Exploiting Local Low Dimensionality of the Atmospheric Dynamics for Efficient Ensemble Kalman Filtering

Edward Ott,[1] Brian R. Hunt, Istvan Szunyogh, Matteo Corazza, Eugenia Kalnay, D.J. Patil, James A. Yorke, Aleksey V. Zimin, and Eric J. Kostelich

University of Maryland College Park, Maryland, USA;

September 9, 2002

[1] *Corresponding author:* Institute for Research in Electronics and Applied Physics, University of Maryland, College Park, MD, 20742-2431, E-mail:eo4@umail.umd.edu

Abstract

Recent studies (Patil et al. 2001, 2002) have shown that, when the Earth's surface is divided up into local regions of moderate size, vectors of the forecast uncertainties in such regions tend to lie in a subspace of much lower dimension than that of the full atmospheric state vector. In this paper we show how this finding can be exploited to formulate a potentially accurate and efficient data assimilation technique. The basic idea is that, since the expected forecast errors lie in a locally low dimensional subspace, the analysis resulting from the data assimilation should also lie in this subspace. This implies that operations only on relatively low dimensional matrices are required. The data assimilation analysis is done locally in a manner allowing massively parallel computation to be exploited. The local analyses are then used to construct global states for advancement to the next forecast time. Potential advantages of the method are discussed.

1 Introduction

Atmospheric *data assimilation* (*analysis*) is the process through which an estimate of the atmospheric state is obtained by using observed data and a dynamical model of the atmosphere (e.g., Daley 1991; Kalnay 2002). The primary goal of data assimilation is to provide *analyses* that can be used as initial conditions in operational numerical weather predictions, but diagnostic studies, such as efforts to detect changes in the Earth's climate, are also often based on analyses instead of raw observed data.

Since the early sixties operational weather prediction centers have been routinely producing objective atmospheric analyses (Bergthorsson and Doos 1959; Cressman 1959; Gandin 1963). The analysis at a given time instant is a maximum likelihood estimate of the atmospheric state, in which a short-term forecast, usually referred to as the *background* or *first guess field*, is used as a *prior* estimate of the atmospheric state (Lorenc 1986). Then, the observations are assimilated into the background field by a statistical interpolation. This interpolation is performed based on the assumptions that *(i)* the uncertainties in the background field and the observations are normally distributed and that *(ii)* the covariance between different components of the background (formally the *the background covariance matrix*) and the covariances between uncertainties in the noisy observations (formally the *observational error covariance matrix*) are known. In reality, however, the background and the observational error covariance matrices cannot be directly computed. The implementation of a data assimilation system, therefore, requires the development of statistical models that can provide estimates of the background and the observational error covariance matrices. The quality of a data assimilation system is primarily determined by the accuracy of these estimated error covariance matrices.

In this paper, our attention is focused on improving the estimate of the background error covariance matrix. Current operational assimilation schemes assume that the background error covariances are *stationary* (do not change in time), *homogeneous* (do not vary with the physical location except for a possible weak dependence on geographical latitude), and *isotropic* (the covariance between errors at different locations decreases at the same rate in each direction with increasing distance). While these assumptions are obviously not always valid for the background errors, computational constraints so far have prevented operational centers from introducing more realistic covariance models. The main difficulty has been to find an algorithm that could accommodate the more realistic assumptions in a computationally efficient way. Nevertheless, it is a well accepted view that the introduction of an adaptive background error covariance matrix should lead to significant improvements in weather analyses and the ensuing forecasts.

The mathematically consistent technique to define an adaptive background covariance matrix is the Kalman Filter (Kalman 1960; Kalman and Bucy 1961) which utilizes the dynamical equations to evolve the most probable state and the error covariance matrix in time. Although the Kalman Filter approach has been successfully implemented for a wide range of applications and has

been considered for atmospheric data assimilation for a long while (Jones 1965; Petersen 1973; Ghil et al. 1981, Dee et al., 1985), the computational cost involved does not allow for an operational implementation in the foreseeable future (see Daley 1991 for details).

The currently most popular approach to reduce the cost of the Kalman Filter is to use an ensemble of background forecasts to estimate the background error covariances. The usefulness of this so-called Ensemble Kalman Filter approach, has been demonstrated for simplified models of the atmosphere in a series of papers (e.g. Evensen 1994; Houtekamer and Mitchell 1998, 2001; Anderson and Anderson 1999; Hamill and Snyder 2000, Anderson 2001, Bishop et al. 2001, Hamill et al. 2001, Whitaker and Hamill 2002). The Ensemble Kalman Filter approach has the additional appeal of providing initial ensemble perturbations that are consistent with the analysis scheme. This is important because currently implemented operational techniques generate initial ensemble perturbations without use of direct information about the analysis errors (Toth and Kalnay 1997; Molteni et al. 1996). These techniques are obviously suboptimal considering the goal of ensemble forecasting, which is to simulate the effect of the analysis uncertainties on the ensuing forecasts.

Many Ensemble Kalman Filter schemes (Evensen and van Leeuwen 1996; Houtekamer and Mitchell 1998, 2001; Hamill and Snyder 2000, 2001) are based on the Monte Carlo approach: an ensemble of analyses is generated by using the ensemble of background fields (that were initiated from the preceding ensemble of analyses) and an ensemble of the randomly perturbed observed data. The main weakness of this approach is that the ensemble size must be large in order to accurately represent the probability distribution of the background errors. Thus a relatively large forecast ensemble has to be evolved in time, limiting the efficiency of the approach. The most recent papers demonstrated, however, that the required size of the ensemble can be reduced by filtering the long distance covariances from the background field (Houtekamer and Mitchell 1998, Hamill et al. 2001).

Our paper is related to previous work that attempts to construct a simplified Kalman Filter by taking into account the dominant unstable directions of the phase space. The first attempt along this line was a paper by Kalnay and Toth (1994), in which the conventional 3D-Var data assimilation was enhanced by a correction in the direction of a single bred vector. In Fisher (1998) the stationary background error covariance matrix was augmented by a reduced-rank adaptive component. Fisher (1998) used the linearized version of the European Centre for Medium Range Weather Forecast (ECMWF) model to determine the directions that were the fastest growing during the 48-hour period ending at the analysis time. Fisher (1998) then assumed that the adaptive component of the background covariance matrix was dominated by components along the fastest growing directions. Another notable attempt was made by Buehner (2001), who developed a technique to obtain low rank estimates of the analysis covariance matrix. Then, he generated an ensemble of normally distributed random analysis perturbations consistent with the analysis covariance matrix. In his approach the minimization was done with global fields. Kalman square-root filters

(Bishop et al. 2001; Anderson 2001; Whitaker and Hamill 2002; Tippett et al. 2002) also use global fields in each step of the analysis, but in those techniques none of the fields are randomized.

In this paper, we propose an alternative approach to reduce the computational cost of the Ensemble Kalman Filter. The development of this new technique was motivated by the finding of Patil et al. (2001, 2002) that synoptic scale ensemble perturbations frequently evolve in locally low dimensional subspaces. In the new algorithm, there can be as few as one ensemble member for each of the low number of orthogonal directions spanning the local subspaces. Thus the ensemble size is small. Furthermore, the matrix computations involved in the analysis are performed in the local low dimensional subspaces leading to a further significant improvement in computational efficiency.

In the following sections we outline our new approach to data assimilation. We believe that this approach holds the possibility of significant improvements in weather forecasts. We restrict the discussion to a simplest version illustrating our basic ideas. Some notable features of our approach are the following:

1. The data assimilation is done purely locally. This means that assimilations in each local region are independent, thus facilitating a massively parallel approach.

2. Computations in each local region are done in relatively low-dimensional subspaces, potentially greatly speeding the computations.

3. With the parallelization feature (1) and the low dimensional feature (2), we anticipate that the frequency of analyses can be increased (e.g., from the current typical four times per day to 24 times per day). This would allow efficient use of observations currently nonsynchronous with the widely spaced analysis times (thus obviating the necessity of computationally costly schemes such as the four-dimensional variational data assimilation technique whose purpose is to account for nonsynchronous ('asynoptic') data).

2 Local vectors and their covariance

A model state of the atmosphere is given by a vector field $\mathbf{x}(\mathbf{r}, t)$ where $\mathbf{r}$ is two dimensional and runs over discrete values $\mathbf{r}_{mn}$ (the grid in the physical space used in the numerical computations). Typically, the two components of $\mathbf{r}$ are the geographical longitude and latitude, and $\mathbf{x}$ at a fixed $\mathbf{r}$ is a vector of all relevant physical state variables of the model (e.g., wind velocity components, temperature, surface pressure, humidity, etc., at all height levels included in the model). Let u denote the dimensionality of $\mathbf{x}(\mathbf{r}, t)$ (at fixed $\mathbf{r}$); e.g., when five independent state variables are defined at 28 vertical levels, $u = 140$.

Data assimilation schemes generally treat $\mathbf{x}(\mathbf{r}, t)$ as a random variable with a time-dependent probability distribution. The distribution is updated over time in two ways: (i) it is evolved according to the model dynamics; and (ii) it is modified periodically to take into account recent atmospheric observations.

