6.8

ENSEMBLE DATA ASSIMILATION WITHOUT PERTURBED OBSERVATIONS

Jeffrey S. Whitaker and Thomas M. Hamill

NOAA-CIRES Climate Diagnostics Center, Boulder, Colorado

1. INTRODUCTION

The ensemble Kalman filter (EnKF), introduced by Evensen (1994), is a Monte-Carlo approximation to the traditional extended Kalman filter (EKF, Cohn 1997). Rather than propagating error covariances using the tangent linear and adjoint of the forecast operator, as in the EKF, the EnKF uses ensemble of forecasts to estimate error covariances, which is much less computationally intensive. By providing flow- and location-dependent estimates of first-guess forecast error, the EnKF can potentially provide analyses and forecasts that are much more accurate than current operational data assimilation schemes which assume that the background error does not vary in time.

It was recognized early in the development of the EnKF for atmospheric data assimilation that in order to maintain sufficient spread in the ensemble and to prevent filter divergence, the observations should be treated as random variables. Therefore, Houtekamer and Mitchell (1998) proposed using perturbed sets of observations to update each ensemble member. The perturbations were generated to be consistent with the error statistics of the observations, which are assumed to be known in any atmospheric data assimilation procedure. Burgers et al. (1998) provided a theoretical justification for perturbing observations and showed that if the observations are not treated as random variables, the ensemble analysis covariances will be systematically underestimated.

In this study we show that there is a source of error arising from using perturbed observations with small sample size, namely, that noise added to generate perturbed observations can be spuriously correlated with the background errors. As discussed in our companion manuscript (Whitaker and Hamill 2001, herafter WH01), the consequences of this error source can be severe when observations are processed serially, leading to systematic underestimation of the analysis error variance. To remove this source of error, we re-examine the formulation of the EnKF and present a modification to currently used algorithms that obviates the need to add random noise to the observations. We note that others have proposed similar but more complex ensemble assimilation algorithms that do not involve perturbing the observations (Lermusiaux and Robinson 1999, Anderson 2001).

Corresponding author address: Dr. Jeffrey S. Whitaker, NOAA-CIRES CDC, R/CDC 1, 325 Broadway, Boulder, CO 80303-3328. jsw@cdc.noaa.gov

Given the need for brevity, much of the detail and documentation of the testing in more complex models is omitted here. Please see WH01 for this additional material.

2. BACKGROUND

a. Ensemble Kalman filter equations

Following the notation of Ide et al. (1997), let $\mathbf{x}^b$ be a background model forecast, $\mathbf{y}^o$ be a set of observations, $\mathbf{H}$ be an operator that converts the model state to the observation space, $\mathbf{P}^b$ be the background error covariance matrix, and $\mathbf{R}$ be the observational error covariance matrix. The minimum error variance estimate of the analyzed state $\mathbf{x}^a$ is then given by the traditional Kalman filter update equation (Lorenc 1986),

$$\mathbf{x}^a = \mathbf{x}^b + \mathbf{K}(\mathbf{y}^o - \mathbf{H}\mathbf{x}^b), \tag{1}$$

where

$$\mathbf{K} = \mathbf{P}^b\mathbf{H}^T(\mathbf{H}\mathbf{P}^b\mathbf{H}^T + \mathbf{R})^{-1}. \tag{2}$$

In the EnKF, $\mathbf{P}^b$ is approximated using the sample covariance from an ensemble of model forecasts. Expressing the variables as an ensemble mean (denoted by an overbar) and a deviation from the mean (denoted by a prime), the update equations may be written as

$$\overline{\mathbf{x}}^a = \overline{\mathbf{x}}^b + \mathbf{K}(\overline{\mathbf{y}}^o - \mathbf{H}\mathbf{x}^b), \tag{3}$$

$$\mathbf{x}'^a = \mathbf{x}'^b + \widetilde{\mathbf{K}}(\mathbf{y}'^o - \mathbf{H}\mathbf{x}'^b), \tag{4}$$

where $\mathbf{P}^b = \overline{\mathbf{x}'^b\mathbf{x}'^{bT}}$, $\mathbf{K}$ is the traditional Kalman gain given by Eq. (2), and $\widetilde{\mathbf{K}}$ is the gain used to update deviations from the ensemble mean. In the EnKF framework, there is no need to compute and store the full matrix $\mathbf{P}^b$. Instead, $\mathbf{P}^b\mathbf{H}^T$ and $\mathbf{H}\mathbf{P}^b\mathbf{H}^T$ are estimated directly using the ensemble (Evensen 1994, Houtekamer and Mitchell 1998).

