A Four-Dimensional Analysis Exactly Satisfying Equations of Motion

ROSS N. HOFFMAN

Atmospheric and Environmental Research, Inc., Cambridge, MA 02139

(Manuscript received 27 June 1984, in final form 28 August 1985)

ABSTRACT

For a discretized deterministic model of the atmosphere, a single point in the model's phase space defines a complete trajectory. It is possible to choose a point which minimizes the differences between the model trajectory starting at the chosen point and all data observed during an analysis period ($-T \leq t \leq 0$). In this way data and model dynamics are combined to yield a four-dimensional analysis exactly satisfying the model equations. This analysis is the solution of the model's equations of motion defined by the optimal initial conditions chosen at $t = -T$. Therefore, provided T is larger than the adjustment time of the model, there should be no need for any initialization at the start of the forecast at $t = 0$.

This report describes some preliminary experiments which use highly simplified filtered and primitive equation models of an atmosphere with f-plane geometry. These simple models are used because of the substantial computational resources required by the minimization method. It is demonstrated that the method is stable in an assimilation cycle, is able to maintain an accurate estimate of the motion field from temperature observations alone and yields a small analysis error. Unfortunately, forecasts made from the four-dimensional analyses exhibit rapid error growth initially; as a result these forecasts are better than ordinary forecasts only for the first 24 h. Beyond 24 h both types of forecasts have the same skill.

1. Introduction

Detailed and accurate analyses are needed as initial conditions for numerical weather prediction (NWP). Observations are made irregularly in space and time, they are inexact, and, if not for satellite observations, some regions of the globe are nearly data voids. Observations are also incomplete in the sense that not all variables are observed, e.g., a sensor may observe temperature but not velocity. In addition, the initial conditions must be carefully adjusted, i.e., initialized so that spurious large amplitude gravity waves are not present in the forecast. At operational NWP centers, four-dimensional (4D) data assimilation is the process that satisfies these requirements, combining current and past data. A dynamical model of the atmosphere provides the necessary time continuity.

The results of a 4D assimilation may be considered a 4D analysis, but not one which satisfies the governing dynamical equations. The 4D assimilation is usually an intermittent process of initialization, prediction, and 3D analysis regularly repeated (e.g., Bengtsson et al., 1982). As a consequence, the time evolution of the 4D assimilation products are discontinuous at the times data is analyzed. The other common method of 4D assimilation, the continuous dynamical assimilation method, involves adding forcing functions to the governing dynamical equations which nudge the model state towards the observations (e.g., Lyne et al., 1982). In this case, the 4D assimilation products are the solution of a model, but not of a model of the atmosphere since the governing dynamical equations have been altered. In both of these cases, the degree to which the 4D assimilation result departs from the true governing dynamical equations is greatest when and where data are most plentiful.

The topic of this paper is, on the other hand, a method of the 4D analysis that produces a result exactly satisfying the governing dynamical equations. That is, the desired analysis will be *the* particular representation of the atmosphere (in space and time) which most closely agrees with the data while satisfying the governing dynamical equations. This characterization quite naturally leads to a minimization problem which may be identified with Sasaki's (1970) variational analysis method for the special case when the model governing equations are used as strong (i.e., exact) constraints. We will call this problem the 4D variational analysis problem. In principle, this problem may be solved as a difficult nonlinear optimization problem with nonlinear constraints. However, a much simpler unconstrained but equivalent problem may be stated: Find the initial conditions at the start of the analysis period such that the corresponding model evolution best fits the data. In this study, the formalism for this method is developed, and some preliminary experiments are described and discussed.

In the 4D variational analysis problem data at the current time influences the analysis at past times, as well as vice versa. As a practical matter, the governing equations referred to are the governing equations of a model and it must be recognized that there is a limit

to how long a 4D analysis can exactly satisfy these equations *and* simultaneously fit the data well. Clearly the atmosphere does not exactly satisfy any model equations. Thus, this study finds exact solutions of model equations which best fit data at several time levels within a reasonably short time interval. The primary motivation for doing this is the belief that it is best to allow the model itself to reach a state of balance by forward time integration. In addition, the use of the governing equations as constraints implies that the governing dynamical equations are themselves a type of a priori information and not just a vehicle for advecting observations. Also, if 4D variational analyses were available operationally, they would be extremely valuable for diagnostic studies of the atmosphere; 4D assimilation results, on the other hand, are always limited by the fact that in conventional systems certain diagnostic relationships are assumed to be valid for the atmosphere. Furthermore, for the purpose of diagnostic studies, the use of past, present, and future data should be beneficial.

In the current study we follow Sasaki (1958, 1970), Tadjbakhsh (1969), Thompson (1969), and Lewis et al. (1983) in using a variational method to analyze data at two or more time levels simultaneously. There have been many other studies of variational analysis methods in which an evolution equation is used as a so-called dynamical constraint. In these other studies, however, data from only a single time level is used and the dynamical constraint is really a balance or diagnostic constraint. For example, one of the most sophisticated of these studies is Achtemeier's (1975) use of the adiabatic primitive equations as constraints in a variational initialization scheme.

The 4D variational analysis problem is also under study by LeDimet and Talagrand (1985) and by Lewis and Derber (1985). In these studies, the adjoint method of solution is used. Bube (1981) has studied the related problem of obtaining solutions of first-order hyperbolic partial differential equations when incomplete data are available for a time interval and model and observational errors are absent.

2. Nonlinear least squares formalism

The nonlinear least-squares formalism needed to fit data by a model evolution is developed here. Data are observed during an analysis period ranging from time $t = -T$ to $t = 0$, which is the start of the forecast period. There are N observations, denoted $\tilde{Z}_n$, $n = 1, \cdots, N$. We define the best model evolution as that which best fits the $\tilde{Z}_n$. The only restriction on the type of observations which may be used is that it must be possible to calculate a unique estimate of each observation from knowledge of the model evolution. This estimate of $\tilde{Z}_n$ is denoted Z_n. The $\tilde{Z}_n$ may be preprocessed in any way or they might be data which is essentially unprocessed, such as radiances, provided it is possible to obtain Z_n diagnostically from the model variables. In this manner, the 4D analysis problem may encompass inverse problems associated with remotely sensed measurements. The Z_n could also involve horizontal and temporal integration of the model variables. For example, if $\tilde{Z}_n$ was the measured height of a river or stream, calculating Z_n might require integrating the precipitation produced by the model.

