forecast \( \mathbf{w}_k^f \), given by (2.4a), will be said to evolve slowly if for all time it is a linear combination of slow modes; i.e., if it has the form

\[
\mathbf{w}_k^f = \mathbf{V}_s \hat{\mathbf{w}}_{sk}, \quad k = 1, 2, 3, \ldots,
\]

(A.1)

\( \mathbf{V}_s \) being the matrix of slow modes defined in (4.12) and \( \hat{\mathbf{w}}_{sk} \) then being a vector of slow mode components.
We want to show that \( \mathbf{w}_k^f \) evolves slowly, independently of the manner in which the true state \( \mathbf{w}_k^o \) is observed, if and only if \( \mathbf{w}_0^a, \mathbf{P}_0^a, \) and \( \mathbf{Q}_k \) have the forms

\[
\mathbf{w}_0^a = \mathbf{V}_s \hat{\mathbf{w}}_{0s}, \quad \mathbf{P}_0^a = \mathbf{V}_s \hat{\mathbf{P}}_{ss} \mathbf{V}_s^*, \quad \mathbf{Q}_k = \mathbf{V}_s \hat{\mathbf{Q}}_k \mathbf{V}_s^*, \quad k = 0, 1, 2, \cdots. \tag{A.2a} \tag{A.2b} \tag{A.2c}
\]

Equations (A.2a) and (A.2b) state simply that the initial analysis and its covariance matrix are initialized. Equation (A.2c) is assumption 2 of section 4, and it is the condition that makes slow evolution of \( \mathbf{w}_k^f \) independent of how the true state is observed.

If one considers it desirable for \( \mathbf{w}_k^f \) to evolve slowly, then one wants this slow evolution to occur independently of how the true state is observed. To amplify this statement we will show also that, while for certain observing patterns slow evolution can occur without (A.2c) being satisfied, such patterns must be specially contrived and generally will lead to growth, possibly unbounded, of the forecast error covariance matrix \( \mathbf{P}_k^f \). This makes it clear that if one accepts both the desirability of slow evolution and the optimality of the Kalman filter, then assumption 2 of section 4 is a logical necessity.

Finally, we will indicate some computational simplifications that occur in the Kalman filter if (A.2a)–(A.2c) all hold.

To begin, and also for later use in appendix B, we transform the Kalman filter equations (2.4a)–(2.4e) to mode space. Thus, given the decomposition (4.2) of the dynamics matrix \( \mathbf{A} \), define the modal component vectors \( \hat{\mathbf{w}}_k^f \) and \( \hat{\mathbf{w}}_k^a \) according to

\[
\mathbf{w}_k^{f,a} = \mathbf{V} \hat{\mathbf{w}}_k^{f,a}, \quad \mathbf{P}_k^{f,a} = \mathbf{V} \hat{\mathbf{P}}_k^{f,a} \mathbf{V}^*, \quad \mathbf{Q}_k = \mathbf{V} \hat{\mathbf{Q}}_k \mathbf{V}^*, \tag{A.3a} \tag{A.3b} \tag{A.3c} \tag{A.3d} \tag{A.3e}
\]

the mode-space covariance matrices \( \hat{\mathbf{P}}_k^f, \hat{\mathbf{P}}_k^a, \) and \( \hat{\mathbf{Q}}_k \) by

\[
\hat{\mathbf{P}}_k^{f,a} = \mathbf{V} \hat{\mathbf{P}}_k^{f,a} \mathbf{V}^*, \quad \hat{\mathbf{Q}}_k = \mathbf{V} \hat{\mathbf{Q}}_k \mathbf{V}^*, \tag{A.4a} \tag{A.4b} \tag{A.4c} \tag{A.4d}
\]

and a fast part

\[
\hat{\mathbf{P}}_k^{f,a} = (I - \hat{\mathbf{K}}_k \hat{\mathbf{H}}_k) \hat{\mathbf{P}}_k^{f,a}. \tag{A.4e}
\]

These are identical in form to (2.4a)–(2.4e). All quantities now wear hats, save \( \mathbf{w}_k^o \) and \( \mathbf{R}_k \), and the transposes have become conjugate transposes.

Now partition \( \hat{\mathbf{A}} \) according to

\[
\hat{\mathbf{A}} = \begin{bmatrix} \hat{\mathbf{A}}_s & 0 \\ 0 & \hat{\mathbf{A}}_f \end{bmatrix}; \tag{A.5a}
\]

\( \hat{\mathbf{A}}_s \) and \( \hat{\mathbf{A}}_f \) are diagonal matrices of "slow" and "fast" eigenvalues, respectively, the corresponding partitioning of \( \mathbf{V} \) having been given by (4.12). Analogously to the partitioning (4.13) of \( \hat{\mathbf{Q}}_k \), partition \( \hat{\mathbf{P}}_k^f, \hat{\mathbf{P}}_k^a \) as

\[
\hat{\mathbf{P}}_k^{f,a} = \begin{bmatrix} \hat{\mathbf{P}}_k^{f,s} & \hat{\mathbf{P}}_k^{f,a} \\ (\hat{\mathbf{P}}_k^{a,s})^* & \hat{\mathbf{P}}_k^{a,a} \end{bmatrix}. \tag{A.5b} \tag{A.5c}
\]

Also, introduce the partitionings

\[
\hat{\mathbf{w}}_k^{f,a} = \begin{bmatrix} \hat{\mathbf{w}}_k^{f,s} \\ \hat{\mathbf{w}}_k^{a,a} \end{bmatrix}, \quad \hat{\mathbf{K}}_k = \begin{bmatrix} \hat{\mathbf{K}}_k^{s} \\ \hat{\mathbf{K}}_k^{a,a} \end{bmatrix}, \quad \hat{\mathbf{H}}_k = [\hat{\mathbf{H}}_k^{s} \hat{\mathbf{H}}_k^{a,a}]. \tag{A.5d} \tag{A.5e} \tag{A.5f} \tag{A.5g}
\]