In the traditional Kalman filter, the analysis error covariance $\mathbf{P}^a$ is reduced from the background amount $\mathbf{P}^b$ by the assimilation of observations:

$$\mathbf{P}^a = (\mathbf{I}-\mathbf{K}\mathbf{H})\mathbf{P}^b(\mathbf{I}-\mathbf{K}\mathbf{H})^T+\mathbf{K}\mathbf{R}\mathbf{K}^T = (\mathbf{I}-\mathbf{K}\mathbf{H})\mathbf{P}^b. \tag{5}$$

However, in the EnKF, if all members are updated with the same observations ($\mathbf{y}'^o = \mathbf{0}$) using the same gain ($\mathbf{K} = \widetilde{\mathbf{K}}$), the covariance of the analyzed ensemble can be shown to be

$$\mathbf{P}^a = (\mathbf{I} - \mathbf{K}\mathbf{H})\mathbf{P}^b(\mathbf{I} - \mathbf{K}\mathbf{H})^T \tag{6}$$

Burgers et al. 1998). The missing term $\mathbf{KRK}^{\mathrm{T}}$ causes $\mathbf{P}^{\mathrm{a}}$ to be systematically underestimated. If random noise is added to the observations so that $\mathbf{y}'^{\mathrm{o}} \neq \mathbf{0}$, the analyzed ensemble variance is

$$\begin{aligned}\mathbf{P}^{\mathrm{a}} = (\mathbf{I}-\mathbf{KH})\mathbf{P}^{\mathrm{b}}(\mathbf{I}-\mathbf{KH})^{\mathrm{T}} + \\ \mathbf{K}(\overline{\mathbf{y}'^{\mathrm{o}}\mathbf{y}'^{\mathrm{o}\mathrm{T}}} - \mathbf{H}\overline{\mathbf{x}'^{\mathrm{b}}\mathbf{y}'^{\mathrm{o}\mathrm{T}}} - \overline{\mathbf{y}'^{\mathrm{o}}\mathbf{x}'^{\mathrm{b}\mathrm{T}}}\mathbf{H}^{\mathrm{T}})\mathbf{K}^{\mathrm{T}} + \\ \overline{\mathbf{x}'^{\mathrm{b}}\mathbf{y}'^{\mathrm{o}\mathrm{T}}}\mathbf{K}^{\mathrm{T}} + \mathbf{K}\overline{\mathbf{y}'^{\mathrm{o}}\mathbf{x}'^{\mathrm{b}\mathrm{T}}}\end{aligned}. \qquad (7)$$

If the observation noise is defined such that $\langle\overline{\mathbf{y}'^{\mathrm{o}}\mathbf{y}'^{\mathrm{o}\mathrm{T}}}\rangle = \mathbf{R}$ (where the brackets denote the expected value), then the expected value of $\mathbf{P}^{\mathrm{a}}$ is equal to that traditional Kalman filter result (Eq. 5), since the expected value of the background-observation error covariance $\langle\overline{\mathbf{x}'^{\mathrm{b}}\mathbf{y}'^{\mathrm{o}\mathrm{T}}}\rangle$ is zero (Burgers et al. 1998). However, for a finite ensemble, observational error and the background-observation error covariances will differ from their expected values due to sampling error. Therefore, errors in the EnKF with perturbed observations may be associated with sample size limitations in the estimation of the background error covariance and the observational error covariance.

b. Single-step data assimilation example

We now demonstrate the problem of spurious correlations between the observation perturbations and the background errors by performing a single data assimilation step where error statistics are known exactly. Our toy model is a one-dimensional periodic domain 40 units wide, discretized into 40 grid points. Background-error variances are 1.0 at each grid point, and the functional dependence of the background-error covariance is as given by Eq. 4.10 in Gaspari and Cohn (1999). This is a function which is approximately Gaussian in shape but which decays monotonically to zero at a separation distance of 10 units. Observations are available at every grid point, and are obtained by adding Gaussian white noise with unit variance to the true state (which is taken to be zero without loss of generality).

Ensembles of first-guess fields are created by generating random fields consistent with the assumed background-error covariances. The sample analysis error covariance obtained from a single step of the EnKF with perturbed observations is compared with that computed from Eq. (5) using the sample $\mathbf{P}^{\mathrm{b}}$. Fig. 1 shows the the absolute error in the estimated analysis error covariance as a function of distance from the analysis point, averaged over all analysis points, for a case with 10 ensemble members and observations at every grid point. The noise added to the observations in the EnKF results in significant higher errors in the estimated analysis error covariance than would be expected from sampling errors in the background-error covariance alone. Using this one-dimensional example, WH01 show that when observations are processed serially in the EnKF, perturbing the observations has an even larger impact. This suggests that methods which do not require perturbed observations, and enforce Eq. (5) directly, should attain higher accuracy for the same ensemble size by removing the error source associated with spurious background-observation error covariances.