The model evolution is determined by $\mathbf{X}$, the initial state at $t = -T$. Therefore, Z_n may be considered to be a function of $\mathbf{X}$. For example, if $\tilde{Z}_n$ was the current temperature measured at 500 mb at one of the model's grid points then Z_n is the model's prediction of that quantity made by integrating the model from time $-T$ to time 0 using $\mathbf{X}$ as initial conditions. Once the optimal $\mathbf{X}$ is found, the analysis at all times from $-T$ to 0 is given by integrating the model in time from these initial conditions. In this work, the optimal $\mathbf{X}$ minimizes the total squared error

$$S(\mathbf{X}) = \sum_n [f_n(\mathbf{X})]^2 \tag{1}$$

where

$$f_n(\mathbf{X}) = w_n[\tilde{Z}_n - Z_n(\mathbf{X})] \tag{2}$$

is the weighted residual for the nth observation. Other definitions of optimal are possible and any reasonable metric might be used to define S, but a squared error norm is most convenient.

The solution of the problem defined here may be considered a "suboptimal" Kalman filter since the Kalman filter may also be defined as a single large least squares problem (Paige and Saunders, 1977). In Kalman filter formalism, the residual, $\tilde{\mathbf{Z}} - \mathbf{Z}$, is called the innovation vector when $\mathbf{Z}$ is calculated using the optimal (filtered) reconstruction of the model state. Here we have assumed no system noise (i.e., model error); this assumption may be viewed as a set of constraints added to the least squares Kalman problem.

a. *The weights*

In specifying the weights, w_n, several factors must be accounted for. First, the weights should take into account the relative size of the measurement errors of the observing systems. Second, the weights must reflect spatial correlations between the observation errors. Third, the weights should depend on the age of the data, thereby accounting for the effect of model errors. A discussion of these last two factors follows. However, the scheme presented here allows for any positive w_n; presumably different weights should be used for different purposes. For this reason, in this preliminary study, the best specification of the weights is not addressed; rather in the experiments reported here the weights are specified in an ad hoc manner (Section 4).

When observational errors at a single time are not independent, e.g., satellite temperatures, their spatial correlations may be accounted for by forming the $\tilde{Z}_n$

as linear combinations of the actual observations such that these linear combinations have associated errors which are expected to be orthogonal. This would require an estimate of the observational error covariance matrix. For example, suppose the actual observations $\tilde{Y}_m$ have errors which are normally distributed with zero mean but which are correlated. Then we wish to choose $\mathbf{A}$ so that

$$\tilde{\mathbf{Z}} = \mathbf{A}\tilde{\mathbf{Y}}$$

has unit error covariance matrix. The proper choice of $\mathbf{A}$ is found by setting $(\mathbf{A}^T\mathbf{A})^{-1}$ equal to the error covariance matrix of the $\tilde{\mathbf{Y}}$. Since the error covariance matrix is real and symmetric, it may be decomposed as $\mathbf{QDQ}^T$ where $\mathbf{Q}$ is orthonormal and $\mathbf{D}$ is diagonal; thus, $\mathbf{A} = \mathbf{D}^{-1/2}\mathbf{Q}^T$. Since $\mathbf{Q}$ and $\mathbf{D}$ contain the eigenvectors and eigenvalues of the error covariance matrix, respectively, the $\tilde{Z}_n$ may be identified as the projection normalized by its expected error of the observation onto the nth empirical orthogonal function of the errors. In most cases, the error covariances must at least in part be modeled; adaptive estimation of some parameters describing the error statistics is possible (e.g., Dee et al., 1985).

When specifying the age dependence of the weights, it is important to note that since the least squares fit automatically accounts for the growth (or decay) of initial state error, the weights need only account for the effect of model error. To see how the variational analysis accounts for initial state error amplification, consider the case with observations only at time $-T$ and time 0 and a model which amplifies errors by a factor A over that time period. Further, suppose the errors at both times are of size ϵ and are uncorrelated. First, consider the forecast through the data at time $-T$ as a 4D analysis. There will be no contribution to S from $t = -T$ but at $t = 0$ the forecast error will be $O(A\epsilon)$ and S will be $O[(A^2 + 1)\epsilon^2]$ since the forecast error and observational error at $t = 0$ are uncorrelated. This is a poor approximation to the 4D variational analysis. On the other hand, the hindcast through the data at $t = 0$ will result in a hindcast error of $O(A^{-1}\epsilon)$ and a value of S of $O[(A^{-2} + 1)\epsilon^2]$. Clearly this is a much better approximation to the 4D variational analysis. We conclude that the 4D variational analysis must be closer to the hindcast through the current data than the forecast through the past data. In this sense, the current data are automatically weighted more heavily when errors are amplified. This argument can be made more explicit by first linearizing the governing dynamical equations about the true evolution and projecting the data onto the eigenvectors of the transition matrix from $t = -T$ to $t = 0$ of the linearized dynamics. The above argument then holds for the coefficients of each mode where A is now the eigenvalue associated with that mode. Further, the 4D variational analysis for the coefficient of each mode is a weighted average of the solutions through the two observations with weights for the current data equal to A^2 times the weight for the past data.

b. Optimization method

Minimizing S is a nonlinear least squares problem. Such problems may have multiple solutions and it may be difficult to find any solution. However, within the context of an assimilation cycle, a good initial estimate, $\mathbf{X}^{(0)}$, will be available from the end of the previous 4D analysis. As a consequence $\mathbf{X}^{(0)}$ contains information for $t < -T$, and may therefore be considered an observation of $\mathbf{X}$. Therefore, as long as there is one more observation, there will be more data than unknowns and minimizing S will be an overdetermined problem.

In the present case, since S is defined by a squared error norm and since a good initial estimate is available, the solution is found by the Gauss-Newton method. In each Gauss-Newton iteration, the f_n are linearized about $\mathbf{X}^{(k)}$ and the resulting linear least squares problem for $\mathbf{P} = (\mathbf{X}^{(k+1)} - \mathbf{X}^{(k)})$ is solved by standard methods. This method is an approximation to a true Newton method for minimizing S. A Newton method finds $\mathbf{X}$ such that the gradient of S vanishes, i.e., such that

$$\frac{1}{2}\frac{\partial S}{\partial X_i} = \sum_n f_n(\mathbf{X})\frac{\partial f_n(\mathbf{X})}{\partial X_i} = \mathbf{0}$$

Expanding f_n and $\partial f_n/\partial X_i$ in Taylor series and neglecting higher order terms yields

$$\sum_j \sum_n \left[\frac{\partial f_n}{\partial X_i}\frac{\partial f_n}{\partial X_j} + f_n\frac{\partial^2 f_n}{\partial X_i \partial X_j}\right] P_j = -\sum_n f_n \frac{\partial f_n}{\partial X_i}$$

where f_n and its derivatives are evaluated at $\mathbf{X}^{(k)}$. If the fit to the observations is good, the f_n will be small and the term involving the second derivatives of f_n may be neglected in the above equation. This yields the Gauss-Newton system of equations. Clearly the Gauss-Newton method will be a poor approximation to the Newton method and may have convergence problems either (1) when $\mathbf{f}$ is large because the initial estimate $\mathbf{X}^{(0)}$ is poor or because the model is inappropriate or (2) when $\partial^2\mathbf{f}$ is large because the problem is strongly nonlinear. Alternatives to the Gauss-Newton method are quasi-Newton methods (Dennis and Moré, 1977). In quasi-Newton methods the $\partial^2\mathbf{f}$ terms are not neglected; instead, the entire term in brackets (the so-called Hessian matrix of S) is approximated. The gradient of S is calculated at each iteration, and is used to update the approximation of the Hessian.