Compared with (A.1) and (A.2), it is clear that what we are to do is to show that

\[
\dot{\hat{\mathbf{w}}}_k^{f,s} = 0, \quad k = 1, 2, 3, \cdots, \tag{A.6a}
\]

for all possible observing patterns \( \mathbf{H}_k \), if and only if all of the following conditions are satisfied:

\[
\dot{\hat{\mathbf{w}}}_0^{f,s} = 0, \quad \dot{\hat{\mathbf{P}}}_0^{f,s} = 0, \quad \dot{\hat{\mathbf{K}}}_k = 0, \quad \hat{\mathbf{C}}_k = 0, \quad k = 0, 1, 2, \cdots. \tag{A.7a} \tag{A.7b} \tag{A.7c} \tag{A.7d} \tag{A.7e}
\]

For convenience we will make one simplifying assumption, namely, that the diagonal matrix \( \hat{\mathbf{A}}_f \) of (A.5a) will be taken to be nonsingular. If \( \hat{\mathbf{A}}_f \) is singular, conditions (A.7) are still sufficient to guarantee (A.6) for all \( \mathbf{H}_k \), and a slight generalization of (A.7) becomes the necessary condition.

Thus, for example, the partitionings allow the state forecast (A.4a) and analysis update (A.4e) to be separated into a slow part

\[
\dot{\hat{\mathbf{w}}}_k^{f,s} = \hat{\mathbf{A}}_s \hat{\mathbf{w}}_{k-1}^{f,s}, \quad \dot{\hat{\mathbf{P}}}_k^{f,s} = \hat{\mathbf{A}}_s \hat{\mathbf{P}}_k^{f,s} \hat{\mathbf{A}}_s^* + \hat{\mathbf{Q}}_{k-1} \tag{A.8a} \tag{A.8b}
\]

and a fast part

\[
\dot{\hat{\mathbf{w}}}_k^{f,a} = \hat{\mathbf{A}}_f \hat{\mathbf{w}}_{k-1}^{f,a}, \quad \dot{\hat{\mathbf{P}}}_k^{f,a} = (I - \hat{\mathbf{K}}_k \hat{\mathbf{H}}_k) \dot{\hat{\mathbf{P}}}_k^{f,a}. \tag{A.9a} \tag{A.9b}
\]
Examination of (A.9) shows that (A.6) implies both (A.7a) and
\[ K_{jk}(w_{k}\hat{w}_{k}^* - \hat{H}_{k}\hat{w}_{k}^*) = 0, \quad k = 1, 2, 3, \cdots, \]  
(A.10)
since \( \hat{A}_{k} \) is nonsingular, while (A.7a) and (A.10) together imply (A.6) for any \( \hat{A}_{k} \).

Equation (A.10) states that each analysis increment is initialized; (A.10) is also equivalent to
\[ \hat{K}_{jk} = 0, \quad k = 1, 2, 3, \cdots, \]  
(A.11)
(with probability one) since the observed-minus-forecast residual is stochastic. Thus, we have shown that (A.6) is equivalent to (A.7a) and (A.11) together. It remains only to show, therefore, that (A.11) is equivalent to (A.7b)–(A.7e).

To do so, first substitute the partitioned matrices into (A.4b)–(A.4d), which gives for \( k = 1, 2, 3, \cdots \)
the equations
\[ \hat{P}_{sk} = \hat{A}_{k}\hat{P}_{sk-1}\hat{A}_{k}^* + \hat{S}_{k-1} \]  
(A.12a)
\[ \hat{P}_{fk} = \hat{A}_{k}\hat{P}_{fk-1}\hat{A}_{k}^* + \hat{F}_{k-1} \]  
(A.12b)
\[ \hat{P}_{ck} = \hat{A}_{k}\hat{P}_{ck-1}\hat{A}_{k}^* + \hat{C}_{k-1} \]  
(A.12c)
\[ M_{k} = \hat{H}_{k}\hat{P}_{k}^*\hat{H}_{k}^* + R_{k} \]  
(A.12d)
\[ K_{sk} = (\hat{H}_{s}\hat{P}_{fk} + \hat{H}_{f}\hat{P}_{ck-1})M_{k}^{-1} \]  
(A.12e)
\[ K_{fk} = (\hat{H}_{s}\hat{P}_{fk} + \hat{H}_{f}\hat{P}_{ck-1})M_{k}^{-1} \]  
(A.12f)
\[ \hat{P}_{sk} = \hat{P}_{sk-1} - K_{sk}M_{k}K_{sk}^* \]  
(A.12g)
\[ \hat{P}_{fk} = \hat{P}_{fk-1} - K_{fk}M_{k}\hat{P}_{fk} \]  
(A.12h)
\[ \hat{P}_{ck} = \hat{P}_{ck-1} - K_{ck}M_{k}\hat{P}_{ck} \]  
(A.12i)

In view of (A.12h) and (A.12i), we see that (A.11) implies
\[ \hat{P}_{fk}^* = \hat{P}_{sk}, \quad \hat{P}_{ck}^* = \hat{P}_{ck}, \quad k = 1, 2, 3, \cdots; \]  
(A.13a), (A.13b)
i.e., that only the slow part \( \hat{P}_{sk}^* \) of the forecast error covariance matrix is altered at each analysis time. On the other hand, if (A.13a) holds, then from (A.12h) it follows that (A.11) is satisfied, since the matrix \( M_{k} \) is positive definite, and \( R_{k} \) being positive definite. Thus (A.11) is equivalent to (A.13a) and (A.13b). Hence, by substituting (A.13a) and (A.13b) into (A.12b) and (A.12c), one finds that (A.11) is also equivalent to
\[ \hat{A}_{f}\hat{P}_{fk-1}\hat{A}_{f}^* + \hat{F}_{k-1}, \quad k = 1, 2, 3, \cdots, \]  
(A.14a)
\[ \hat{A}_{f}\hat{P}_{ck-1}\hat{A}_{f}^* + \hat{C}_{k-1}, \quad k = 1, 2, 3, \cdots, \]  
(A.14b)