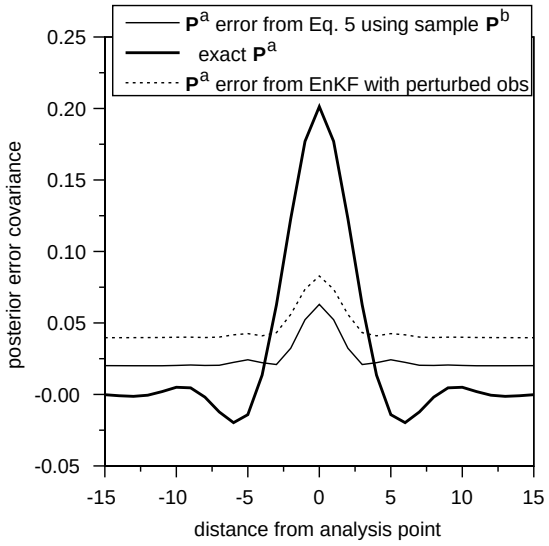

Figure 1. Absolute error in the analysis error covariance averaged over all analysis points, as a function of distance from the analysis point, for the EnKF (dotted), and computed with eq. (5) using the sample background-error covariance (light solid line). The thick solid line is the exact analysis error covariance computed from eq. (5) using the exact background-error covariance. A 10-member ensemble with observations at every grid point is used.

3. AN ENSEMBLE SQUARE ROOT FILTER.

We propose that instead of adding noise to the observations to obtain the correct $\mathbf{P}^{\mathrm{a}}$, $\widetilde{\mathbf{K}}$ be defined such that Eq. (5) is satisfied. Substituting $\widetilde{\mathbf{K}}$ into Eq. (6), and requiring that the resulting expression be equal to the "correct" $\mathbf{P}^{\mathrm{a}}$ (the right-hand side of Eq. (5)), yields an equation for $\widetilde{\mathbf{K}}$

$$\widetilde{\mathbf{K}}\mathbf{H}\mathbf{P}^{\mathrm{b}}\mathbf{H}^{\mathrm{T}}\widetilde{\mathbf{K}}^{\mathrm{T}} - \mathbf{P}^{\mathrm{b}}\mathbf{H}^{\mathrm{T}}\widetilde{\mathbf{K}}^{\mathrm{T}} - \widetilde{\mathbf{K}}\mathbf{H}\mathbf{P}^{\mathrm{b}} + \mathbf{K}\mathbf{H}\mathbf{P}^{\mathrm{b}} = \mathbf{0}. \qquad (8)$$

Letting $\mathbf{A} = \mathbf{I} - \widetilde{\mathbf{K}}\mathbf{H}$, the above may be written as

$$\mathbf{A}\mathbf{P}^{\mathrm{b}}\mathbf{A}^{\mathrm{T}} = \mathbf{P}^{\mathrm{b}} - \mathbf{K}\mathbf{H}\mathbf{P}^{\mathrm{b}}, \qquad (9)$$

which has a solution

$$\mathbf{A}\sqrt{\mathbf{P}^{\mathrm{b}}} = \sqrt{\mathbf{P}^{\mathrm{b}} - \mathbf{K}\mathbf{H}\mathbf{P}^{\mathrm{b}}}. \qquad (10)$$

Since the calculation of $\mathbf{A}$ involves the square root of the background-error covariance matrix, this is essentially a Monte-Carlo implementation of a square-root filter

(Maybeck 1979). For this reason, we call this algorithm the ensemble square root filter, or EnSRF.

When sequentially processing independent observations, $\mathbf{K}$, $\widetilde{\mathbf{K}}$, $\mathbf{HP}^{\mathrm{b}}$ and $\mathbf{P}^{\mathrm{b}}\mathbf{H}^{\mathrm{T}}$ are all vectors with the same length as the model state vector, and $\mathbf{HP}^{\mathrm{b}}\mathbf{H}^{\mathrm{T}}$ is a scalar. Thus, as first noted by Potter (1964), when observations are processed one at a time Eq. (8) becomes a scalar quadratic which can solved for each each element of the vector $\widetilde{\mathbf{K}}$ independently, yielding

$$\widetilde{\mathbf{K}} = \left(1 + \sqrt{\frac{\mathbf{R}}{\mathbf{HP}^{\mathrm{b}}\mathbf{H}^{\mathrm{T}} + \mathbf{R}}}\right)^{-1} \mathbf{K}. \qquad (11)$$