In this discussion it has been assumed that the $f_n(\mathbf{X})$ have continuous second derivatives. For the model described in the next section, this is a justifiable assumption; in general it is not. This assumption is not justified for any model which includes an "on–off switch" in a parameterization, such as a convective adjustment. When this assumption does not hold the finite difference Gauss-Newton method may still be implemented

but convergence problems and sensitivity to step length should be anticipated. Further, when this assumption does not hold, it is still possible to state and solve the 4D analysis problem since special minimization methods are available for nondifferentiable functions (e.g., Dixon, 1980); however, such general methods are more computationally expensive.

The Gauss-Newton system of equations may be written as

$$\mathbf{J}^T\mathbf{J}\,\mathbf{P} = -\mathbf{J}^T\mathbf{f}$$

where $\mathbf{f}$ is the vector of residuals and $\mathbf{J}$ is the Jacobian matrix; $J_{ni} = \partial f_n/\partial X_i$. To solve this system $\mathbf{J}$ is decomposed into $\mathbf{QR}$-form, where $\mathbf{R}$ is a right triangular matrix and $\mathbf{Q}^T\mathbf{Q} = \mathbf{I}$. For overdetermined systems, as is the case here, $\mathbf{R}$ is nonsingular and the system reduces to

$$\mathbf{R}\,\mathbf{P} = -\mathbf{Q}^T\mathbf{f}$$

which is easily solved by back substitution.

In practice $\mathbf{J}$ is approximated by the uncentered difference

$$J_{ni}(\mathbf{X}) = h_i^{-1}\{f_n(\mathbf{X} + h_i\mathbf{u}_i) - f_n(\mathbf{X})\}$$

where $\mathbf{u}_i$ is the ith unit vector and h_i is the step size appropriate for X_i. In this study, h_i is specified as $\epsilon_0(\gamma_0 s_i)$ where ϵ_0 is a small number, γ_0 is a constant equal to 3/40 and s_i is an approximate climate standard deviation of X_i. As discussed below (Section 3b), $\gamma_0 s_i$ is the nominal size of the measurement errors. Note that once $f_n(\mathbf{X})$ is known, it requires one model integration to determine one column of $\mathbf{J}$. Convergence is deemed to have occurred when the change in S is small: i.e., when

$$(S^{(k+1)} - S^{(k)})^2/(S^{(k+1)}\,S^{(k)}) < \epsilon_1{}^2.$$

In the experiments reported below, the values of ϵ_0 and ϵ_1 used are 10^{-5} and 10^{-6}. In all cases, only 3 or 4 iterations were required. For each case, the last gradient of S calculated corresponds to the next to last iterate of $\mathbf{X}$; in all cases, the magnitude of this gradient was smaller than 10^{-4} times the magnitude of the gradient of S at $\mathbf{X}^{(0)}$.

3. Experimental design

a. Dynamical models

The primitive equation (PE) and quasi-geostrophic (QG) versions of a simple two layer, f-plane spectral model (Hoffman, 1981; hereafter H) are used in the experiments described below. The model is based on the energetically consistent formulation of Lorenz (1960). Physical processes included are: 1) linear exchange of heat between the layers, 2) linear exchange of momentum between the layers, 3) linear dissipation of momentum in the lower layer (representing the effect of surface friction) and 4) Newtonian heating of the lower layer. These processes have inverse time constants of $fh_0/2$, $f(k_1 - k_0)/2$, $2fk_0$ and $2fh_1$, respectively. The values of the parameters h_0, h_1, k_0 and k_1 used in the experiments reported here and in Hoffman and Kalnay (1983; hereafter HK) are $2k_0 = k_1 = 0.016$ and $h_0 = h_1 = 0.018$. Taking $f = (3\ h)^{-1}$ the corresponding time scales are approximately 2, 3.5, 1 and 0.5 weeks for the above physical processes, respectively. These values are appropriate for the particular spectral truncation used in this study; they are not estimates of these time scales in the real atmosphere, but rather they were chosen to yield aperiodic behavior and thus to simulate an important feature of the real atmosphere. As in HK, the temperature distribution which yields zero Newtonian heating varies with both horizontal coordinates. The nondimensional, horizontally continuous governing equations are (H):

$$\frac{\partial}{\partial t}\theta = -J(\psi,\theta) + \sigma_0\nabla^2\chi - h_1(\theta - \theta^R)$$

$$\frac{\partial}{\partial t}\nabla^2\psi = -J(\psi,\nabla^2\psi) - J(\rho,\nabla^2\rho) + \delta_{\mathrm{PE}}[\nabla\cdot(\nabla^2\rho\nabla\chi + \nabla^2\chi\nabla\rho) + J(\chi,\nabla^2\chi)] - k_0\nabla^2(\psi - \rho)$$

$$\frac{\partial}{\partial t}\nabla^2\rho = -J(\psi,\nabla^2\rho) - J(\rho,\nabla^2\psi) + \nabla^2\chi + \delta_{\mathrm{PE}}\nabla\cdot(\nabla^2\psi\nabla\chi) - k_1\nabla^2\rho + k_0\nabla^2\psi$$

$$\delta_{\mathrm{PE}}\frac{\partial}{\partial t}\nabla^2\chi = \nabla^2\theta - \nabla^2\rho - \delta_{\mathrm{PE}}\{\nabla\cdot(\nabla^2\psi\nabla\rho + \nabla^2\rho\nabla\psi) - \nabla^2(\nabla\psi\cdot\nabla\rho) + \nabla^2[J(\psi,\chi)] + J(\chi,\nabla^2\psi) + k_1\nabla^2\chi\}$$

$$\frac{\partial}{\partial t}\sigma_0 = -\overline{\theta\nabla^2\chi} - h_0\sigma_0 - h_1\sigma_0 + h_1\theta_0$$