Finally, from (A.12f) it is seen that (A.11) is equivalent to
\[ \hat{H}_{k}\hat{P}_{ck} + \hat{H}_{f}\hat{P}_{fk} = 0, \quad k = 1, 2, 3, \cdots, \]  
(A.15)
since the matrix \( M_{k}^{-1} \) is nonsingular. This can be true for all observing patterns \( H_{k} \) if and only if in fact
\[ \hat{P}_{fk} = 0, \quad \hat{P}_{ck} = 0, \quad k = 1, 2, 3, \cdots, \]  
(A.16a), (A.16b)
or, equivalently in light of (A.13),
\[ \hat{P}_{fk}^* = 0, \quad \hat{P}_{ck}^* = 0, \quad k = 1, 2, 3, \cdots. \]  
(A.17a), (A.17b)

Actually, (A.17b) is redundant because it is implied by (A.17a), as \( \hat{P}_{k} \) is a covariance matrix and is, therefore, positive semidefinite. Thus, (A.17b) holds for all \( H_{k} \) if and only if (A.17a) holds. It is clear from (A.14a), though, that (A.17a) is equivalent to (A.7b) and (A.7d), using again the assumption that \( \hat{A}_{k} \) is nonsingular. Conditions (A.7c) and (A.7e) are redundant, again because \( P_{0}^* \) and \( Q_{k} \) are covariance matrices. Therefore, (A.11) holds for all \( H_{k} \) if and only if (A.7b)–(A.7e) hold, and this completes our argument.

That (A.2a)–(A.2c) imply slow evolution is intuitively clear: (A.2b) and (A.2c) imply that at no time does the forecast error have a fast component to be corrected by the observed data, so naturally the Kalman filter yields analysis increments devoid of fast components, and slow evolution for all time then follows from (A.2a). Conversely, if the evolution is slow, then (A.2a) must hold and the observations must contain no information concerning fast-mode forecast errors; i.e., (A.15) must hold. If no observations can contain such information then it must be the case that there are no fast-mode forecast errors; i.e., (A.15) implies (A.2b) and (A.2c).

In particular it is possible to obtain slow evolution without satisfying (A.2b) and (A.2c), as long as (A.15) holds. Such a situation is not at all desirable though, for as we have shown, slow evolution also implies (A.14). Fast-mode components of the forecast error, even if present, will not get corrected if (A.15) holds; the reason being that (A.15) states that no fast components could be discerned by the data. For instance, if \( P_{f0}^* = 0 \) but \( F_{k} \) is nonzero and time independent, then \( \hat{P}_{fk} \) grows indefinitely in time according to (A.14a). This growth could even be unbounded if \( \hat{A}_{f} \) has any entries with modulus greater than or equal to one; i.e., if there are any neutrally stable or unstable fast modes, although unbounded growth would obviously be an artifact of the linear theory. Still, suffice it to say that if the forecast error does have a fast component, then one certainly would like to have observations capable of correcting such a component; such a correction cannot be made if (A.15) holds.

Satisfying (A.15) without satisfying (A.2b) and (A.2c) also imposes generally an unrealistic constraint on the observing pattern \( H_{k} \). Suppose for example that
(A.7a)–(A.7c) and (A.7e) all hold, but that $\hat{F}_k$ is positive definite; the model error has a nonzero component along each fast mode, but these are not correlated with any slow-mode components. Then from (A.13) and (A.14) it follows that each $\hat{P}_{ik}^{f,a}$ is zero and each $P_{ik}$ is positive definite, hence invertible, so that (A.15) becomes simply $H_k = 0$. That is, by use of (A.3f), (A.5g) and (4.12), we have

$$H_k V_f = 0,$$  \hspace{1cm} (A.18)

which is to say that every fast mode is a nullvector of $H_k$: the condition (A.15) for no fast-mode components of the forecast error to be discernible by the data in this case is simply that no fast modes can be observed directly.

Condition (A.18) is probably not physically realizable with local observations of winds and heights, although it might conceivably be the case that $H_k V_f = 0$ for some of the fast modes $V_f$. Modes $f$ for which $H_k V_f = 0$ are said to be unobservable. The condition

$$H_k V_f \neq 0$$  \hspace{1cm} (A.19)

for all eigenvectors $V_f$ of the dynamics matrix $A$ guarantees complete observability of the dynamical system, and leads to asymptotic stability of the Kalman filter; i.e., to impossibility of the aforementioned growth of $P_k^{f,a}$ [cf., Kalman and Bucy (1961); Kailath (1980); and Cohn and Dee (1988)].