Here, $\mathbf{HP}^{\mathrm{b}}\mathbf{H}^{\mathrm{T}}$ and $\mathbf{R}$ are scalars representing the background and observational error variance at the observation location. The quantity multiplying $\mathbf{K}$ in Eq. (11) is a scalar between 0 and 1. This means that, in order to obtain the desired analysis error covariance, one must use a modified Kalman gain to update deviations from the ensemble mean that is reduced in magnitude relative to the traditional Kalman gain. Thus, deviations from the mean are reduced less in the analysis using $\widetilde{\mathbf{K}}$ than they would be using $\mathbf{K}$. In the EnKF, the excess variance reduction caused by using $\mathbf{K}$ to update deviations from the mean is compensated for by the introduction of noise to the observations. In the EnSRF, the mean and departures from the mean are updated independently according to Eqs. (2), (3), (4), and (11), with $\mathbf{y}'^{o} = \mathbf{0}$. If observations are processed one at a time, the EnSRF requires no more computation than the traditional EnKF with perturbed observations.

4. RESULTS WITH THE 40-VARIABLE LORENZ MODEL.

The model of Lorenz and Emanuel (1998) is governed by the equation

$$\frac{dX_i}{dt} = (X_{i+1} - X_{i-2})X_{i-1} - X_i + F, \qquad (12)$$

where $i = 1, ..., N$ with cyclic boundary conditions. Here we use $N = 40$, $F = 8$ and a fourth-order Runge-Kutta time integration scheme with a time step of 0.05 units. For this parameter setting, the leading Lyapunov exponent implies an error doubling time of about 8 time steps, and the fractal dimension of the attractor is about 27 (Lorenz and Emanuel 1998). For our assimilation experiments, each state variable is observed directly, and observations have uncorrelated errors with unit variance. A 10-member ensemble is used, an observations are assimilated every time step for 50,000 time steps (after a spin-up period of 1000 time steps). We use the following statistics to measure the relative performance of the EnKF and EnSRF. The time-averaged RMS error of the ensemble mean is denoted as (E_1), and E_2 denotes the time-averaged RMS error of each ensemble member. The RMS ratio $R = E_1/E_2$ is a measure of how similar the truth is to a randomly selected member of the ensemble (Anderson 2001). If the truth is statistically indistinguishable from any ensemble member then the expected value of $\langle R \rangle = \sqrt{(M+1)/2M}$, or approximately 0.74 for a 10-member ensemble (Murphy 1988). If the actual $R < \langle R \rangle$, there is too much spread in the ensemble, and if $R > \langle R \rangle$, there is not enough spread.

Sampling error in the EnKF can cause filter divergence, so some extra processing of the ensemble covariances may be necesary if the number of ensemble members is less than the number of degrees of freedom present in the system being analyzed. The two techniques used here are distance-dependent covariance filtering (Houtekamer and Mitchell 2001, Hamill et al. 2001) and covariance inflation (Anderson and Anderson 1999). Distance-dependent covariance filtering counters the tendency for ensemble variance to be excessively reduced by spurious long-range correlations between analysis and observations points by applying a filter that forces the ensemble covariances to go to zero at some distance L from the observation being assimilated. The function we use to perform the filtering is given by Eq. 4.10 in Gaspari and Cohn (1999).

Occasionally, sampling error will cause the background error variances to be underestimated, and the analysis system will weight the first-guess forecasts too heavily. Due to the nonlinear relationship between $\mathbf{K}$ and $\mathbf{P}^{\mathrm{b}}$, underestimation of the background error variances can have a relatively more severe impact on analysis error than overestimating the variances by the same amount. To compensate for this, covariance inflation simply inflates the deviations from the ensemble mean first-guess by a small constant factor r for each member of the ensemble, before the computation of the background error covariances and before any observations are assimilated. See Hamill and Whitaker (2001) for more details.