$$\frac{\partial}{\partial t}\theta_0 = -h_1\theta_0 + h_1\sigma_0.$$

Here θ_0 is the difference between the average over the two layers of the horizontally averaged potential temperature and the horizontally averaged radiative temperature in the lower layer; σ_0 is half the difference of the horizontally averaged potential temperature in the upper and lower layers; θ is the deviation from the horizontal average of the potential temperature averaged over the layers; θ^R is the radiative temperature corresponding to θ; ψ is the streamfunction averaged over the layers; ρ is half the difference of the streamfunction in the upper and lower layers; χ is the velocity potential function in the lower layer; J is the Jacobian operator; the overbar represents a horizontal average; δ_{PE} is 1 for the PE and 0 for the QG version of the model. The scales for length, time and temperature are L, f^{-1} and $L^2f^2(c_pb)^{-1}$ where c_p is the specific heat of air at constant pressure and b is a constant ($\sim$0.124) which arises from the vertical discretization (Lorenz, 1960). As in the previous studies, (H, HK) the variables are represented spectrally with a double fourier series,

$$\zeta(x, y) = \sum_{\mathbf{I}} \zeta_{\mathbf{I}} \exp(i[mx + ny])$$

for $\zeta = \theta, \psi, \chi$ and ρ and where $\mathbf{I} = (m, n)$. This representation is truncated severely, retaining only wavevectors $\pm(0, 2), \pm(0, 4), \pm(-4, 2), \pm(4, 2), \pm(4, 0), \pm(8, 0)$. In our phase space description (as well as in the actual computer model) only those spectral coefficients corresponding to wavevectors with the plus sign are retained as prognostic variables. The other spectral coefficients are diagnostic; $\zeta_{-\mathbf{I}}$ must be equal to the complex conjugate of $\zeta_{\mathbf{I}}$ if $\zeta(x, y)$ is real. Because of extreme truncation, the state vector for the PE and QG models have only 50 and 26 real entries, respectively. The QG state vector is defined here to contain σ_0, θ_0, and the real and imaginary parts of the prognostic spectral coefficients of θ and ψ. In the QG model, ρ is obtained diagnostically from the thermal wind equation $\nabla^2\theta = \nabla^2\rho$ and χ is obtained from an omega equation derived from the time derivative of the thermal wind equation. (Actually, $\nabla^2\rho$ is the curl of the thermal wind between the two layers.) In this work, $\theta^R = \theta^*(2\cos 2y + 0.5\cos 4x)$, with $\theta^* = 0.008$ which correponds to $\sim$22 K if L is taken to be the radius of the earth.

A convenient measure of global forecast or analysis error, Δ, is defined in terms of the usual Euclidean distance in the QG phase space between forecast and verification. In terms of the model variables this may be written as

$$\Delta^2 = (\delta\theta_0)^2 + (\delta\sigma_0)^2 + 2^{-1}\overline{(\delta\theta)^2} + 2^{-1}\overline{(\delta\psi)^2}$$

where $\delta\zeta$ is the error in ζ. The factor of ½ arises because only spectral coefficients corresponding to wavenumbers in one half of the complex wavenumber plane are included in the definition of phase space. A similar measure defined in terms of temperature alone is

$$D^2 = (\delta\theta_0)^2 + (\delta\sigma_0)^2 + 2^{-1}\overline{(\delta\theta)^2}.$$

It turns out that $\delta\sigma_0$ and $\delta\theta_0$ are usually quite small and the results presented below would not be changed qualitatively if these terms were not included in the definitions of Δ and D. For interpretation of the results presented below, consider the following examples: If the forecast of σ_0, θ_0 and ψ is perfect then a value of Δ of 0.001 is equivalent to an rms temperature error of $\sim$3.95 K. On the other hand, if the forecast of σ_0, θ_0 and θ is perfect then a value of Δ of 0.001 is equivalent to an rms wind error of $\sim$4.17 m s^{-1}, taking 5 as a reasonable value of the typical wavenumber. Roughly speaking then, a nondimensional result of $\Delta = 250 \times 10^{-6}$ is equivalent to a 1 K rms temperature error or a 1 m s^{-1} rms wind error.

b. *Observing systems; weights*

The formalism of Section 2 allows for arbitrary observing systems and leaves open the problem of specifying weights. As a matter of convenience, in the experiments reported here a greatly simplified observing system and ad hoc weights have been used. For each case, for each analysis, the time origin, $t = 0$, is the time associated with the most recent observation, $t = -T$ is the start of the analysis interval, i.e. the time associated with the initial conditions, $\mathbf{X}$, and τ is the time increment between the observing times,

$$t_k = -(K - k)\tau, \quad k = 1, \cdots, K.$$

At each t_k a fixed set of model variables are observed. Each observation is the true value obtained from nature plus a measurement error. The total number of observations in the analysis interval is N. Nature is simulated by either the PE or QG model and forecasts are made using the QG model. The measurement errors are independent for each variable and are temporally white noise with zero mean and standard deviation γs_n where γ is nominally 3/40 and s_n is an approximate climate standard deviation of the corresponding model variable. This value of γ represents rms temperature errors of $\sim$1 K and rms wind errors of $\sim$2 m s^{-1}.

The weight function which appears in Eq. (2) is defined to be

$$w_n = \alpha(t_n)/(\gamma s_n)$$

where t_n is the time at which observation n is made, and

$$\alpha(t) = \begin{cases} 1 - |t|/\tau_M, & -\tau_M \leq t \leq 0 \\ 0, & \text{otherwise.} \end{cases}$$

Here τ_M is the time scale for observations to lose their value; it may be thought of as the memory time of the system. The form of α and the nominal value of τ_M of 16 days is suggested by the observation that the ensemble rms forecast error (Δ) of the QG model grows roughly linear with time and reaches the expected value of a climate mean forecast at $\sim$16 days (cf. HK, Fig. 4a). Although this observation is for the case when the PE is nature, this formulation is also used in the perfect model experiments described here. For the assimilation experiment of Section 4a, the values of the model variables at the end of the previous 4D analysis at $t = -T$ are treated as observations. For these observations the associated weights, $w_n = (K + 1)^{1/2}\alpha(T)/\gamma s_n$ include a factor of $(K + 1)^{1/2}$ to account in an approximate fashion for the fact that these data are a sort of "average" of $K + 1$ "independent measurements" obtained at the K observing times in the interval $-2T < t \leq -T$ and at the end of the second previous 4D analysis at $t = -2T$. For the nominal values of $J = 8$ and $T/\tau_M = 1/4$ the ratio of the weights for the $t = -T$ analyses to the $t = 0$ observations is 9/4. This rather large weight on the previous analysis is used only in the perfect model experiment.