When (A.2a)–(A.2c) are all satisfied, there results a considerable simplification of the Kalman filter, at least in mode space, since we have shown that (A.2a)–(A.2c) imply that $w_k^{f,a}$ and $P_k^{f,a}$ have the forms

$$w_k^{f,a} = V_f \hat{w}_k^{f,a}, \quad (A.20a), \quad (A.20b)$$

$$P_k^{f,a} = V_f \hat{P}_k^{f,a} V_f^*, \quad (A.21a), \quad (A.21b)$$

respectively, for each $k = 1, 2, 3, \ldots$; only these slow-mode components, $\hat{w}_k^{f,a}$ and $\hat{P}_k^{f,a}$, which have much smaller dimension than $w_k^{f,a}$ and $P_k^{f,a}$, need be calculated. According to (A.8) and (A.12), these are given recursively by

$$\dot{\hat{w}}_k^{f,a} = A * \hat{w}_k^{f,a} + B * \hat{Y}_k^{f,a} + P_k^{f,a} \hat{V}_f$$  \hspace{1cm} (A.22a)

$$\dot{\hat{P}}_{fjk}^{f,a} = A * \hat{P}_{fjk}^{f,a} + A * \hat{P}_{fjk}^{f,a} A^* + \hat{Q}_{fjk}$$  \hspace{1cm} (A.22b)

$$\dot{K}_k = \hat{K}_k (H_k H_k^* + \hat{R}_k)^{-1}$$  \hspace{1cm} (A.22c)

$$\dot{\hat{P}}_k^{f,a} = (I - K_k H_k) \hat{P}_k$$  \hspace{1cm} (A.22d)

$$\dot{\hat{w}}_k^{f,a} = \hat{K}_k \hat{w}_k^{f,a}$$  \hspace{1cm} (A.22e)

In (A.22c) we have used the fact that, according to (A.21a) and (A.5b) and (A.5g), the matrix $M_k$ of (A.12d) can be written as

$$M_k = H_k H_k^* \hat{R}_k + \hat{R}_k$$  \hspace{1cm} (A.23)

a similar simplification has been used in (A.22e). The matrix $H_k$ appears nowhere in (A.22a)–(A.22c) even though $H_k \neq 0$; the Kalman filter behaves exactly as if no fast modes are observable, because in fact there are none to observe. The idea here can be thought of as an example of the idea of “reduction of the control variable” discussed by Le Dimet and Talagrand (1986).

The modal Kalman filter equations (A.22) present a potentially large computational advantage over the physical-space filter (2.4). In addition to the lower dimensionality already noted, the dynamics matrix $A*$ in (A.22b) is diagonal and, under assumption 1 of section 4, the matrix $S_k$ would also be diagonal. Equations (A.22) also have the same form as the physical-space filter (2.4) so that, for example, the update equations (A.22c)–(A.22e) can still be implemented by processing observations serially (cf., Parrish and Cohn 1985 appendix B), which eliminates the need for solving directly the linear system represented by (A.22c). The additional overhead is that of calculating the eigendecomposition (4.2) and the transformed observing matrix $H_k = H_k V_f$. The formulation and computational efficiency of a slow-mode filter like (A.22) for fully nonlinear dynamics remain to be seen.

**APPENDIX B**

Propagation and Nonpropagation of Covariance Matrices

**a. Motivation**

The forecast/analysis error covariance matrix $P_k^{f,a}$ can be said to propagate according to the dynamics (2.4b) in much the same way that the estimated state $w_k^{f,a}$ can be said to propagate according to the dynamics (2.4a). Using expression (4.5) for the elements of the diagonal matrix $A*$, and the fact that the elements of $A*$ are just the complex conjugates of those of $A$, the mode-space counterparts (A.4a) and (A.4b) of (2.4a) and (2.4b) may be written componentwise as

$$\dot{w}_k^{f,a} = \rho \exp (\delta t) \hat{w}_k^{f,a}$$  \hspace{1cm} (B.1a)

$$\dot{P}_{fjk}^{f,a} = \rho \exp (\delta t) \hat{P}_{fjk}^{f,a} + \hat{Q}_{fjk}$$  \hspace{1cm} (B.1b)

where $\hat{w}_k^{f,a}$ denotes the $j$th element of $\hat{w}_k^{f,a}$, and $\hat{P}_{fjk}^{f,a}$ and $\hat{Q}_{fjk}$ denote the $(i,j)$th elements of $\hat{P}_{fjk}^{f,a}$ and $\hat{Q}_k$, respectively. The modal decompositions (A.3a)–(A.3d) themselves may be written as

$$w_k^{f,a} = \sum_{j=1}^{n} \hat{w}_k^{f,a} V_j$$  \hspace{1cm} (B.2a)

$$P_k^{f,a} = \sum_{i=1}^{n} \sum_{j=1}^{n} \hat{P}_{fjk}^{f,a} V_i V_j^*$$  \hspace{1cm} (B.2b)

$v_j$ being the $j$th mode (cf. (4.4)]. Equations (B.1a) and (B.2a) show that mode $v_j$ of the state estimate propagates with frequency $\nu_j$, while (B.1b) and (B.2b) show that the "mode," or outer product, $v_j v_j^*$, of the esti-
mation error covariance matrix propagates with frequency \( \nu_i - \nu_j \).

It is clear from (B.1a) that neither the stationary modes (those for which \( \nu_i = 0 \)) nor the nullvectors of \( A \) (those for which \( \rho_j = 0 \)), if any, can be said to propagate. All other modes do propagate. Thus, the condition for \( \mathbf{w}^a \) not to propagate (to time \( t_{k+1} \)) can be stated as:

\[
\text{either } \nu^a_{jk} = 0, \quad \text{or } \nu_j = 0, \quad \text{or } \rho_j = 0, \quad \text{or } \rho_i = 0, \quad \text{or } \nu_i = \nu_j,
\]

(B.3a)-(B.3c)

for each \( j = 1, 2, \cdots, n \). This would be unusual, of course, and also not very interesting.