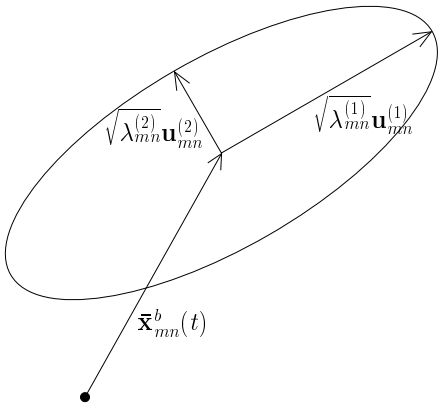

Figure 1: Probability ellipsoid for $\mathbf{x}^b_{mn}$.

We do our analysis locally in model space (grid point by grid point). In this section we introduce our local coordinate system and the approximations we make to the local probability distribution of $\mathbf{x}(\mathbf{r}, t)$. Since all the analysis operations take place at a fixed time t, we will suppress the t dependence of all vectors and matrices introduced henceforth.

Motivated by the works of Patil et al. (2001, 2002) we introduce at each point *local vectors* $\mathbf{x}_{mn}$ of the information $\mathbf{x}(\mathbf{r}_{m+m',n+n'}, t)$ for $-l \leq (m', n') \leq l$. That is, $\mathbf{x}_{mn}$ specifies the model atmospheric state within a $(2l+1)$ by $(2l+1)$ patch of grid points centered at $\mathbf{r}_{mn}$. The dimensionality of $\mathbf{x}_{mn}$ is $(2l+1)^2 u$. We represent the construction of local vectors via a linear operator $\mathbf{M}_{mn}$,

$$\mathbf{x}_{mn} = \mathbf{M}_{mn}\mathbf{x}(\mathbf{r}, t). \tag{1}$$

(See Comment 1 of Section 5 for a generalization.) We now consider local vectors obtained from the model as forecasts, using initial conditions distributed according to the result of the previous analysis, and we denote these by $\mathbf{x}^b_{mn}$ (where the superscript b stands for "background"). Let $F_{mn}(\mathbf{x}^b_{mn})$ be our approximation to the probability density function for $\mathbf{x}^b_{mn}$ at the current analysis time t. A fundamental assumption is that this probability distribution can be usefully approximated as Gaussian,

$$F_{mn}(\mathbf{x}^b_{mn}) \sim \exp\big[-\frac{1}{2}(\mathbf{x}^b_{mn} - \bar{\mathbf{x}}^b_{mn})^T(\mathbf{P}^b_{mn})^{-1}(\mathbf{x}^b_{mn} - \bar{\mathbf{x}}^b_{mn})\big], \tag{2}$$

where $\mathbf{P}^b_{mn}$ and $\bar{\mathbf{x}}^b_{mn}$ are the *local background error covariance matrix* and most probable state associated with $F_{mn}(\mathbf{x}^b_{mn})$ (see Comment 2 of Section 5). Graphically, the level set

$$F_{mn}(\mathbf{x}^b_{mn}) = e^{-1/2}F_{mn}(\bar{\mathbf{x}}^b_{mn}) \tag{3}$$

is an ellipsoid as illustrated in Figure 1. The equation of this *probability ellipsoid* is

$$(\mathbf{x}^b_{mn} - \bar{\mathbf{x}}^b_{mn})^T(\mathbf{P}^b_{mn})^{-1}(\mathbf{x}^b_{mn} - \bar{\mathbf{x}}^b_{mn}) = 1. \tag{4}$$

As explained subsequently, the rank of the $(2l+1)^2u$ by $(2l+1)^2u$ covariance matrix $\mathbf{P}^b_{mn}$ for our approximate probability distribution function F_{mn} is much less than $(2l+1)^2u$. Let

$$k = \text{rank}(\mathbf{P}^b_{mn}); \tag{5}$$

($k = 2$ in Figure 1). Thus $\mathbf{P}^b_{mn}$ has a $(2l+1)^2u - k$ dimensional null space $\bar{\mathbb{S}}_{mn}$ and the inverse $(\mathbf{P}^b_{mn})^{-1}$ is defined for vectors $(\mathbf{x}^b_{mn} - \bar{\mathbf{x}}^b_{mn})$ lying in the k dimensional subspace $\mathbb{S}_{mn}$ orthogonal to $\bar{\mathbb{S}}_{mn}$, with F_{mn} defined to be zero if $(\mathbf{x}^b_{mn} - \bar{\mathbf{x}}^b_{mn})$ is not in $\mathbb{S}_{mn}$.

In the data assimilation procedure we describe in this paper, the background error covariance matrix $\mathbf{P}^b_{mn}$ and the most probable background state $\bar{\mathbf{x}}^b_{mn}$ are derived from a $k'+1$ member ensemble of global state field vectors $\{\mathbf{x}^{b(i)}(\mathbf{r},t)\}$, $i = 1, 2, \cdots, k'+1$; $k' \geq k \geq 1$. The most probable state is given by

$$\bar{\mathbf{x}}^b_{mn} = \mathbf{M}_{mn}\big[(k'+1)^{-1}\sum_{i=1}^{k'+1}\mathbf{x}^{b(i)}(\mathbf{r},t)\big]. \tag{6}$$

To obtain the local background error covariance matrix $\mathbf{P}^b_{mn}$ that we use in our analysis, we first consider a matrix $\mathbf{P}^{b'}_{mn}$ given by

$$\mathbf{P}^{b'}_{mn} = {k'}^{-1}\sum_{i=1}^{k'+1}\delta\mathbf{x}^{b(i)}_{mn}\big(\delta\mathbf{x}^{b(i)}_{mn}\big)^T, \tag{7}$$

where

$$\delta\mathbf{x}^{b(i)}_{mn} = \mathbf{M}_{mn}\mathbf{x}^{b(i)}(\mathbf{r},t) - \bar{\mathbf{x}}^b_{mn}(\mathbf{r},t). \tag{8}$$

Patil et al. (2001, 2002), using 30-pair ensembles of bred vectors (Toth and Kalnay 1993, 1997), have found that forecast errors in the mid-latitude extratropics tend to lie in a low dimensional subset of the $(2l+1)^2u$ dimensional local vector space. Thus we anticipate that we can approximate the background error covariance matrix by one of much lower rank than $(2l+1)^2u$, and this motivates our assumption that an ensemble of size of $k'+1$, where $k'+1$ is substantially less than $(2l+1)^2u$, will be sufficient to yield a good approximate representation of the background covariance matrix. Let the eigenvalues of the matrix $\mathbf{P}^{b'}_{mn}$ be denoted by $\lambda^{(j)}_{mn}$, where the convention for labeling the index j is

$$\lambda^{(1)}_{mn} \geq \lambda^{(2)}_{mn} \geq \ldots \geq \lambda^{(k)}_{mn} \geq \cdots \geq \lambda^{(k')}_{mn}. \tag{9}$$

Since $\mathbf{P}^{b'}_{mn}$ is a symmetric matrix, it has k' orthonormal eigenvectors $\{\mathbf{u}^{(j)}_{mn}\}$ corresponding to the k' eigenvalues (9). Thus

$$\mathbf{P}^{b'}_{mn} = \sum_{j=1}^{k'}\lambda^{(j)}_{mn}\mathbf{u}^{(j)}_{mn}(\mathbf{u}^{(j)}_{mn})^T. \tag{10}$$

We approximate $\mathbf{P}^{b'}_{mn}$ by truncating the sum at $k \leq k'$

$$\mathbf{P}^b_{mn} = \sum_{j=1}^{k}\lambda^{(j)}_{mn}\mathbf{u}^{(j)}_{mn}\big(\mathbf{u}^{(j)}_{mn}\big)^T. \tag{11}$$

In terms of $\mathbf{u}_{mn}^{(j)}$ and $\lambda_{mn}^{(j)}$, the principal axes of the probability ellipsoid (Figure 1) are given by

$$\sqrt{\lambda_{mn}^{(j)}}\mathbf{u}_{mn}^{(j)}. \tag{12}$$

The basic justification for the approximation of the covariance by $\mathbf{P}_{mn}^{b}$ is our supposition, supported by Patil et al. (2002), that the error variance in all other directions is much less than the variance,

$$\sum_{j=1}^{k}\lambda_{mn}^{(j)}, \tag{13}$$

in the directions $\{\mathbf{u}_{mn}^{(j)}\}$, $j = 1, 2, \ldots, k$. In practice we envision that the ensemble size k' is of the order of k.