The definitions of w_n and $\alpha(t)$ given above are ad hoc. Other definitions are possible. However, in the next section, it is demonstrated that there is very little sensitivity of the method to large changes in τ_M. We might even consider ordinary forecasts made from the observations at $t = 0$ to be the limiting case of $\tau_M \to 0$. Further, the results discussed in section 4a are con-

sistent with including the factor of $(K + 1)^{1/2}$ in the second definition of w_n above, since the ratio of observational to analysis error is roughly 3 (Fig. 4). The specification of w_n and especially of $\alpha(t)$ given here is far from satisfactory; however, the lack of sensitivity to τ_M suggests that the choice of weights, within reason, is not crucial.

4. Results

a. An assimilation experiment without model error

A series of assimilation experiments using the QG model to generate nature and as the prediction model are described here. Such "identical twin" experiments have no external source of prediction error. These experiments demonstrate that 1) the method is stable within an assimilation, 2) the analysis errors are much smaller than the measurement errors, and 3) stream function can be satisfactorily maintained from observations of the temperature variables. Also, the sensitivity of the results to the magnitude of the measurement error, to the length of the analysis interval and to the interval between observing times is described.

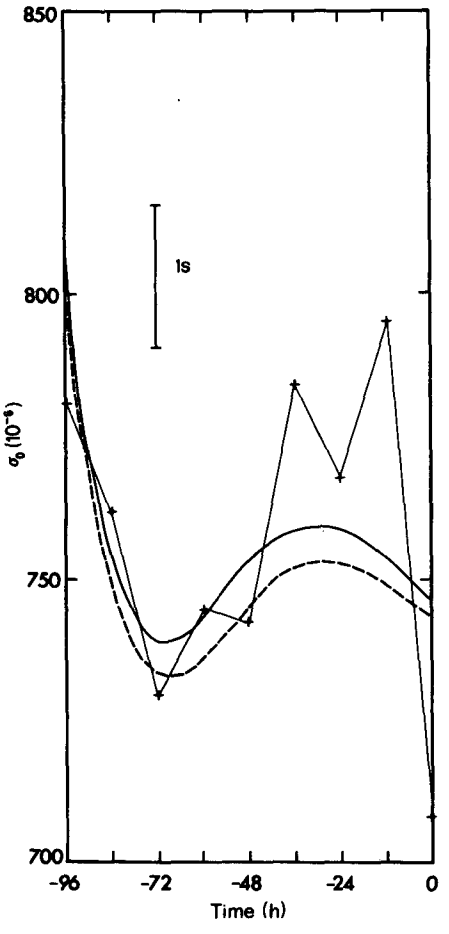

FIG. 1. Evolution of static stability σ_0 during an analysis interval from the standard assimilation experiment. The heavy solid line is the true evolution, observations are plotted as crosses which are connected by light line segments for clarity and the heavy dashed line is the analysis evolution.

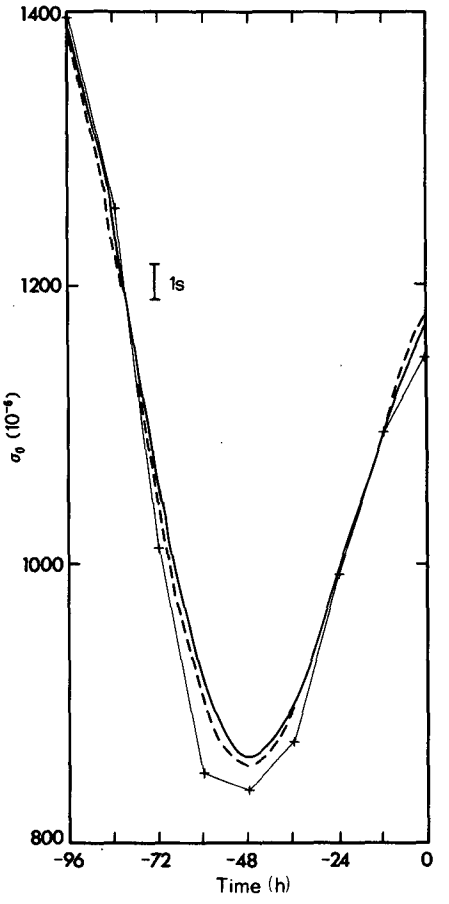

FIG. 2. As in Fig. 1 but for a second case. Note the different vertical scale.

The standard experiment uses the nominal parameter values $T = 4$ days, $\tau = 12$ h, $K = 8$ and $\gamma = 3/40$. Each assimilation experiment is 128 days long. Since the analyses are continuous, i.e., $t = 0$ of one analysis is $t = -T$ for the next, there are 32 analyses in the standard experiment. The experiment begins with a perfect $\mathbf{X}^{(0)}$. The variables observed are σ_0, θ_0 and θ; thus there are $N = 26 + 14K$ observations. Some results are also presented for sensitivity experiments doubling the errors ($\gamma = 3/20$), doubling the interval between observations ($\tau = 24$ h, $K = 4$) and halving the cycle length ($T = 2$ days, $K = 4$). These experiments will be referred to as the 2γ, 2τ and $T/2$ experiments, respectively.

The evolution of σ_0 for two representative analysis intervals is shown in Figs. 1 and 2. In these figures, the 1 s error bars show the standard deviation of the σ_0 measurement errors, which is 25×10^{-6} or 0.07 K dimensionally; σ_0 has a small measurement error since

it is a globally averaged quantity. The 4D analysis is very successful at reconstructing the true evolution. It must be remembered that the 4D analysis makes use of the observations of θ_0 and θ, as well as the observations of σ_0, although it is only the latter which are displayed. In Fig. 3 the ensemble rms errors for σ_0 are shown. These results are typical of the results obtained for other variables and other experiments.