In a sense it is much easier for \( \mathbf{P}^a_k \) not to propagate, because the propagation frequency in (B.1b) is the difference \( \nu_i - \nu_j \). In particular, \( \mathbf{P}^a_k \) does not propagate if \( \mathbf{P}^a_k \) is diagonal; i.e., if the analysis errors are uncorrelated in mode space. More generally, it is clear that the condition for \( \mathbf{P}^a_k \) not to propagate (to time \( t_{k+1} \)) is that, for each pair of indices \( i, j \) with \( i \neq j \), either

\[
\hat{\mathbf{P}}^a_{ijk} = 0, \quad \text{or } \nu_i = \nu_j, \quad \text{or } \rho_i = 0, \quad \text{or } \rho_j = 0; \quad \text{(B.5a)-(B.5c)}
\]

the diagonal elements \( \hat{\mathbf{P}}^a_{iik} \) are unrestricted, as are the off-diagonal elements \( \hat{\mathbf{P}}^a_{ijk} \) for which any of (B.5a)-(B.5c) hold. If \( \mathbf{P}^a_k \) does not propagate, then the individual nonzero modal components \( \hat{\mathbf{P}}^a_{ijk} \) can still grow, remain constant, or dissipate, accordingly as \( \rho_i \rho_j > 1 \), \( \rho_i \rho_j < 1 \), and can be augmented by \( \hat{Q}_{ijk} \) [cf., (B.1b)], but their phases cannot change.

We have introduced the notion of lack of propagation for covariance matrices primarily because it gives additional intuitive appeal to the notion of lack of mode-space correlation. For this reason, in this Appendix we wish to draw attention to the similarity of these two notions, rather than to the difference between them. It is clear that lack of mode-space correlation implies lack of propagation, while the two notions are in fact equivalent if none of (B.5a)-(B.5c) ever occur; i.e., if the dynamics matrix \( A \) is such that distinct modes have distinct frequencies and if \( A \) also has no null vectors. It is frequently the case that none of (B.5a)-(B.5c) hold for a given linearized atmospheric model, as will be discussed later.

Our main point will be to show that, were it not for the taking of observations, it would not be unusual for each \( \mathbf{P}^a_k \), \( k = 0, 1, 2, \cdots \), not to propagate. Specifically, we will show that assumption 1 of section 4 is essentially an assumption that the portion of each \( \mathbf{P}^a_k \) due to accumulated model error does not propagate. Further, for all “realistic” observing patterns, the introduction of observations causes \( \mathbf{P}^a_k \) to propagate, by introducing correlations in mode space. In particular, it will be shown that nonpropagation would generally require the number of observations \( p_k \) at any given observing time \( t_k \) to be at least as large as the number of state variables \( n \), or at least as large as the number of slow modes in case assumption 2 of section 4 is also made, and it would further require certain relationships between observation error variances to be satisfied. Also we will indicate how assumption 1 of section 4 leads to certain computational simplifications of the Kalman filter, in physical space as well as in mode space, making that assumption even more worthwhile.

b. Propagation of model error covariances

Equations (2.4b)-(2.4d) may be combined to give

\[
\mathbf{P}^a_k = \mathbf{A} \mathbf{P}^a_{k-1} \mathbf{A}^T + \mathbf{Q}_{k-1} - \mathbf{G}_k, \quad \text{(B.6)}
\]

where

\[
\mathbf{G}_k = \mathbf{K}_k \mathbf{H}_k \mathbf{P}^f_k \mathbf{H}_k^T (\mathbf{H}_k \mathbf{P}^f_k \mathbf{H}_k^T + \mathbf{R}_k)^{-1} \mathbf{H}_k \mathbf{P}^f_k. \quad \text{(B.7)}
\]

Iterating (B.6) back to time \( t_0 \) gives

\[
\mathbf{P}^a_k = \mathbf{A}^p \mathbf{P}^a_0 \mathbf{A}^{T_k} + \mathbf{Q}^e_k - \mathbf{G}^e_k, \quad \text{(B.8)}
\]

where

\[
\mathbf{Q}^e_k = \sum_{l=0}^{k-1} \mathbf{A}^l \mathbf{Q}_{k-1-l} \mathbf{A}^{T_l}, \quad \text{(B.9)}
\]

is the contribution to \( \mathbf{P}^a_k \) solely from the model error covariances, and

\[
\mathbf{G}^e_k = \sum_{l=0}^{k-1} \mathbf{A}^l \mathbf{G}_{k-l} \mathbf{A}^{T_l}, \quad \text{(B.10)}
\]

is the contribution from the observations. The matrix \( \mathbf{G}^e_k \) also depends implicitly upon the model error covariances, since \( \mathbf{G}_{k-l} \) depends on \( \mathbf{P}^a_{k-l} \) and therefore on \( \mathbf{Q}_{k-l-1} \), but \( \mathbf{G}^e_k = 0 \) if there are no observations through time \( t_k \).

If one defines \( \hat{\mathbf{Q}}^e_k \) via

\[
\hat{\mathbf{Q}}^e_k = \mathbf{V} \hat{\mathbf{Q}}^e_k \mathbf{V}^*, \quad \text{(B.11)}
\]

then, according to the definition of nonpropagation for \( \mathbf{P}^a_k \), the condition for the contribution from \( \hat{\mathbf{Q}}^e_k \) not to propagate (to time \( t_{k+1} \)) is that, for each \( i \neq j \) with

\[
\hat{Q}^e_{ijk} = 0, \quad \text{(B.12)}
\]

or one of (B.5a)-(B.5c) holds. From (4.2), (4.8), (B.9), and (B.11), one has

\[
\hat{Q}^e_k = \sum_{l=0}^{k-1} \hat{\mathbf{A}}^l \hat{Q}_{k-l-1} \hat{\mathbf{A}}^{*l}. \quad \text{(B.13)}
\]

Therefore, the condition for \( \hat{\mathbf{Q}}^e_k \) not to propagate to time \( t_{k+1} \) for each \( k = 1, 2, 3, \cdots \), is simply that either
or that one of (B.5a)–(B.5c) holds, for each \( i \) and \( j \) with \( i \neq j \) and for each \( k = 0, 1, 2, \ldots \).