For the purpose of subsequent computation, we consider the coordinate system for the k dimensional space $\mathbb{S}_{mn}$ determined by the basis vectors $\{\mathbf{u}_{mn}^{(j)}\}$. We call this the *internal coordinate system* for $\mathbb{S}_{mn}$. To change between the internal coordinates and those of the local space, we introduce the $(2l+1)^2u$ by k matrix,

$$\mathbf{Q}_{mn} = \left\{\mathbf{u}_{mn}^{(1)}|\mathbf{u}_{mn}^{(2)}|\cdots|\mathbf{u}_{mn}^{(k)}\right\}. \tag{14}$$

We denote the projection of vectors into $\mathbb{S}_{mn}$ and the restriction of matrices to $\mathbb{S}_{mn}$ by a superscribed circumflex (hat). Thus for a $(2l+1)^2u$ dimensional column vector $\mathbf{w}$, the vector $\hat{\mathbf{w}}$ is a k dimensional column vector given by

$$\hat{\mathbf{w}} = \mathbf{Q}_{mn}^{T}\mathbf{w}. \tag{15}$$

Note that this operation consists of both projecting $\mathbf{w}$ into $\mathbb{S}_{mn}$ and changing to the internal coordinate system. Similarly, for a $(2l+1)^2u$ by $(2l+1)^2u$ matrix $\mathbf{Y}$, the matrix $\hat{\mathbf{Y}}$ is k by k and given by

$$\hat{\mathbf{Y}} = \mathbf{Q}_{mn}^{T}\mathbf{Y}\mathbf{Q}_{mn}. \tag{16}$$

To go back to the original $(2l+1)^2u$ dimensional local vector space, note that $\mathbf{Q}_{mn}^{T}\mathbf{Q}_{mn} = \mathbf{I}$ while $\mathbf{Q}_{mn}\mathbf{Q}_{mn}^{T}$ represents projection on $\mathbb{S}_{mn}$, i.e. it has null space $\bar{\mathbb{S}}_{mn}$ and acts as the identity on $\mathbb{S}_{mn}$. Then, for $\mathbf{w}$ in $\mathbb{S}_{mn}$,

$$\mathbf{w} = \mathbf{Q}_{mn}\hat{\mathbf{w}} \tag{17}$$

and, if $\mathbf{Y}$ is symmetric with null space $\bar{\mathbb{S}}_{mn}$,

$$\mathbf{Y} = \mathbf{Q}_{mn}\hat{\mathbf{Y}}\mathbf{Q}_{mn}^{T}. \tag{18}$$

Note that

$$\hat{\mathbf{P}}_{mn}^{b} = \mathrm{diag}\big(\lambda_{mn}^{(1)}, \lambda_{mn}^{(2)}, ..., \lambda_{mn}^{(k)}\big), \tag{19}$$

and thus it is trivial to invert.

3 Data assimilation

With Section 2 as background, we now consider the assimilation of observational data to obtain a new specification of the probability distribution of the local vector. In what follows, the notational convention of Ide et al. (1997) is adopted whenever it is possible.

Let $\mathbf{x}^a_{mn}$ be the random variable at the current analysis time t representing the local vector after knowledge of the observations is taken into account. For simplicity, we assume that all observations collected for the current analysis were taken at the same time t (see Comment 7 in Section 5). Let $\mathbf{y}^o_{mn}$ be the vector of current observations within the local region, and assume that the errors in these observations are normally distributed with covariance matrix $\mathbf{R}_{mn}$. We also assume that the expected observation vector $\bar{\mathbf{y}}^o_{mn}$ can be written as a linear operator $\mathbf{H}_{mn}$ times the true local state of the atmosphere. (If there are s scalar observations in the local $(2l+1)$ by $(2l+1)$ region at analysis time t, then $\bar{\mathbf{y}}^o_{mn}$ is s dimensional and the rectangular matrix $\mathbf{H}_{mn}$ is s by $(2l+1)^2u$). Then, since we have assumed the background (pre-analysis) state $\mathbf{x}^b_{mn}$ to be normally distributed, it will follow below that $\mathbf{x}^a_{mn}$ is also normally distributed. Its distribution is determined by the most probable state $\bar{\mathbf{x}}^a_{mn}$ and the associated covariance matrix $\mathbf{P}^a_{mn}$. The data assimilation step determines $\bar{\mathbf{x}}^a_{mn}$ (the *local analysis*) and $\mathbf{P}^a_{mn}$ (the *local analysis covariance matrix*).

Since our approximate background covariance matrix $\mathbf{P}^b_{mn}$ has null space $\bar{\mathbb{S}}_{mn}$, perturbations $\Delta\mathbf{x}^a_{mn} = \big(\mathbf{x}^a_{mn} - \bar{\mathbf{x}}^b_{mn}\big)$ outside the k-dimensional subspace $\mathbb{S}_{mn}$ are not allowed. Accordingly, we do data assimilation in $\mathbb{S}_{mn}$, i.e., we consider $\Delta\mathbf{x}^a_{mn}$ to lie in $\mathbb{S}_{mn}$. Thus the data assimilation is done by minimizing the quadratic form,

$$\begin{aligned} J\big(\Delta\hat{\mathbf{x}}^a_{mn}\big) &= \big(\Delta\hat{\mathbf{x}}^a_{mn}\big)^T\big(\hat{\mathbf{P}}^b_{mn}\big)^{-1}\Delta\hat{\mathbf{x}}^a_{mn} \\ &+ \big(\hat{\mathbf{H}}_{mn}\Delta\hat{\mathbf{x}}^a_{mn} + \mathbf{H}_{mn}\bar{\mathbf{x}}^b_{mn} - \mathbf{y}^o_{mn}\big)^T\mathbf{R}^{-1}_{mn} \times \\ &\quad \big(\hat{\mathbf{H}}_{mn}\Delta\hat{\mathbf{x}}^a_{mn} + \mathbf{H}_{mn}\bar{\mathbf{x}}^b_{mn} - \mathbf{y}^o_{mn}\big). \end{aligned} \tag{20}$$

Here $\hat{\mathbf{H}}_{mn} = \mathbf{H}_{mn}\mathbf{Q}_{mn}$ maps $\mathbb{S}_{mn}$ to the observation space, using the internal coordinate system for $\mathbb{S}_{mn}$ introduced in the previous section, so that $\Delta\mathbf{x}^a_{mn} = \mathbf{Q}_{mn}\Delta\hat{\mathbf{x}}^a_{mn}$. The most probable value of $\Delta\hat{\mathbf{x}}^a_{mn}$,

$$\Delta\hat{\bar{\mathbf{x}}}^a_{mn} = \hat{\mathbf{P}}^a_{mn}\hat{\mathbf{H}}^T_{mn}\mathbf{R}^{-1}_{mn}\big(\mathbf{y}^o_{mn} - \mathbf{H}_{mn}\bar{\mathbf{x}}^b_{mn}\big), \tag{21}$$

is the minimizer of $J\big(\Delta\hat{\mathbf{x}}^a_{mn}\big)$, where the analysis covariance matrix $\hat{\mathbf{P}}^a_{mn}$ is the inverse of the matrix of second derivatives (Hessian) of $J\big(\Delta\hat{\mathbf{x}}^a_{mn}\big)$ with respect to $\Delta\hat{\mathbf{x}}^a_{mn}$,

$$\hat{\mathbf{P}}^a_{mn} = \big[\big(\hat{\mathbf{P}}^b_{mn}\big)^{-1} + \hat{\mathbf{H}}^T_{mn}\mathbf{R}^{-1}_{mn}\hat{\mathbf{H}}_{mn}\big]^{-1}. \tag{22}$$

For computational purposes, we prefer to use the alternate form,

$$\hat{\mathbf{P}}^a_{mn} = \hat{\mathbf{P}}^b_{mn}\big[\mathbf{I} + \hat{\mathbf{H}}^T_{mn}\mathbf{R}^{-1}_{mn}\hat{\mathbf{H}}_{mn}\hat{\mathbf{P}}^b_{mn}\big]^{-1}, \tag{23}$$

both in place of (22) and in computing (21). A potential numerical advantage of (23) over (22) is that (22) involves the inverse of $\hat{\mathbf{P}}^b_{mn}$, which may be problematic if $\hat{\mathbf{P}}^b_{mn}$ has a small eigenvalue.

Another alternative is to compute (21) and (22) in terms of the "Kalman gain" matrix

$$\hat{\mathbf{K}}_{mn} = \hat{\mathbf{P}}^b_{mn}\hat{\mathbf{H}}^T_{mn}\big(\hat{\mathbf{H}}_{mn}\hat{\mathbf{P}}^b_{mn}\hat{\mathbf{H}}^T_{mn} + \mathbf{R}_{mn}\big)^{-1}. \tag{24}$$

Then it can be shown (e.g., Kalnay 2002, p. 171) that (21) and (22)/(23) are equivalent to

$$\Delta\hat{\bar{\mathbf{x}}}^a_{mn} = \hat{\mathbf{K}}_{mn}\big(\mathbf{y}^o_{mn} - \mathbf{H}_{mn}\bar{\mathbf{x}}^b_{mn}\big), \tag{25}$$

and

$$\hat{\mathbf{P}}^a_{mn} = \big(\mathbf{I} - \hat{\mathbf{K}}_{mn}\hat{\mathbf{H}}_{mn}\big)\hat{\mathbf{P}}^b_{mn}. \tag{26}$$

Again, the inverse of $\hat{\mathbf{P}}^b_{mn}$ is not required.

Though (21) and (23) are mathematically equivalent to (24)–(26), the former approach may be significantly more efficient computationally for the following reasons. In both cases, one must invert and s by s matrix, where s is the number of local observations. While these matrices are considerably smaller than those involved in global data assimilation schemes, they may still be reasonably large. Generally the s by s matrix $\mathbf{R}_{mn}$ whose inverse is required in (23) will be diagonal or close to diagonal, and thus less expensive to invert than the matrix inverted in (24). (Furthermore, in some cases one may be able to treat $\mathbf{R}_{mn}$ as time-independent and avoid recomputing its inverse for each successive analysis.) The additional inverse required in (23) is of a k by k matrix, where the rank k of $\mathbf{P}^b_{mn}$ is relatively small.