The ensemble rms global analysis errors, $\langle\Delta^2\rangle^{1/2}$ and $\langle D^2\rangle^{1/2}$ are shown in Fig. 4. Comparing curves "D" and "Δ" to the measurement error shows how well the 4D analysis filters the measurement error and maintains an accurate stream function. For the analysis, $\langle D^2\rangle^{1/2}$ is less than 25% of the typical measurement error; the squared error of the observed quantities is reduced by 95%. Also, for the analysis, $\langle\Delta^2\rangle^{1/2}$ is only greater than $\langle D^2\rangle^{1/2}$ by a factor of two. Relative to the natural variability of ψ and θ, ψ, is analyzed as well as θ. (If ψ was also observed with the same relative accuracy as θ, the expected value of $\langle\Delta^2\rangle^{1/2}$ for the observations would be 556×10^{-6}.) It is noteworthy that the analysis error is smallest in the center of the analysis, close to the center of mass of the data. This behavior is typical of methods which analyze (i.e., smooth) data over an interval (Gelb, 1974, Chapter 5). Also, the analysis error is smaller at $t = -T$ than at $t = 0$, which reflects the information content of the previous analysis

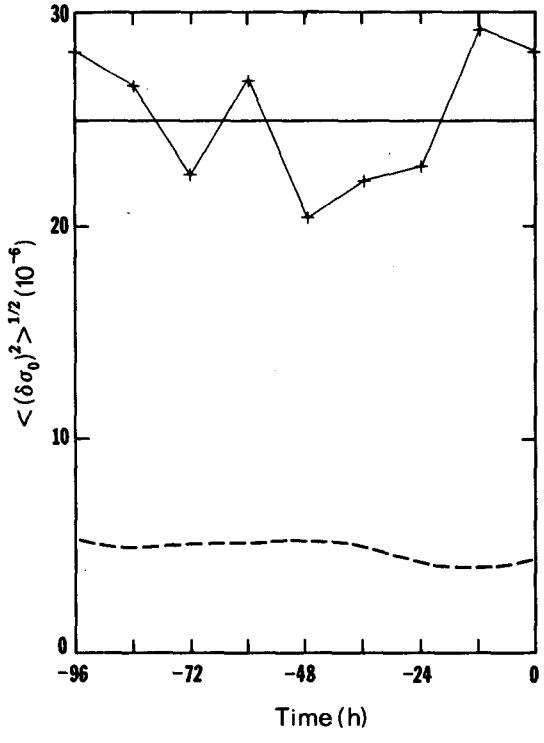

FIG. 3. Ensemble rms analysis and measurement error of static stability, $\langle(\delta\sigma_0)^2\rangle^{1/2}$, as a function of time during the analysis interval for the standard assimilation experiment (heavy dashed line and crosses, respectively). The heavy solid horizontal line at 25×10^{-6} marks the expected rms measurement error.

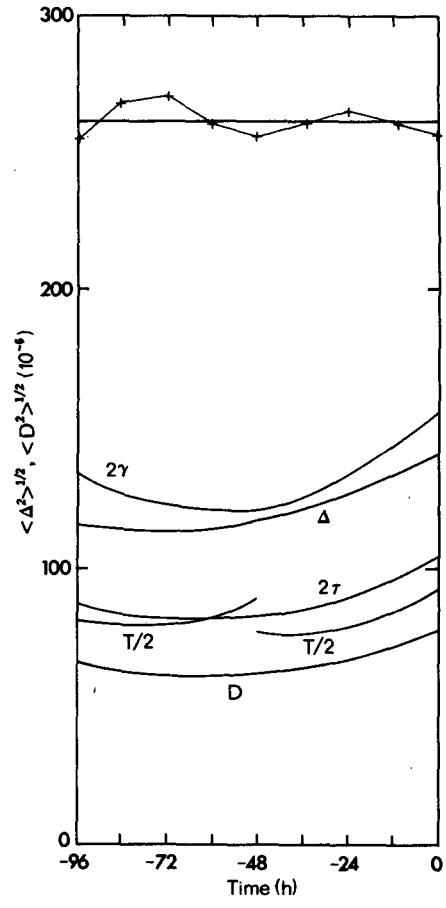

FIG. 4. Ensemble rms global analysis and measurement errors, $\langle\Delta^2\rangle^{1/2}$ and $\langle D^2\rangle^{1/2}$ as functions of time during the analysis interval for the assimilation experiments. The curves labeled Δ and D are these errors for the standard assimilation experiment. Crosses mark $\langle D^2\rangle^{1/2}$ for the observations. These vary only a small amount from the horizontal line plotted at the expected value of $\langle D^2\rangle^{1/2}$ for the observations (261×10^{-6}). Curves labeled 2γ, 2τ and $T/2$ are $\langle D^2\rangle^{1/2}$ for the respective sensitivity experiments.

at $t = -T$. Results for $\langle D^2\rangle^{1/2}$ for the sensitivity experiments are also displayed in this figure. As expected, all three changes had a detrimental effect on the analysis. Doubling the measurement errors (experiment 2γ) approximately doubled the analysis errors. Halving the amount of data used (experiments 2τ and $T/2$) did not have a corresponding large effect on the analysis. For the purpose of this plot, for experiment $T/2$ each standard analysis interval was divided into two intervals, one from -96 to -48 h and one from -48 to 0 h. The two halves of curve 3 are therefore not identical. They are very similar, showing that the 32 analysis intervals used here is a sufficient number to obtain stable results for global ensemble statistics like $\langle D^2\rangle^{1/2}$.

b. Analysis and forecast experiments with model error

An ensemble of analyses and forecasts is examined here. In this experiment σ_0, θ_0, θ and ψ are observed.

Nature is generated using the PE; forecasts are made with the QG model. Thus there is a source of external prediction error due to model imperfections. The 4D analyses obtained are satisfactory; the analysis errors are much smaller than the measurement errors. However, beyond 24 h, forecasts based on the 4D analyses are only as skillful as forecasts based on the $t = 0$ observations.

An ensemble of 50 isolated analysis intervals, well separated in time, is used here. The nature run and experimental design are precisely the same as in HK. Here $T = 42$ h, $\tau = 6$ h, K = 8 and $\gamma = 3/40$. In each analysis, perfect $\mathbf{X}^{(0)}$ are used for an initial estimate of the optimal $\mathbf{X}$. This strategem makes it easy to find the optimal $\mathbf{X}$, but note that these $\mathbf{X}^{(0)}$ are used only as an initial estimate and are not treated as observations; thus $N = 26K$. This approach is reasonable since in this experiment small changes in the initial $\mathbf{X}^{(0)}$ would not alter the minimizing analyses and since any application will be within an assimilation.

The evolution of global analysis/forecast error obtained from this experiment is shown in Fig. 5 for $t = -42$ h to $t = 96$ h. For comparison, forecasts made from the observations at $t = 0$ and from error-free values of the variables from the nature run at $t = 0$ are also displayed. These three evolutions are termed the 4D analysis forecast (4DAF), the ordinary dynamical forecast (ODF), and the perfect initial conditions forecast (PIF), respectively. Beyond about 24 h the ensemble rms forecast error grows roughly linearly with time at the same rate for all three forecasts. At $t = 0$ the analysis error for the 4D analysis is less than half the expected rms measurement error. However, the forecast error for the PIF and 4DAF are already growing roughly linearly at $t = 0$ at the same rate observed at later times. For the ODF on the other hand, there is no initial error growth rate and it takes 24 h for this growth rate to build up.