In particular, under assumption 1 of section 4 that each \( \mathbf{Q}_k \) be diagonal, the contribution to each \( \mathbf{P}_k^a \) from \( \mathbf{Q}_k \) does not propagate. The nonpropagation condition is independent of how modes belonging to repeated eigenvalues are chosen, for if \( \mathbf{Q}_i = \mathbf{Q}_j \) for some \( i \neq j \), then \( \mathbf{Q}_i = \mathbf{Q}_j \) and \( \mathbf{Q}_{ijk} \) is allowed to be nonzero. If none of (B.5a)–(B.5c) hold, then assumption 1 is equivalent to assuming that each \( \mathbf{Q}_k \) does not propagate.

For many linearized atmospheric models in fact, it is the case that none of (B.5a)–(B.5c) hold. Distinct modes having exactly identical frequencies (B.5a) occur in realistic models only by sheer coincidence. Even in very simple models, usually the only distinct modes with identical frequencies will be the stationary modes \( (\mathbf{Q}_i = 0) \) of the model, if in fact there is more than a single stationary mode. The discretization itself may also artificially introduce distinct modes with identical frequencies (cf., Cohn and Dee 1988), at least for very simple dynamics.

Conditions (B.5b) and (B.5c) can arise only as an artifact of discretization, at least for hyperbolic dynamics, since hyperbolic dynamics are supposed to be reversible. Our dynamics matrix \( \mathbf{A} \) has 50 artificial null modes, as discussed in section 4. The usual way that null modes arise is by use of dissipative discretizations, for which the shortest wavelength modes are sometimes nullvectors. Although this situation is artificial, one should observe that our assumption (2.2a) on the true state reads

\[
\mathbf{\hat{v}}_{jk} = \mathbf{\hat{b}}_{jk-1}
\]

for modes with \( \rho_j = 0 \); i.e., the modal component of the true state for any such mode is simply white noise, which is not unreasonable. One might even imagine purposefully employing a dynamics matrix \( \mathbf{A} \) for which all modes with wavelength shorter than some cutoff wavelength are nullvectors. This would simplify computation of the evolution equation (2.4b), or at least the mode–space counterpart (A.4b), by lowering the dimensionality.

The nonpropagation assumption (B.14) and (B.5) itself, and assumption 1 even more so, also happens to simplify (A.4b) computationally, since then each \( \mathbf{Q}_k \) has few nonzero elements. This leads also to a simplification in physical space (2.4b) if \( \mathbf{Q}_k \) is time independent, at least in the usual case that the model time step is shorter than the time between observations. If there are no observations in between times \( t_{k-1} \) and \( t_k \), say, then one has

\[
\mathbf{P}_k = \mathbf{A}^T \mathbf{A} + \mathbf{Q}^{(i)}
\]

where

\[
\mathbf{Q}^{(i)} = \mathbf{V} (\sum_{j=0}^{l-1} \mathbf{\hat{A}} \mathbf{\hat{Q}} \mathbf{\hat{A}}^*) \mathbf{V}^*
\]

if \( \mathbf{Q} \) is diagonal, then (B.16b) can be written as

\[
\mathbf{Q}^{(i)} = \mathbf{V} (\sum_{j=0}^{l-1} (\mathbf{\hat{A}} \mathbf{\hat{Q}} \mathbf{\hat{A}}^*)^j) \mathbf{V}^*,
\]

and the geometric series can be summed explicitly, while if \( \mathbf{Q} \) is only nonpropagating then it is still easy to account separately for the possibly nonzero off-diagonal elements of the sum in (B.16b). In particular, if the dynamics are conservative (all \( \rho_j = 1 \) and \( \mathbf{Q} \) is diagonal, then one has simply \( \mathbf{Q}^{(i)} = \mathbf{Q} \), so that

\[
\mathbf{P}_k = \mathbf{A}^T \mathbf{P}_{k-1}^e \mathbf{A} + \mathbf{Q}^e.
\]

The matrix \( \mathbf{A}^T \mathbf{P}_{k-1}^e \mathbf{A} \) can also be approximated by carrying out one time step of a large time step scheme, as was done by Dee (1990b).

While assumption 1, or an assumption of nonpropagation, will be at best only approximately true in practice, there is some computational advantage to making such an assumption. There is also a large body of literature (cf., Jazwinski 1970) concerning the effect on filter performance of making such simplifying assumptions. While we shall not review that literature here, our remarks to follow are intended to provide just an intuitive, if not quantitative, rationale for making such simplifying assumptions on \( \mathbf{Q}_k \). Thus it will be shown that taking observations nearly always results in propagation of \( \mathbf{P}_k^a \); if the "amount" of propagation due to observations is much larger than the "amount" due to actual model error, then one would expect that taking \( \mathbf{Q}_k \) to be nonpropagating should have little effect on filter performance.

c. Propagation due to observations

Equations (2.4c) and (2.4d) can be combined to give

\[
\mathbf{P}_k = \mathbf{P}_{k-1}^e - \mathbf{G}_k,
\]