Finally, going back to the local space representation, we have

$$\bar{\mathbf{x}}^a_{mn} = \mathbf{Q}_{mn}\Delta\hat{\bar{\mathbf{x}}}^a_{mn} + \bar{\mathbf{x}}^b_{mn}, \tag{27}$$

$$\mathbf{P}^a_{mn} = \mathbf{Q}_{mn}\hat{\mathbf{P}}^a_{mn}\mathbf{Q}^T_{mn}. \tag{28}$$

(See Comment 3 of Section 5.) For future reference let $\big\{\mathbf{v}^{(j)}_{mn}\big\}$ denote the set of k orthonormal eigenvectors of $\mathbf{P}^a_{mn}$, and let $\big\{\mu^{(j)}_{mn}\big\}$ denote the corresponding set of k eigenvalues. Thus

$$\mathbf{P}^a_{mn} = \sum_{j=1}^{k} \mu^{(j)}_{mn}\mathbf{v}^{(j)}_{mn}\big(\mathbf{v}^{(j)}_{mn}\big)^T. \tag{29}$$

4 Updating the background field

We now wish to use the analysis information, $\mathbf{P}^a_{mn}$ and $\bar{\mathbf{x}}^a_{mn}$, to obtain an ensemble of global analysis fields $\big\{\mathbf{x}^{a(i)}(\mathbf{r},t)\big\}$; $i = 1, 2, \cdots, k'+1$. Once these fields are determined, they can be used as initial conditions for the atmospheric model. Integrating these global fields forward in time to the next analysis time

$t + \Delta t$, we obtain the background ensemble $\left\{\mathbf{x}^{b(i)}(\mathbf{r}, t + \Delta t)\right\}$. This completes the loop, and, if the procedure is stable, it can be repeated for as long as desired. Thus at each analysis time we are in possession of a global initial condition that can be used for making forecasts of the desired durations.

Our remaining task is to specify the global analysis fields $\left\{\mathbf{x}^{a(i)}(\mathbf{r}, t)\right\}$ from our analysis information, $\mathbf{P}^a_{mn}$ and $\bar{\mathbf{x}}^a_{mn}$. Let

$$\mathbf{x}^{a(i)}_{mn} = \bar{\mathbf{x}}^a_{mn} + \delta\mathbf{x}^{a(i)}_{mn} \tag{30}$$

denote $k' + 1$ local analysis vectors, where

$$\sum_{i=1}^{k'+1} \delta\mathbf{x}^{a(i)}_{mn} = 0, \tag{31}$$

and

$$\mathbf{P}^a_{mn} = {k'}^{-1} \sum_{i=1}^{k'+1} \delta\mathbf{x}^{a(i)}_{mn} \left(\delta\mathbf{x}^{a(i)}_{mn}\right)^T. \tag{32}$$

Thus, by (31), the local analysis state $\bar{\mathbf{x}}^a_{mn}$ (determined in Section 3) is the mean over the local analysis ensemble $\left\{\mathbf{x}^{a(i)}_{mn}\right\}$, and, by (32), $\left\{\mathbf{x}^{a(i)}_{mn}\right\}$ gives a representation of the analysis error covariance matrix.

There are many choices for $\left\{\delta\mathbf{x}^{a(i)}_{mn}\right\}$ that satisfy (31) and (32), and in this section we will describe several possible methods for computing a set of solutions to these equations. See also (Tippett et al. 2002) for different approaches to this problem in the global setting. In a given forecasting scenario, one could compare the accuracy and speed of these methods in order to choose among them. There are two main criteria we have in mind in formulating these methods.

First, since we wish to specify global fields that we think of as being similar to physical fields, we desire that these fields be slowly varying in m and n. That is, if $\mathbf{P}^a_{mn}$ is slowly varying, we do not want to introduce any artificial rapid variations in the individual $\delta\mathbf{x}^{a(i)}_{mn}$ through our method of constructing a solution of (31) and (32). For this purpose we regard the background vectors as physical states, and hence slowly varying in m and n. (This is a reasonable assumption since the background ensemble is obtained from evolution of the atmospheric model from time $t - \Delta t$ to time t.)

Second, the method for computing $\left\{\delta\mathbf{x}^{a(i)}_{mn}\right\}$ should be numerically stable and efficient. As in the previous section, we prefer to avoid matrix operations (which in this section include square roots as well as inverses) on s by s (local observation space) matrices, except for $\mathbf{R}_{mn}$, which we expect to be diagonal or near-diagonal and thus significantly easier to invert than a typical s by s matrix. Otherwise, these operations should be done on matrices that are k by k or $k' + 1$ by $k' + 1$.

Thus we are motivated to express the vectors $\delta\mathbf{x}^{a(i)}_{mn}$ in the form

$$\delta\mathbf{x}^{a(i)}_{mn} = \mathbf{Z}_{mn}\delta\mathbf{x}^{b(i)}_{mn}, \tag{33}$$

where the matrix $\mathbf{Z}_{mn}$ can be thought of as a generalized 'rescaling' of the original background fields. This 'rescaling' can be viewed as being similar to the techniques employed in the breeding method (Toth and Kalnay, 1993) and in the Ensemble Transform Kalman Filter approach (Bishop et al., 2001; Wang and Bishop, 2002). If $\mathbf{Z}_{mn}$ varies slowly with m and n, then so will $\delta\mathbf{x}^{a(i)}_{mn}$. Furthermore, (31) will automatically be satisfied, because the background perturbations $\delta\mathbf{x}^{b(i)}_{mn}$ sum to zero; see (6) and (8). Finally, due to (7), the analysis perturbations given by (33) will satisfy (32) if and only if

$$\mathbf{P}^{a}_{mn} = \mathbf{Z}_{mn}\mathbf{P}^{b'}_{mn}\mathbf{Z}^{T}_{mn}. \tag{34}$$

Let

$$\mathbf{Z}_{mn} = \mathbf{Q}_{mn}\hat{\mathbf{Z}}_{mn}\mathbf{Q}^{T}_{mn}, \tag{35}$$

where $\hat{\mathbf{Z}}_{mn}$ is a k by k matrix to be determined. From (10), (11), and (14) we have

$$\hat{\mathbf{P}}^{b}_{mn} = \mathbf{Q}^{T}_{mn}\mathbf{P}^{b}_{mn}\mathbf{Q}_{mn} = \mathbf{Q}^{T}_{mn}\mathbf{P}^{b'}_{mn}\mathbf{Q}_{mn}, \tag{36}$$

so that (34) is equivalent to

$$\mathbf{P}^{a}_{mn} = \mathbf{Q}_{mn}\hat{\mathbf{Z}}_{mn}\mathbf{Q}^{T}_{mn}\mathbf{P}^{b'}_{mn}\mathbf{Q}_{mn}\hat{\mathbf{Z}}^{T}_{mn}\mathbf{Q}^{T}_{mn} = \mathbf{Q}_{mn}\hat{\mathbf{Z}}_{mn}\mathbf{P}^{b}_{mn}\hat{\mathbf{Z}}_{mn}\mathbf{Q}^{T}_{mn}. \tag{37}$$

Thus we need $\hat{\mathbf{Z}}_{mn}$ to satisfy

$$\hat{\mathbf{P}}^{a}_{mn} = \hat{\mathbf{Z}}_{mn}\hat{\mathbf{P}}^{b}_{mn}\hat{\mathbf{Z}}^{T}_{mn}; \tag{38}$$

then $\mathbf{Z}_{mn} = \mathbf{Q}_{mn}\hat{\mathbf{Z}}_{mn}\mathbf{Q}^{T}_{mn}$ solves (34). Equation (38) has infinitely many solutions. In order for the result to be consistent from one grid point to the next, it is important that we use an algorithm for computing a particular solution that depends continuously on $\hat{\mathbf{P}}^{a}_{mn}$ and $\hat{\mathbf{P}}^{b}_{mn}$.

Solution 1. Mathematically, the simplest way to express a solution to (38) is

$$\hat{\mathbf{Z}}_{mn} = \left(\hat{\mathbf{P}}^{a}_{mn}\right)^{1/2}\left(\hat{\mathbf{P}}^{b}_{mn}\right)^{-1/2}, \tag{39}$$

where in (39), by the notation $\left(\mathbf{M}\right)^{1/2}$, we mean the positive symmetric square root of $\mathbf{M}$. Recall that $\hat{\mathbf{P}}^{b}_{mn}$ is diagonal, so the inverse square root is easily computed. Then, from (10) and (29),

$$\mathbf{Z}_{mn} = \sum_{j=1}^{k}\sum_{j'=1}^{k}\mathbf{C}^{(jj')} \tag{40}$$

$$\mathbf{C}^{(jj')} = \sqrt{\frac{\mu^{(j)}_{mn}}{\lambda^{(j')}_{mn}}}\left[\left(\mathbf{v}^{(j)}_{mn}\right)^{T}\mathbf{u}^{(j')}_{mn}\right]\left[\mathbf{v}^{(j)}_{mn}\left(\mathbf{u}^{(j')}_{mn}\right)^{T}\right]. \tag{41}$$

To summarize the result at this point, the $k'+1$ local analysis fields are given by (30), (33), (40), and (41). We emphasize that $\mathbf{Z}_{mn}$ defined in this way will be

slowly varying in m and n if $\mathbf{P}^a_{mn}$ and $\mathbf{P}^b_{mn}$ are. (See Comment 9 in Section 5 for further discussion.)