Because of these disappointing forecast results, an attempt was made to give more weight to the most current observations by using $\tau_M = 4$ days. As seen in Fig. 5, the effect is small although the analysis error at $t = 0$ is slightly reduced. The difference between the two 4D analyses and forecasts is very small.

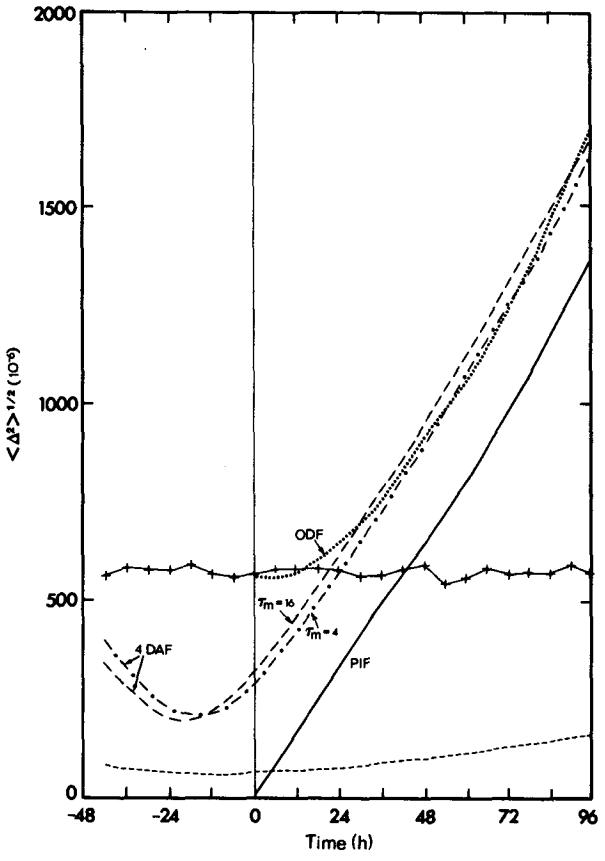

FIG. 5. Ensemble rms global error $\langle \Delta^2 \rangle^{1/2}$ for the analysis/forecast experiments as functions of time for the ODF (dotted line), PIF (solid line) and 4DAF (dashed line). A second 4DAF (dash–dotted line) using $\tau_M = 4$ days is also shown. The measurement error (crosses) has an expected value of 580×10^{-6}. The ensemble rms global difference between the two 4DAF is also shown (short dashed line).

5. Discussion: Growth of error

In general, the evolution of error is governed by the following four factors:

(i) A source of error due to observational errors;
(ii) A source of error due to modeling errors;
(iii) The growth of error because forecasts initially close together diverge;
(iv) The saturation or leveling off of the error growth curve at long forecast times when the prediction becomes uncorrelated with nature.

This last factor does not become important for the simple models used here until roughly two weeks (HK), well beyond the right edge of Fig. 5. Therefore, for the purpose of discussion, it is useful to idealize the initial growth of error as a process governed by linear dynamics with initial conditions due to observational errors, inhomogeneously forced by modeling errors. The general solution to this idealized problem may be described as the sum of exponential solutions of the homogeneous problem and a particular solution of the inhomogeneous problem.

Eventually, the exponentially growing solutions dominate the evolution of error for each individual forecast. However, the time of maximum error growth varies considerably from case to case, resulting in an ensemble average error growth which is nearly linear in time from 36 h until saturation becomes important (HK, Fig. 7).

At the start of the forecast, the initial conditions will project on both exponentially growing and decaying solutions. The random initial conditions used in the ODF experiments project on the decaying solutions just enough to yield zero initial error growth. Initial errors may even decay by this mechanism (Lorenz,

personal communication, 1976), as was seen in some of the classical predictability experiments of Charney et al. (1966). This is also seen in our 4DAF experiments if we consider, for the moment, the 4DAF as a forecast beginning at $t = -T$. That is, we may view the 4DAF as a very good but older forecast. By $t = 0$, the exponentially growing solutions dominate the error of the 4DAF.

From this viewpoint, the 4DAF is better than the forecast made with "perfect" initial conditions. In Fig. 5, a curve for the PIF starting from $t = -42$ h would be nearly identical to the one drawn but would be shifted to the left by 42 h. Compared to this curve, the 4DAF is superior past $t = -24$ h, i.e., beyond 18 h into the forecast. Thus, there are QG forecasts better than the PIF. The perfect initial conditions are perfect only in the sense that they represent a QG state which perfectly agrees with the data, they are not optimal for the purpose of forecasting the PE evolution. Leith (1980) has described how best to choose QG initial conditions to forecast the PE evolution; essentially one finds a QG state which, when nonlinearly balanced, yields a PE state which agrees with the data.

Finally, we note that the two 4DAF for $\tau_m = 4$ and 16 diverge very slowly as shown by the short dashed line in Fig. 5. This is unexpected; for random initial errors we found (HK), that the ensemble average error growth rate obtained by comparing pairs of QG forecasts is about ¾ of that obtained by comparing QG and PE forecasts.

6. Summary and concluding remarks

Four-dimensional analyses are obtained in this study by finding initial conditions at the start of each analysis ($t = -T$) which result in a model evolution which best fits all available data. In general, the fit is a weighted fit; in the present study, the dependence of the weights on the age of the data is ad hoc. In future work, methods of prescribing optimal weights should be explored. This method is tested using simple spectral nonlinear models of the atmosphere. It is demonstrated that 1) the 4D analysis errors are much smaller than the measurement errors, 2) the method is stable in the sense that within an assimilation the solution to the nonlinear least squares problem converges without difficulty, and 3) observations of temperature alone are sufficient to maintain an accurate estimate of the velocity field. Forecasts based on these analyses are better than ordinary forecasts made from the observations at the end of the analysis interval ($t = 0$) for the first 24 h of the forecast. Beyond 24 h both forecasts have the same skill.

With regard to these 4D analyses, there are two further concerns which must be addressed. First, is the method applicable to realistic systems? Second, what is the relationship between initialization and 4D analysis? These questions are now discussed in turn.

The method of solution used here is not economical; in fact it is not even feasible for a realistic model. The principal source of inefficiency is calculating the Jacobian matrix **J**, which requires M model integrations, where M is the number of variables required to specify the initial conditions for the model. For example, a dry 18-layer global spectral model with 30 wave triangular truncation has M equal to roughly 50 000 prognostic variables. For M equal to 50 000, if 1000 processors are available, each capable of integrating the model in 10 minutes, it would require about 21 hours just to determine **J** once by the brute force method used here. Even if speed is of no concern, storage and factorization of a matrix as large as **J** presents difficulties. One might argue that if the primary factor limiting NWP becomes observational error, then the rate at which the complexity and resolution of NWP models increase might become small compared to the rate at which computer resources increase. In this case, the method proposed here might become feasible. However, I suspect that the complexity and resolution of NWP models will keep pace with available computer resources.