where the contribution \( \mathbf{G}_k \) due to the observations was defined in (B.7) and depends on \( \mathbf{P}_{k-1}^e \) as well as on \( \mathbf{H}_k \) and \( \mathbf{R}_k \). The accumulated contribution \( \mathbf{G}_k \) was defined in (B.10). To show that the accumulated contribution at all times \( t_k \) does propagate to time \( t_{k+1} \), it suffices to show that the single contribution \( \mathbf{G}_k \) at any fixed observing time \( t_k \) propagates to time \( t_{k+1} \); hence that \( \mathbf{P}_k^a \) in (B.18) propagates even if \( \mathbf{P}_{k-1}^e \) does not. To do this, we will suppose (at some fixed time \( t_k \)) that both \( \mathbf{P}_k^a \) and \( \mathbf{P}_{k-1}^e \) are nonpropagating and demonstrate that this leads to unrealistic conditions on \( \mathbf{H}_k \) and \( \mathbf{R}_k \). Since we will not need to refer to any other times, the time index \( k \) will be omitted in most of what follows.
To simplify matters we will assume for the moment that
\[ \nu_i \neq \nu_j \quad \text{if} \quad i \neq j, \quad (B.19a) \]
and
\[ p_j > 0 \quad \text{for each} \quad j = 1, 2, \ldots, n; \quad (B.19b) \]
\( \hat{\mathbf{P}}' \) and \( \hat{\mathbf{P}}^* \) are, therefore, diagonal. Also assume for now that
\[ \mathbf{H} \nu \neq 0 \quad \text{for every eigenvector} \ \nu \ \text{of} \ \mathbf{A}; \quad (B.19c) \]
i.e., that every mode is observable (cf., the discussion in appendix A), and also that \( \mathbf{P}' \) is nonsingular, \( (B.19d) \)
or equivalently, that \( \hat{\mathbf{P}}' \) is nonsingular.

Equations (A.4c) and (A.4d) can be combined to give
\[ \hat{\mathbf{P}}' = \hat{\mathbf{P}}' - \mathbf{H} \mathbf{H}^* (\mathbf{H} \hat{\mathbf{P}}' \mathbf{H}^* + \mathbf{R}^{-1} \mathbf{H} \hat{\mathbf{P}}')^{-1} \mathbf{H} \hat{\mathbf{P}}', \quad (B.20) \]
Since \( \hat{\mathbf{P}}' \) is nonsingular, the Woodbury formula (cf., Hager 1989) can be applied here to give
\[ (\hat{\mathbf{P}}')^{-1} = (\hat{\mathbf{P}}')^{-1} + \mathbf{D}, \quad (B.21) \]
where
\[ \mathbf{D} = \mathbf{H}^* \mathbf{R}^{-1} \mathbf{H}, \quad (B.22) \]
as can be verified by multiplying (B.20) and (B.21) together. Recall that \( \mathbf{R} \) is \( p \times p \)—being the number of observations—so that in particular
\[ \text{rank} \ \mathbf{D} \leq p. \quad (B.23) \]
On the other hand, \( \mathbf{D} \) is \( n \times n \) since \( \mathbf{H} \) is \( p \times n \). We are going to show that \( \mathbf{D} \) is nonsingular; i.e., that
\[ \text{rank} \ \mathbf{D} = n. \quad (B.24) \]
From (B.23) it then follows that
\[ p \geq n, \quad (B.25) \]
which is our first “unrealistic” condition on the observations.

The matrix \( \mathbf{D} \) is diagonal according to (B.21), since \( \hat{\mathbf{P}}' \) and \( \hat{\mathbf{P}}^* \) and therefore their inverses, are diagonal. To show that \( \mathbf{D} \) is nonsingular it, therefore, suffices to show that the diagonal elements of \( \mathbf{D} \) are nonzero. To do so, observe that (B.22) can be written as
\[ \mathbf{D} = (\mathbf{R}^{-1/2} \mathbf{H}^*) (\mathbf{R}^{-1/2} \mathbf{H}), \quad (B.26) \]
where \( \mathbf{R}^{-1/2} \) is the unique real symmetric square root of \( \mathbf{R}^{-1} \), which exists because \( \mathbf{R} \) is real symmetric positive definite. According to (B.26), the \( j \)th diagonal element of \( \mathbf{D} \) is just the Euclidean length of the \( j \)th column of \( \mathbf{R}^{-1/2} \mathbf{H} \). Assumption (B.19c) states that every column of \( \mathbf{H} \), defined in (A.3f), has at least one nonzero element. Since \( \mathbf{R}^{-1} \), and therefore \( \mathbf{R}^{-1/2} \), is nonsingular, it follows that each column of \( \mathbf{R}^{-1/2} \mathbf{H} \) also has at least one nonzero element, and therefore has nonzero Euclidean length. Hence, each diagonal element of \( \mathbf{D} \) is nonzero, and (B.25) follows.

It is fairly straightforward to modify the arguments leading to (B.25) in case any of assumptions (B.19) are not satisfied. One finds that the minimum number of observations (B.25) is reduced by a small amount if the number of multiple frequencies (B.19a), zero eigenvalues (B.19b), and unobserved modes (B.19c) are small, as is usual. Assumption (B.19d) is violated directly by assumption 2 of section 4, for as shown in appendix A, each \( \mathbf{P}_{k} \) then has the form (A.21a). In this case, one can simply replace (A.4c) and (A.4d) by (A.22c) and (A.22d) in the arguments above, leading to
\[ (\hat{\mathbf{P}}_k')^{-1} = (\hat{\mathbf{P}}_k')^{-1} + \mathbf{D}_{sk}, \quad (B.27) \]
\[ \mathbf{D}_{sk} = \mathbf{H}^* \mathbf{R}_k^{-1} \mathbf{H}_k = \mathbf{V}^* \mathbf{H}_x \mathbf{R}_k^{-1} \mathbf{H}_x \mathbf{V}_x, \quad (B.28) \]
in place of (B.21) and (B.22). The result, then, is that given (A.2b) and (A.2c), (B.19a) and (B.19b), and \( \mathbf{H}_x \mathbf{V}_x \neq 0 \) for every slow eigenvector \( \nu \) of \( \mathbf{A} \), \( (B.29a) \)
and
\[ \hat{\mathbf{P}}_k' \] is nonsingular, \( (B.29b) \)
then the number of observations \( p_k \) must equal or exceed the number of slow modes. This is still a large number of observations. Condition (B.29b) holds if either \( \mathbf{P}_{k0} \) is nonsingular or if at least one \( \mathbf{S}_l \), \( l < k \), is nonsingular. If this is not the case, then the minimum number of observations is again reduced.