More generally, we can replace (39) by

$$\hat{\mathbf{Z}}_{mn} = \sqrt{\hat{\mathbf{P}}^a_{mn}}\sqrt{\hat{\mathbf{P}}^{b^{-1}}_{mn}}^T, \tag{42}$$

where for a positive definite symmetric matrix $\mathbf{M}$, we mean by $\sqrt{\mathbf{M}}$ any matrix for which $\sqrt{\mathbf{M}}\sqrt{\mathbf{M}}^T = \mathbf{M}$. Note that this equation does not uniquely determine $\sqrt{\mathbf{M}}$, and that given any solution $\sqrt{\mathbf{M}} = \mathbf{W}$, the most general solution is $\sqrt{\mathbf{M}} = \mathbf{WU}$ where $\mathbf{U}$ is any orthogonal matrix. Furthermore, by considering all possible matrices $\sqrt{\hat{\mathbf{P}}^a_{mn}}$ we obtain all possible solutions $\hat{\mathbf{Z}}_{mn}$ of (39). The square root of $\hat{\mathbf{P}}^a_{mn}$ may be computed by any method, such as Cholesky factorization, provided the result depends continuously on $\hat{\mathbf{P}}^a_{mn}$.

Solution 2. Another way to express a solution to (38), from (Andrews 1968), is

$$\hat{\mathbf{Z}}_{mn} = \mathbf{I} - \hat{\mathbf{P}}^b_{mn}\hat{\mathbf{H}}^T_{mn}\big(\mathbf{W}^{-1}_{mn}\big)^T\big(\mathbf{W}_{mn} + \sqrt{\mathbf{R}_{mn}}\big)^{-1}\hat{\mathbf{H}}_{mn}, \tag{43}$$

where

$$\mathbf{W}_{mn} = \sqrt{\hat{\mathbf{H}}_{mn}\hat{\mathbf{P}}^b_{mn}\hat{\mathbf{H}}^T_{mn} + \mathbf{R}_{mn}} \tag{44}$$

and in each equation any matrix square root, as defined above, may be used. A mathematically equivalent formulation which shifts matrix operations from the observation space into the (generally lower-dimensional) space $\mathbb{S}_{mn}$ is

$$\hat{\mathbf{Z}}_{mn} = \mathbf{I} - \hat{\mathbf{P}}^b_{mn}\big(\hat{\mathbf{W}}^{-1}_{mn}\big)^T\left[\hat{\mathbf{W}}_{mn} + \sqrt{\big(\hat{\mathbf{H}}^T_{mn}\mathbf{R}^{-1}_{mn}\hat{\mathbf{H}}_{mn}\big)^{-1}}\right]^{-1}, \tag{45}$$

where

$$\hat{\mathbf{W}}_{mn} = \sqrt{\hat{\mathbf{P}}^b_{mn} + \big(\hat{\mathbf{H}}^T_{mn}\mathbf{R}^{-1}_{mn}\hat{\mathbf{H}}_{mn}\big)^{-1}}. \tag{46}$$

Taking the inverse of $\hat{\mathbf{H}}^T_{mn}\mathbf{R}^{-1}_{mn}\hat{\mathbf{H}}_{mn}$ may, however, be problematic if the observations are sufficiently sparse in the local region that this k by k matrix does not have full rank.

Solution 3. An earlier method that is similar to Solution 2 but involves $k'+1$ by $k'+1$ dimensional matrices is the "Potter method"; see e.g. (Bierman 1977). Recall that we envision the ensemble size $k'+1$ to be not significantly larger than k, so this method should be similarly efficient. For this solution we abandon (33) and instead solve (31) and (32) as follows. We will write $\delta\mathbf{x}^{a(i)}_{mn} = \mathbf{Q}_{mn}\delta\hat{\mathbf{x}}^{a(i)}_{mn}$, where $\delta\hat{\mathbf{x}}^{a(i)}_{mn}$ lies in $\mathbb{S}_{mn}$. Let $\delta\mathbf{x}^{b(i)}_{mn} = \mathbf{Q}_{mn}\delta\hat{\mathbf{x}}^{b(i)}_{mn}$, and form the $(2l+1)^2u$ by $k'+1$ matrices

$$\hat{\mathbf{X}}^b_{mn} = \big\{\delta\hat{\mathbf{x}}^{b(1)}_{mn}|\delta\hat{\mathbf{x}}^{b(2)}_{mn}|\cdots|\delta\hat{\mathbf{x}}^{b(k'+1)}_{mn}\big\} \tag{47}$$

and

$$\hat{\mathbf{X}}^a_{mn} = \big\{\delta\hat{\mathbf{x}}^{a(1)}_{mn}|\delta\hat{\mathbf{x}}^{a(2)}_{mn}|\cdots|\delta\hat{\mathbf{x}}^{a(k'+1)}_{mn}\big\}. \tag{48}$$

To solve (31) and (32) we require that

$$\sum_{i=1}^{k'+1} \delta\hat{\mathbf{x}}_{mn}^{a(i)} = 0, \tag{49}$$

and

$$\hat{\mathbf{P}}_{mn}^{a} = k'^{-1} \sum_{i=1}^{k'+1} \delta\hat{\mathbf{x}}_{mn}^{a(i)} \big(\delta\hat{\mathbf{x}}_{mn}^{a(i)}\big)^T. \tag{50}$$

We wish to express

$$\hat{\mathbf{X}}_{mn}^{a} = \hat{\mathbf{X}}_{mn}^{b} \hat{\mathbf{Y}}_{mn} \tag{51}$$

for an appropriate matrix $\mathbf{Y}_{mn}$. This differs from (33) in that we multiply $\mathbf{X}_{mn}^{b}$ on the right by $\mathbf{Y}_{mn}$ rather than on the left by $\mathbf{Z}_{mn}$. Note that

$$\hat{\mathbf{P}}_{mn}^{b} = k'^{-1} \hat{\mathbf{X}}_{mn}^{b} \big(\hat{\mathbf{X}}_{mn}^{b}\big)^T \tag{52}$$

and to satisfy (32) we require that

$$\hat{\mathbf{P}}_{mn}^{a} = k'^{-1} \hat{\mathbf{X}}_{mn}^{b} \hat{\mathbf{Y}}_{mn} \hat{\mathbf{Y}}_{mn}^{T} \big(\hat{\mathbf{X}}_{mn}^{b}\big)^T. \tag{53}$$

Potter's expression for the general solution to this equation is

$$\hat{\mathbf{Y}}_{mn} = \sqrt{\mathbf{I} - k'^{-1} \hat{\mathbf{X}}_{mn}^{b^T} \hat{\mathbf{H}}_{mn}^{T} \big[\hat{\mathbf{H}}_{mn} \hat{\mathbf{P}}_{mn}^{b} \hat{\mathbf{H}}_{mn}^{T} + \mathbf{R}_{mn}\big]^{-1} \hat{\mathbf{H}}_{mn} \hat{\mathbf{X}}_{mn}^{b}}. \tag{54}$$

Here the square root is taken of a $k'+1$ by $k'+1$ matrix, but the inverse is of an s by s matrix, where s is the dimension of the local observation space. An equivalent way to write (54) in our setting is

$$\hat{\mathbf{Y}}_{mn} = \sqrt{\mathbf{I} - k'^{-1} \hat{\mathbf{X}}_{mn}^{b^T} \hat{\mathbf{V}}_{mn} \hat{\mathbf{H}}_{mn}^{T} \mathbf{R}_{mn}^{-1} \hat{\mathbf{H}}_{mn} \hat{\mathbf{X}}_{mn}^{b}}, \tag{55}$$

where

$$\hat{\mathbf{V}}_{mn} = \big[\mathbf{I} + \hat{\mathbf{H}}_{mn}^{T} \mathbf{R}_{mn}^{-1} \hat{\mathbf{H}}_{mn} \hat{\mathbf{P}}_{mn}^{b}\big]^{-1}. \tag{56}$$

Now aside from $\mathbf{R}_{mn}$, we need only invert a k by k matrix, and furthermore it is the same matrix that was inverted in (23). As previously discussed, although $\mathbf{R}_{mn}$ is s by s, its inverse is easily computed even when s is much larger than k.

As we will now explain, it is important in our setting to use the positive symmetric square root in (54) or (55). For the ensemble perturbations defined by (51) to satisfy (50), any matrix square root (as defined in Solution 1) can be used. However, the perturbations so defined may not satisfy (49). To understand this, note that (49) is equivalent to saying the sum of the columns of $\hat{\mathbf{X}}_{mn}^{a}$ is zero, i.e., $\hat{\mathbf{X}}_{mn}^{a}\mathbf{v} = 0$ where $\mathbf{v}$ is a column vector of $k'+1$ ones. Since the sum of the background ensemble perturbations, which are defined by (8), is zero, we know that $\hat{\mathbf{X}}_{mn}^{b}\mathbf{v} = 0$. It then follows from (54) or (55) that $\big(\mathbf{Y}_{mn}\mathbf{Y}_{mn}^{T}\big)\mathbf{v} = \mathbf{v}$. If we choose the positive symmetric square root in (55), then $\mathbf{v}$ is also an eigenvector of $\mathbf{Y}_{mn}$ with eigenvalue one. Then by (51), we indeed have $\mathbf{X}_{mn}^{a}\mathbf{v} = 0$ as desired.