The method of solution may be made more practical if we can reduce the size of **J** and/or speed up its calculation. It is possible to reduce the size of **J** by limiting the number of independent degrees of freedom which are allowed to be altered. For example, we might include a normal-mode initialization procedure at $-T$ as part of our definition of the model. In this case only some of the normal mode coefficients need to be specified since the coefficients of the initialized modes are generated diagnostically. As an extreme example, if all gravity modes are initialized, M would be reduced to one-third its original size. As for speeding up the calculation of **J**, the fact that interactions in a grid point model are confined to nearby neighbors and that in a spectral model many interaction coefficients are null (Ellsaesser, 1966) should be exploited in calculating **J**. To begin with, consider the advection of a quantity during a single forward time step of a grid point model: the value of that quantity at the end of the step depends only on a small fraction of the model variables at the start of the step. As a consequence, the single time step transition matrix is sparse and much easier to calculate directly than by a finite difference approximation (which would require M integrations, where M is again the number of prognostic variables). Multistep transition matrices could be constructed by multiplying single time step transition matrices. These observations suggest that a more careful and exact calculation of **J** would be considerably faster.

Another procedure for finding the 4D analysis which minimizes S is to use gradient methods such as steepest descent or conjugate gradient methods. Gradients of S with respect to **X** may be calculated by using the adjoint method (LeDimet and Talagrand, 1985). In

general this method also requires calculating transition matrices since they are involved in the dynamics of the adjoint problem. However, when the dynamics of the adjoint problem are written out explicitly as in Lewis and Derber (1985), they may be no more complicated that the original model, and the cost of calculating $\partial S/\partial X_i$ may be as small as the cost of integrating the model.

The method studied here potentially eliminates the need for initialization and avoids rapid adjustments at the start of the forecast. The analyses obtained use the model dynamics to achieve balance. In general, a balanced state is a naturally occurring state. The appropriate balance depends upon the scale and phenomena under consideration. For example, a balanced representation of the global atmosphere would not contain large amplitude fast gravity waves. To the extent that a model simulates the atmospheric phenomena under consideration, states obtained by integrating a model will be balanced, once transients associated with the particular initial conditions have died out.

Therefore, as long as T is large enough, the model should settle down to a balanced state by the start of the forecast. This would eliminate the need for normal mode initialization procedures, and thereby concerns of how to include physical processes like latent heat release and friction in such procedures.

Acknowledgments. I thank M. Ghil, E. Lorenz, S. Cohn and the reviewers of this journal for their comments on the original manuscript. In particular, I thank John Lewis and John Derber for suggesting the two time level example discussed in Section 2. Most of this work was performed while the author was a National Research Council Resident Research Associate at the NASA Goddard Space Flight Center Laboratory for Atmospheres Modeling and Simulation Branch. Additional support was provided under NASA contract NAS-5-27297 to the Universities Space Research Association visiting scientist program and by Air Force Geophysics Laboratory Air Force Systems Command contract F19628-83-C-0027. I thank L. Gibson for carefully typing the manuscript and drafting the figures for this work.

REFERENCES

Achtemeier, G. L., 1975: On the initialization problem: A variational adjustment method. *Mon. Wea. Rev.*, **103,** 1089–1103.

Bengtsson, L., M. Kanamitsu, P. Kållberg and S. Uppala, 1982: FGGE 4-dimensional data assimilation at ECMWF. *Bull. Amer. Meteor. Soc.*, **63,** 29–43.

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

Charney, J. G., R. G. Fleagle, H. Riehl, V. E. Lally and D. Q. Wark, 1966: The feasibility of a global observation and analysis experiment. *Bull. Amer. Meteor. Soc.*, **47,** 200–220.

Dee, D. P., S. E. Cohn, A. Dalcher and M. Ghil, 1985: An efficient algorithm for estimating noise covariances in distributed systems. *IEEE Trans. Automatic Control,* **AC-30, 11.**

Dennis, J. E., and J. J. Moré, 1977: Quasi-Newton methods, motivation and theory. *SIAM Rev.*, **19,** 46–89.

Dixon, L. C. W., 1980: Reflections on nondifferentiable optimization, Part 1, Ball gradient. *J. Opt. Th. Appl.*, **32,** 123–133.

Ellsaesser, H. W., 1966: Evaluation of spectral versus grid methods of hemispheric numerical weather prediction. *J. Appl. Meteor.*, **5,** 246–262.

Gelb, A., Ed., 1974: *Applied Optimal Estimation,* MIT Press, Cambridge, MA 02138, 374 pp.

Hoffman, R. N., 1981: Alterations of the climate of a primitive equation model produced by filtering approximations and subsequent tuning and stochastic forcing. *J. Atmos. Sci.*, **38,** 514–530.

——, and E. Kalnay, 1983: Lagged average forecasting, an alternative to Monte Carlo forecasting. *Tellus,* **35A,** 100–118.

LeDimet, F.-X., and O. Talagrand, 1985: A general variational formalism in meteorology. *Tellus,* (in press).

Leith, C. E., 1980: Nonlinear normal model initialization and quasi-geostrophic theory. *J. Atmos. Sci.*, **37,** 958–968.

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

——, C. M. Hayden and A. J. Schreiner, 1983: Adjustment of VAS and RAOB geopotential analysis using quasi-geostrophic constraints. *Mon. Wea. Rev.*, **111,** 2058–2067.

Lorenz, E. N., 1960: Energy and numerical weather prediction. *Tellus,* **12,** 364–373.

Lyne, W. H., R. Swinbank and N. T. Birch, 1982: A data assimilation experiment and the global circulation during the FGGE special observing periods. *Quart. J. Roy. Meteor. Soc.*, **108,** 573–594.

Paige, C. C., and M. A. Saunders, 1977: Least squares estimation of discrete linear dynamic systems using orthogonal transformation. *SIAM J. Numer. Anal.*, **14,** 180–193.

Sasaki, Y., 1958: An objective analysis based on the variational method. *J. Meteor. Soc. Japan,* **36,** 77–88.

——, 1970: Some basic formalisms in numerical variational analysis. *Mon. Wea. Rev.*, **98,** 875–883.

Tadjbakhsh, I., 1969: Utilization of time-dependent data in running solution of initial value problems. *J. Appl. Meteor.*, **8,** 389–391.

Thompson, P., 1969: Reduction of analysis error through constraints of dynamical consistency. *J. Appl. Meteor.*, **8,** 738–742.