Inequalities like (B.25) on the number of observations are only necessary conditions for the contribution \( \mathbf{G} \) to \( \mathbf{P}^0 \) not to propagate. According to (B.21), (B.27), and (B.28), on the other hand, a condition that \( \mathbf{D} \) defined in (B.22) be diagonal, along with \( \hat{\mathbf{P}}' \), is always sufficient; this condition is also necessary in the presence of assumptions (B.19). To express more concretely the condition that \( \mathbf{D} \) be diagonal, suppose that the observations are all located at grid points and that the observation errors are uncorrelated with one another. In this case the matrix \( \mathbf{D} \) defined by
\[ \mathbf{D} = \mathbf{H}^* \mathbf{R}^{-1} \mathbf{H}, \quad (B.30) \]
is itself diagonal. From (B.22), (B.30), and (A.3f) we then have
\[ \mathbf{D} = \mathbf{V}^* \mathbf{D} \mathbf{V}, \quad (B.31) \]
with both \( \mathbf{D} \) and \( \mathbf{D} \) diagonal.

The matrix \( \mathbf{V}^* \) is nonsingular since \( \mathbf{V} \) is nonsingular, and the matrix \( \mathbf{N} \), defined by
\[ \mathbf{N} = (\mathbf{V}^* \mathbf{V})^{-1}, \quad (B.32) \]
is symmetric positive definite. Definition (B.32) allows (B.31) to be written as
\[ \mathbf{V} \mathbf{D} = \mathbf{N}^{-1} \mathbf{D} \mathbf{V}. \quad (B.33) \]

Premultiplying (B.30) by \( \mathbf{V}^{-1} (\mathbf{V} \mathbf{V}^*) \) and postmultiplying by \( \mathbf{V} \) gives

\[ \mathbf{V}^* \mathbf{N} \mathbf{V} = I, \quad (B.34) \]

so \( \mathbf{N} \) is just the matrix with respect to which the modes are orthonormal. For the shallow-water dynamics, it follows from (4.27) and (4.29) that \( \mathbf{N} \) can be viewed as a discretization of the diagonal matrix \( \mathbf{M} \) defined in (4.18), times a scaling constant. The matrix \( \mathbf{N} \) need not be precisely diagonal for any particular discretization of (3.1), and it is not for ours. For the sake of argument, however, let us presume \( \mathbf{N} \) to be diagonal, in fact

\[ \mathbf{N} = c \mathbf{N}, \quad (B.35) \]

where \( \mathbf{N} \) is the discretization of \( \mathbf{M} \) described following (4.17) and \( c \) is a scaling constant \([ c = \Delta x \Delta y / 2e \text{ in the notation of (4.17)} \]).

Equation (B.33) now reads

\[ c \mathbf{V} \mathbf{D} = \mathbf{N}^{-1} \mathbf{D} \mathbf{V}, \quad (B.36) \]

with \( \mathbf{D} \) and \( \mathbf{N}^{-1} \mathbf{D} \) both diagonal. This is satisfied if both \( c \mathbf{D} \) and \( \mathbf{N}^{-1} \mathbf{D} \) are equal to a constant times the identity matrix,

\[ c \mathbf{D} = \mathbf{N}^{-1} \mathbf{D} = \alpha \mathbf{I}, \quad (B.37) \]

and it is straightforward to show that in fact (B.37) is the only solution of (B.36) if, for example, \( \mathbf{V} \) has at least one column all of whose elements are nonzero. This is the case for our discretization, and for any natural one: the eigenvectors are products of (complex) Fourier modes in the zonal direction, which do not vanish anywhere, and solutions of an auxiliary eigenproblem in the meridional direction, at least one of which vanishes nowhere.

Equation (B.37) is our restatement of the condition that \( \mathbf{D} \) be diagonal, in the case we have described. Our previous necessary condition (B.25) does also follow from (B.37), since (B.37) implies in particular that \( \mathbf{D} \) is nonsingular. From (B.37), and (B.30), it also follows that \( \mathbf{H} = \mathbf{I} \); i.e., that observations are taken at all grid points, and that therefore

\[ \mathbf{R} \mathbf{N} = \alpha^{-1} \mathbf{I}. \quad (B.38) \]

According to (4.18), if we denote the observation error variances by \( \sigma_y^2 (y_j) \), \( \sigma_v^2 (y_j) \), and \( \sigma_w^2 (y_j) \), then (B.38) reads

\[ \sigma_y^2 (y_j) \Phi (y_j) = \sigma_v^2 (y_j) \Phi (y_j) = g^2 \sigma_w^2 (y_j) = \alpha^{-1}, \quad (B.39) \]

where \( \Phi (y_j) \) is given by (3.4). In particular, the wind observation variances are proportional to the height observation variances, with a latitude-dependent constant of proportionality. For example, at the southern boundary \( y = 0 \), (B.39) gives

\[ \sigma_y = g^{-1} \sqrt{\Phi_0} \sigma_v. \quad (B.40) \]

For radiosonde observations of winds (3.15b) this gives \( \sigma_y \approx 49.5 \) m, which is larger than the actual value (3.15c) by a factor of about 4.5. This factor would be even larger had we used an equivalent depth \( \Phi_0 \) corresponding to external modes.

Thus, there is a relative lack of height information in our radiosonde-only experiment, and even more so with the addition of wind profilers. At least in the absence of dissipation and modeling error, one would expect a visible propagation of height observation error into the wind field and, oppositely, propagation of relatively accurate wind information into the height field.

REFERENCES