Construction of the global fields. Regardless of which of these solution methods is chosen, it now remains to construct an ensemble of global fields $\left\{\mathbf{x}^{a(i)}(\mathbf{r},t)\right\}$ that can be propagated forward in time to the next analysis time. For this purpose we set

$$\mathbf{x}^{a(i)}(\mathbf{r}_{mn},t) = \mathbf{J}\mathbf{x}^{a(i)}_{mn}, \tag{57}$$

where $\mathbf{J}$ maps the $(2l+1)^2u$ dimensional local vector to the u dimensional vector giving the atmospheric state at the grid point $\mathbf{r}_{mn}$ in the center of the patch, and we apply (57) globally; i.e., at all grid points (m,n) of the atmospheric model. Thus (57) defines an ensemble of $k'+1$ global analysis fields $\mathbf{x}^{a(i)}(\mathbf{r},t)$ for $i = 1,2,\ldots,k'+1$.

A more elaborate alternative to (57) for generating $\mathbf{x}^{a(i)}(\mathbf{r}_{mn},t)$ is to take into account all the atmospheric states at the point $\mathbf{r}_{mn}$ obtained from each of the $(2l+1)^2$ local vectors $\mathbf{x}^{a(i)}_{m-m',n-n'}$ $(|m'| \le l, |n'| \le l)$ that include the point $\mathbf{r}_{mn}$. For example, the u by u error covariance in each of these $(2l+1)^2$ states at $\mathbf{r}_{mn}$ can be estimated from the corresponding analysis error covariance matrices $\mathbf{P}^{a(i)}_{m-m',n-n'}$. Using this information, the most probable $\mathbf{x}^{a(i)}(\mathbf{r}_{mn},t)$ can then be obtained from a minimization procedure like that in Section 3.

5 Comments

1. The value of l is a parameter of the procedure and might be selected on the basis of which value yields the best forecasts. Also, in formulating our local vectors via (1) we have used a square patch from the grid. That is, we use the vectors $\mathbf{x}(\mathbf{r}_{m+m',n+n'},t)$ for $-l \le (m',n') \le l$. One could also employ other choices, for example, a circular patch with a given radius.

2. We have used the hypothesis that the forecast errors are Gaussian in formulating our procedure. This hypothesis is also inherent in current operational data assimilation schemes. In particular, these schemes use minimization of a quadratic form, such as (20), and this is strictly valid only for Gaussian distributions. An assumption of Gaussian forecast errors can only be justified if they evolve linearly; if they evolve nonlinearly, a Gaussian distribution at time t will evolve to a non-Gaussian distribution at time $t+\Delta t$. It should also be mentioned that the validity of the linearity assumption for the atmospheric model equations has been intensely investigated in the last decade. The most frequently cited papers on this subject (Vukicevic 1991; Errico et al. 1993) concluded that the linearized time evolution operator typically well approximates the evolution of the initial condition uncertainties up to 24-72 hours. The most recent estimates (Gilmour et al. 2001) are less optimistic, suggesting that the limit of linearity for current operational ensembles is about 12-24 hours, but considering that a typical background field is a 6-hour forecast (or an even shorter forecast; see Comment 7) the linearity assumption seems to be very reasonable. Moreover, we view as beneficial the restoration of the nonlinear dynamics via the nonlinear model integration of $\mathbf{x}^{a(i)}(\mathbf{r},t)$ to

obtain $\mathbf{x}^{b(i)}(\mathbf{r}, t+\Delta t)$. In particular, retention of the linear operator could lead to very fast nonphysical growth (to large amplitude) for some of the perturbations that might otherwise become saturated at low amplitude by the full nonlinear dynamics.

3. Since the minimization in (20) is done in localized regions, there is no need for distance dependent filtering to eliminate the influence of spurious long distance correlations.

4. (a) It may prove useful to enhance the probability of error in directions that formally show only very small error probability, e.g., by replacing $\mu_{mn}^{(i)}$ in (29) [$\lambda_{mn}^{(i)}$ in (19)] by $\max\left(\mu_{mn}^{(i)}, \varepsilon\right)$ [$\max\left(\lambda_{mn}^{(i)}, \varepsilon\right)$] where ε is a small positive constant. Other means of doing this are also possible (e.g., add ε to each eigenvalue). In any case, following such a small modification of $\mathbf{P}^a_{mn}$ [$\mathbf{P}^b_{mn}$], for consistency, one could also make small modifications to the ensemble perturbations $\delta\mathbf{x}_{mn}^{a(i)}$ [$\delta\mathbf{x}_{mn}^{b(i)}$] so as to preserve the relationship (32) [(7)]. (Again, similar to the discussion in Section 4, the choice of these small modifications is not unique. See the Appendix for a specific procedure for making the modified perturbations.) One reason why one might wish to employ such procedures is to account for model error. That is, our procedure has assumed that the model truly replicates the real atmospheric dynamics. If this was so, then a prediction of low error by our technique might be thought to be reliable. However, in the presence of model error, a prediction of very small error is probably unreliable, and ε might be thought of as a typical error due to the fact that the model is not perfect. Another unmodeled potential contributor to forecast error is subgrid scale dynamical fluctuations (e. g., small scale turbulence, butterflies, etc.) which may be regarded as acting in a way that is essentially similar to nondeterministic noise. Another reason for putting a floor on the eigenvalues is that it has the beneficial effect of preventing divergence of the method.

   One source of divergence could be collapse of the dimension of the space spanned by $\delta\hat{\mathbf{x}}_{mn}^{a(i)}$ [$\delta\hat{\mathbf{x}}_{mn}^{b(i)}$]. In particular, if an eigenvalue of the covariance matrix $\hat{\mathbf{P}}^a_{mn}$ [$\hat{\mathbf{P}}^b_{mn}$] becomes zero, then the ensemble perturbations $\delta\hat{\mathbf{x}}_{mn}^{a(i)}$ [$\delta\hat{\mathbf{x}}_{mn}^{b(i)}$] will no longer span a k-dimensional subspace. By putting a floor on the eigenvalues *and* making corresponding changes in the ensemble perturbations (e.g., see the Appendix) we expect that collapses of the dimension of the space spanned by the ensemble perturbations can be prevented. Another potential means for achieving this is to make small random changes to the global analysis fields $\mathbf{x}^{a(i)}(\mathbf{r}, t)$ at each grid point location. This method, although introducing a small nonsmooth field (the random perturbations), nevertheless, offers advantages in terms of simplicity and ease of implementation.

   (b) Another possible modification of the method is to weight the influence of the background and observations, giving decreasing weight to points

within the local patch that are further from the center of the patch. In this way one could, for example, gradually taper the influence of points to zero at the edge of the local patch (whose size might be correspondingly increased). This is in contrast with the scheme as stated in sections 2-4, where the influence of observational and background information drops abruptly to zero at the boundary of the patch.

5. Most operational weather prediction centers use three-dimensional variational data assimilation (3D-Var) schemes, in which the background covariance matrix is static, homogeneous, and isotropic, while the minimization of the quadratic form is done for global fields. These schemes use an elliptic *balance equation* to define the covariance between the hydrodynamical state variables (vorticity, divergence, temperature, surface pressure) in the background covariance matrix. Although this approach is necessary to filter unrealistic gravity wave components from the analyses and the ensuing forecasts, the true atmosphere (and a state-of-the-art model based on the atmospheric primitive equations) may deviate realistically from a balanced state. One major advantage of our technique (and, in general, of the Ensemble Kalman Filter approach) is that, instead of a prescribed balance equation, the nonlinear evolution of the ensemble defines the covariance between the different state variables.

6. Since the analysis at the different locations is prepared independently in our technique, it could potentially be efficiently implemented on the massively parallel computers currently used by operational prediction centers.

7. Operational data assimilation systems assimilate observed data from a time window, defined by the time interval $(t - \Delta t/2, t + \Delta t/2)$. For the global forecasting systems, the usual value of Δt is 6 hours, reflecting the fact that traditional components of the global observing network (e.g., radiosonde and surface observing stations) take observations at regular times separated by 6-12 hours intervals (0000 UTC, 0600 UTC, 1200 UTC, and/or 1800 UTC). The handling of observations taken at other times is a major challenge for the 3D-Var schemes. This is an important problem since the number of these (asynoptic time) observations is dramatically growing. Typical examples are remotely sensed (satellite and radar) observations, reports by commercial aircraft and ships, and adaptive observations collected during operational winter storm and hurricane reconnaissance programs. The accurate handling of these data will require the preparation of analyses more frequently (e.g., 24 times a day) than in current operational practice (4 times a day). We expect that the computational efficiency of our technique would allow this even in an operational environment. Moreover, by performing analyses more frequently the need for a technique like 4D-Var (see Comment 8) is obviated.

8. Another direction of development is *four-dimensional variational data assimilation* (4D-Var, Lions 1971; Talagrand 1981), a technique already

implemented operationally at the European Centre for Medium Range Weather Forecasts (ECMWF; Rabier et al. 2000; Mahfouf et al., 2000; Klinker et al. 2000) and Meteo France. In 4D-Var data assimilation the analysis is chosen so that the trajectory started from the analysis closely fits the observed data not only at the analysis time but also over an extended (6-12-hour) period. In other words, this technique fits a phase trajectory instead of a phase point to the observations. Early formulations of the 4D-Var problem assumed that the linearized dynamical equation would impose a strong constraint on the analysis, thus making it less sensitive to the particular choice of the background error covariance matrix (e.g., Daley 1991). However, experience accumulated during the last decade shows that obtaining a reliable estimate of the background error covariance matrix is no less important in 4D-Var than in other data assimilation techniques.

9. Note that not all matrix square root definitions yield global continuity. One particular important mechanism for non-global-continuity of matrix square roots is that the eigenvectors of a globally continuous, symmetric, non-negative matrix, $\mathbf{M}(\mathbf{r})$, may not be definable in a globally continuous manner. In particular, for smooth variation of $\mathbf{M}(\mathbf{r})$ in two dimensions, it can be shown that there will generically be isolated points in space where two of the eigenvalues of $\mathbf{M}(\mathbf{r})$ are equal. Following previous terminology in the field of quantum chaos (e.g., Ott 2002), we call such points "diabolical points" (e.g., Berry 1983). Assume that two eigenvalues of $\mathbf{M}(\mathbf{r})$ denoted $\xi_1(\mathbf{r})$ and $\xi_2(\mathbf{r})$, are equal at the diabolical point $\mathbf{r} = \mathbf{r}_d$, and denote their associated orthonormal eigenvectors by $\mathbf{v}_1(\mathbf{r})$ and $\mathbf{v}_2(\mathbf{r})$. Now consider starting at a point $\mathbf{r}_o \neq \mathbf{r}_d$ and following a continuous path C that encircles $\mathbf{r}_d$ and return to $\mathbf{r}_o$. Then it can be shown that, with continuous variation of $\mathbf{v}_1(\mathbf{r})$ and $\mathbf{v}_2(\mathbf{r})$ along the path, their directions are flipped by 180° upon return to $\mathbf{r}_o$. This presents no contradiction, since orthonormal eigenvectors are arbitrary up to within a change of sign, but it shows that $\mathbf{v}_1(\mathbf{r})$ and $\mathbf{v}_2(\mathbf{r})$ cannot be defined in a globally continuous manner. The positive symmetric square root $\big[\mathbf{M}(\mathbf{r})\big]^{1/2}$,

$$\big[\mathbf{M}(\mathbf{r})\big]^{1/2} = \sum_j \xi_j^{1/2}(\mathbf{r})\mathbf{v}_j(\mathbf{r})\mathbf{v}_j^T(\mathbf{r}),$$

is globally continuous because $\mathbf{v}_j(\mathbf{r})\mathbf{v}_j^T(\mathbf{r})$ returns to itself upon circuit around a diabolical point, even though $\mathbf{v}_j(\mathbf{r})$ may flip by 180°. Thus the solution for $\mathbf{Z}_{mn}$ given in (39)–(41) will be to be globally continuous, since the positive symmetric square root is used. The Cholesky square root will also yield global continuity. On the other hand, as an example of one of the choices that is unsatisfactory, the matrix square root choice,

$$\sqrt{\mathbf{M}(\mathbf{r})} = \big[\mathbf{v}_1(\mathbf{r}) \mid \mathbf{v}_2(\mathbf{r}) \mid \cdots\big]^T \mathrm{diag}\big(\xi_1^{1/2}(\mathbf{r}), \xi_2^{1/2}(\mathbf{r}), \cdots\big),$$

is clearly not globally continuous if diabolical points are present.

Acknowledgments
This work was supported by the W. M. Keck Foundation, by the Office of Naval Research (Physics), and by the National Science Foundation (Grant #0104087).

Appendix: An implementation of comment 4(a)

In comment 4(a) of section 5 we considered the modification of $\mathbf{P}^a_{mn}$ or $\mathbf{P}^b_{mn}$ to prevent the occurrence of small eigenvalues in these matrices. Furthermore, we considered the possibility of an accompanying modification of the corresponding ensemble perturbations, so as to preserve the relation

$$\hat{\mathbf{P}}_{mn} = \frac{1}{k'} \sum_{i=1}^{k'+1} \delta\hat{\mathbf{x}}^{(i)}_{mn} \left(\delta\hat{\mathbf{x}}^{(i)}_{mn}\right)^T. \tag{58}$$

In the above equation

$$\delta\hat{\mathbf{x}}^{i}_{mn} = \mathbf{Q}^T_{mn} \delta\mathbf{x}^{i}_{mn}, \tag{59}$$

and we have suppressed the superscript a or b with the understanding that (58) and (59) can apply to either the analysis or background.

We consider the example where $\hat{\mathbf{P}}_{mn}$ is changed to a new covariance matrix by addition of a small perturbation in the form,

$$\hat{\mathbf{P}}^*_{mn} = \hat{\mathbf{P}}_{mn} + \frac{\varepsilon\Lambda}{k}\mathbf{I}_k, \qquad \varepsilon > 0, \tag{60}$$

where $\mathbf{I}_k$ denotes the $k \times k$ unit matrix, and Λ is the trace of $\hat{\mathbf{P}}_{mn}$; i.e., it is the sum of its eigenvalues, and thus represents the total variance of the ensemble. Hence (60) increases the total variance by the factor $(1 + \varepsilon)$, where we regard ε as small, $1 \gg \varepsilon > 0$. More importantly, for small ε, the additional variance represented in (60) results in a relatively small change in the largest eigenvalues of $\hat{\mathbf{P}}_{mn}$, but prevents any eigenvalue from dropping below $\varepsilon\Lambda/k$, thus effectively providing a floor on the variance in any eigendirection. Having modified $\hat{\mathbf{P}}_{mn}$ to $\hat{\mathbf{P}}^*_{mn}$ via (60), we now consider the modification of the ensemble perturbations, $\delta\hat{\mathbf{x}}^{(i)}_{mn}$, to another set of ensemble perturbations, $\delta\hat{\mathbf{x}}^{(i)*}_{mn}$, with the perturbed covariance,

$$\hat{\mathbf{P}}^*_{mn} = \frac{1}{k'} \sum_{i=1}^{k'+1} \delta\hat{\mathbf{x}}^{(i)*}_{mn} \left(\delta\hat{\mathbf{x}}^{(i)*}_{mn}\right)^T. \tag{61}$$

From (58), (59) and (61) we have

$$\frac{1}{k'} \sum_{i=1}^{k'+1} \delta\hat{\mathbf{x}}^{(i)*}_{mn} \left(\delta\hat{\mathbf{x}}^{(i)*}_{mn}\right)^T = \frac{1}{k'} \sum_{i=1}^{k'+1} \delta\hat{\mathbf{x}}^{(i)}_{mn} \left(\delta\hat{\mathbf{x}}^{(i)}_{mn}\right)^T + \frac{\varepsilon\Lambda}{k}\mathbf{I}_k. \tag{62}$$

Letting

$$\delta\hat{\mathbf{x}}^{(i)*}_{mn} = \delta\hat{\mathbf{x}}^{(i)}_{mn} + \varepsilon\Lambda\delta\hat{\mathbf{z}}^{(i)}_{mn}, \tag{63}$$

our task is now to find sensible ensemble modifications $\delta\hat{\mathbf{z}}^{(i)}_{mn}$ that satisfy (62). Substituting (63) in (62), we obtain

$$\sum_{i=1}^{k'+1} \left[\delta\hat{\mathbf{x}}^{(i)}_{mn}\left(\delta\hat{\mathbf{z}}^{(i)}_{mn}\right)^T + \delta\hat{\mathbf{z}}^{(i)}_{mn}\left(\delta\hat{\mathbf{x}}^{(i)}_{mn}\right)^T + \varepsilon\Lambda\delta\hat{\mathbf{z}}^{(i)}_{mn}\left(\delta\hat{\mathbf{z}}^{(i)}_{mn}\right)^T\right] = \frac{k'}{k}\mathbf{I}_k. \tag{64}$$

The above represents $(k^2+k)/2$ linear scalar equations in the $k(k'+1)$ scalars making up the small ensemble modification $\delta\hat{\mathbf{z}}^{(i)}_{mn}$. Thus the system (64) is under-determined. To make the choice for the modifications $\delta\hat{\mathbf{z}}^{(i)}_{mn}$ unique, we recall that we think of the the background ensemble members as being similar to physical fields and that we have attempted to construct our analysis ensemble so as to preserve the physical nature (e.g., continuity) of the background as much as possible. Thus we desire to make the ensemble modifications as small as possible, and to thereby not destroy any favorable physical properties of the derived background or analysis ensemble members. Thus we shall seek the solution of (64) that minimizes

$$\sum_{i=1}^{k'+1} \left(\delta\hat{\mathbf{z}}^{(i)}_{mn}\right)^T \delta\hat{\mathbf{z}}^{(i)}_{mn}. \tag{65}$$

We solve this problem by introducing the orthonormal eigenvectors of $\hat{\mathbf{P}}_{mn}$, which we denote $\mathbf{w}^{(j)}_{mn}$, and then doing the constrained minimization using Lagrange multipliers. (In terms of our previous notation, $\mathbf{w}^{(j)}_{mn} = \mathbf{u}^{(j)}_{mn}$ for the background, and $\mathbf{w}^{(j)}_{mn} = \mathbf{v}^{(j)}_{mn}$ for the analysis.) The result for $\delta\hat{\mathbf{z}}^{(i)}_{mn}$ is

$$\delta\hat{\mathbf{z}}^{(i)}_{mn} = \left[\sum_{j=1}^{k} \xi^{(j)}_{mn}\mathbf{w}^{(j)}_{mn}\left(\mathbf{w}^{(j)}_{mn}\right)^T\right]\delta\hat{\mathbf{x}}^{(i)}_{mn}, \tag{66}$$

where

$$\xi^{(j)}_{mn} = \frac{\sqrt{1+\varepsilon\Lambda/k\eta^{(j)}_{mn}}-1}{\varepsilon\Lambda}, \tag{67}$$

and $\eta^{(j)}_{mn}$ is the eigenvalue of $\hat{\mathbf{P}}_{mn}$ corresponding to $\mathbf{w}^{(j)}_{mn}$; that is, $\hat{\mathbf{P}}^{(j)}_{mn}\mathbf{w}^{(j)}_{mn} = \eta^{(j)}_{mn}\mathbf{w}^{(j)}_{mn}$. Note that (66) preserves the necessary condition, $\sum_i \delta\hat{\mathbf{z}}^{(i)}_{mn} = 0$, satisfied by the original ensemble perturbations $\delta\hat{\mathbf{x}}^{(i)}_{mn}$.

References

Anderson, J. L., 2001: An ensemble adjustment filter for data assimilation. *Mon. Wea. Rev.,* **129,** 2884-2903.

Anderson, J. L., and S. L. Anderson, 1999: A Monte Carlo implementation of the nonlinear filtering problem to produce ensemble assimilations and forecasts. *Mon. Wea. Rev.,* **127,** 2741-2758.

Andrews, A., 1968: A square root formulation of the Kalman covariance equations. *AIAA J.,* **6,** 1165-1166.

Bergthorsson, P., B. Doos, 1955: Numerical weather map analysis. *Tellus,* **7,** 329-340.

Berry, M. V., 1983: Semiclassical mechanics of regular and irregular motion. In *Chaotic Behavior of Deterministic Systems,* R. H. G. Helleman and G. Ioos, eds., North-Holland, Amsterdam.

Bierman, G. J., 1977: *Factorization Methods for Discrete Sequential Estimation.* Mathematics in Science and Engineering, Academic Press, New York.

Bishop, C. H., B. J. Etherton, and S. Majumdar, 2001: Adaptive sampling with the Ensemble Transform Kalman Filter. Part I: Theoretical aspects. *Mon. Wea. Rev.,* **129,** 420-436.

Buehner, M., 2001: Improving the 3D-Var forecast error statistics: Ensemble statistics from the EFS and elements of a simplified Kalman Filter based on a variational approach. (manuscript)

Cressman, G., 1959: An operational objective analysis system. *Mon. Wea. Rev.,* **88,** 327-342.

Daley, R., 1991: *Atmospheric data analysis.* Cambridge University Press, New York.

Dee, D., S. Cohn, A. Dalcher, and M. Ghil, 1985: An efficient algorithm for estimating noise covariances in distributed systems. *IEEE Trans. Automatic Control,* **30,** 1057-1065.

Errico, R. M., T. Vukicevic, and K. Raeder, 1993: Examination of the accuracy of a tangent-linear-model. *Tellus,* **45A,** 462-477.

Evensen, G., 1994: Sequential data assimilation with a nonlinear quasi-geostrophic model using Monte Carlo methods to forecast error statistics. *J. Geophys. Res.,* **99**(C5), 10 143-10 162.

Evensen, G., and P. J. van Leeuwen, 1996: Assimilation of Geosat altimeter data for the Agulhas current using the ensemble Kalman Filter with a quasi-geostrophic model. *Mon. Wea. Rev.,* **124,** 85-96.

Fisher, M., 1998: Development of a simplified Kalman Filter. ECMWF Research Department Tech. Memo. 260, 16 pp. [Available from European Centre for Medium-Range Weather Forecasts, Shinfield Park, Reading, Berkshire, RG2 9AX, United Kingdom.]

Gandin, L., 1963: *Objective analysis of meteorological fields.* Gidromet, Leningrad; English translation: Israel Program for Scientific Translation, Jerusalem,

1965.

Ghil, M., S. Cohn, J. Tavantzis, K. Bube, and E. Isaacson, 1981: Applications of estimation theory to numerical weather prediction. In *Dynamic meteorology: data assimilation methods,* L. Bengtsson, M. Ghil, and E. Kallen, eds. Springer-Verlag, New York, 139-224.

Gilmour, I., L. A. Smith, R. Buizza, 2001: Linear regime duration: Is 24 hours a long time in synoptic weather forecasting? *J. Atmos. Sci.,* **58,** 3525-3539.

Hamill, T. M., and C. Snyder, 2000: A hybrid ensemble Kalman Filter-3D variational analysis scheme. *Mon. Wea. Rev.,* **128,** 2905-2919.

Hamill, T. M., J. Whitaker, and C. Snyder, 2001: Distance-dependent filtering of background error covariance estimates in an Ensemble Kalman Filter. *Mon. Wea. Rev.,* **129,** 2776-2790.

Houtekamer, P. L., and H. L. Mitchell, 2001: A sequential ensemble Kalman Filter for atmospheric data assimilation. *Mon. Wea. Rev.,* **129,** 796-811.

Houtekamer, P. L., and H. L. Mitchell, 1998: Data assimilation using an ensemble Kalman Filter technique. *Mon. Wea. Rev.,* **126,** 796-811.

Ide, K., P. Courtier, M. Ghil, and A. C. Lorenc, 1997: Unified notation for data assimilation: Operational, sequential, and variational. *J. Meteor. Soc. Japan,* **75**(1B), 181-189.

Jones, R., 1965: An experiment in nonlinear prediction. *J. Appl. Meteor.,* **4,** 701-705.

Kalman, R., 1960: A new approach to linear filtering and prediction problems. *Trans. ASME, Ser. D, J. Basic Eng.,* **82,** 35-45.

Kalman, R., R. Bucy, 1961: New results in linear filtering and prediction theory. *Trans. ASME, Ser. D, J. Basic Eng.,* **83,** 95-108.

Kalnay, E., 2002: *Atmospheric modeling, data assimilation, and predictability.* Cambridge University Press (in press).

Kalnay, E., and Z. Toth, 1994: Removing growing errors in the analysis. Preprints, *10th AMS Conference on Numerical Weather Prediction,* Portland, OR, 212-215.

Klinker, E., F. Rabier, G. Kelly, and J.-F. Mahfouf, 2000: The ECMWF operational implementation of four-dimensional variational assimilation. III: Experimental results and diagnostics with operational configuration. *Quart. J. Roy. Meteor. Soc.,* **126,** 1191.

Lions, J., 1971: *Optimal control of systems governed by partial differential equations.* Springer-Verlag, Berlin.

Lorenc, A., 1986: Analysis methods for numerical weather prediction. *Quart. J. Roy. Meteor. Soc.,* **112,** 1177-1194.

Mahfouf, J.-F., and F. Rabier, 2000: The ECMWF operational implementation of four-dimensional variational assimilation. II: Experimental results with improved physics. *Quart. J. Roy. Meteor. Soc.,* **126,** 1171-1190.

Molteni, F., R. Buizza, T. N. Palmer, and T. Petroliagis, 1996: The ECMWF ensemble prediction system: Methodology and validation. *Quart. J. Roy. Meteor. Soc.,* **122,** 73-119.

Ott, E., 2002: *Chaos in dynamical systems,* (second edition). Cambridge University Press, Cambridge, chapter 11.

Patil, D. J., I. Szunyogh, B. R. Hunt, E. Kalnay, J. A. Yorke, and E. Ott, 2002: Identification of local low dimensionality of atmospheric dynamics. To be submitted to Tellus.

Patil, D. J., B. R. Hunt, E. Kalnay, J. A. Yorke, and E. Ott, 2001: Local low dimensionality of atmospheric dynamics. *Phys. Rev. Lett.,* **86,** 5878-5881.

Petersen, D., 1973: A comparison of the performance of quasi-optimal and conventional objective analysis schemes. *J. Appl. Meteor.,* **12,** 1093-1101.

Rabier, F., H. Jarvinen, E. Klinker, J.-F. Mahfouf, and A. Simmons, 2000: The ECMWF operational implementation of four-dimensional variational assimilation. I: Experimental results with simplified physics. *Quart. J. Roy. Meteor. Soc.,* **126,** 1143-1170.

Talagrand, O., 1981: On the mathematics of data assimilation. *Tellus,* **33,** 43-60.

Tippett, M. K., J. L. Anderson, C. H. Bishop, T. M. Hamill, and J. S. Whitaker, 2002: Ensemble square-root filters. (under review).

Toth, Z., and E. Kalnay, 1997: Ensemble forecasting at NCEP and the breeding method. *Mon. Wea. Rev.,* **127,** 1374-1377.

Toth, Z., and E. Kalnay, 1993: Ensemble forecasting at NMC: The generation of perturbations. *Bull. Amer. Meteorol. Soc.,* **74,** 2317-2330.

Vukicevic, T., 1991: Nonlinear and linear evolution of initial forecast errors. *Mon. Wea. Rev.,* **119,** 1602-1611.

Wang, X., and C. H. Bishop, 2002: A comparison of breeding and Ensemble Transform Kalman Filter ensemble forecast schemes. Preprints, *Symposium on Observations, Data Assimilation, and Probabilistic Prediction*, Orlando, FL., Amer. Meteor. Soc., J28-J31.

Whitaker, J. S., and T. H. Hamill, 2002: Ensemble Data Assimilation without

perturbed observations. *Mon. Wea. Rev.,* **130,** 1913-1924.