Fig. 2 shows E_1, averaged over 50,000 assimilation cycles, for the EnKF and EnSRF, as a function of the covariance inflation factor and the length scale of the covariance localization filter. The shaded areas on these plots indicate regions in parameter space where the filter has diverged, i.e., has drifted into a regime where it effectively ignores observations. For both the EnKF and EnSRF, filter divergence occurs for a given covariance filter length scale L when r is less than a critical value. However, the EnSRF appears to be less susceptible to filter divergence, since the critical value of the covariance inflation factor is always less for a given L. Overall, for almost any parameter value, the error in the EnSRF is less than the EnKF. The minimum error in the EnSRF is 0.16, which occurs at $L = 24$ for $r = 1.03$. For the EnKF, the minimum error is 0.21, which occurs at L=15 for $r = 1.08$. These results are consistent with those shown for the one-dimensional model in Fig. 1, which demonstrated that the extra terms in Eq. (7) involving background-observation

error covariances reflect increased sampling error for the same ensemble size when noise is added to the observations in the EnKF. This extra sampling error makes the EnKF more susceptible to filter divergence, and reduces the accuracy of the filter. Effectively, the signal/noise ratio present in the EnSRF ensemble is higher than in the EnKF ensemble, which means that the EnSRF is able to extract more useful information from the observations for a given ensemble size. This is consistent with the fact that larger values of L benefit the EnSRF, but are detrimental to the EnKF. For greatly separated observation and analysis grid points, there may only be a very small 'true' covariance, or signal. The EnSRF is able to extract and use of this signal, which is overwhelmed by the extra noise associated with the perturbed observations in the EnKF.

Fig. 3 shows the RMS ratio for the same set of experiments. For a 10-member ensemble, the expected value for an ensemble which faithfully represents the true underlying probability distribution is 0.74. For nearly all parameter settings in this model, the EnSRF has a lower RMS ratio than the EnKF, indicating that the variance in the EnKF ensemble is smaller relative to ensemble mean analysis error. This behavior is a consequence of the additional noise introduced into the background-error covariance estimate by the perturbed observations. Because of the extra sampling error associated with the perturbed observations, there is more temporal variability in the background error variances in the EnKF than in the EnSRF. The nonlinear relationship between the $\mathbf{K}$ and $\mathbf{P}^{b}$ causes underestimation of $\mathbf{P}^{b}$ to haves a relatively larger impact on $\mathbf{P}^{a}$ than overestimation of $\mathbf{P}^{b}$. Therefore, over many assimilation cycles, the net result of the extra temporal variability in the EnKF background error variance estimates is an under-weighting of the observations, and an analysis ensemble whose variance is smaller relative to ensemble mean error.

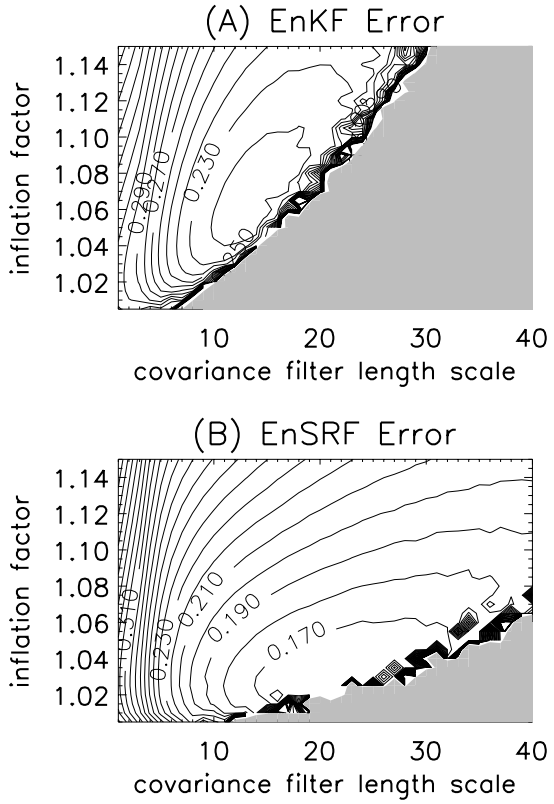

Figure 2. Ensemble mean error as a function of the distance at which the covariance filter goes to zero, and the covariance inflation factor, for the EnKF (A) and the EnSRF (B). Results are for a 10 member ensemble averaged over 50,000 assimilation cycles using the model of Lorenz and Emanuel (1998), with observations of every state variable. Observations have unit error variance. Shaded regions indicate regions in parameter space where the filter diverges.

5. CONCLUSIONS.

We have implemented an ensemble square-root filter, or EnSRF, based upon the algorithm of Potter (1964), which involves processing the observations serially. This implementation is attractive because, in addition to being algorithmically simple, it avoids the need to compute matrix square roots and thus is requires no more computation than the EnKF with perturbed observations and serial observation processing.

The benefits of ensemble data assimilation without perturbed observations have been demonstrated by comparing the EnKF and the EnSRF using the Lorenz and Emanuel (1998) model, and using an idealized primitive equation GCM in our companion paper (WH01). The EnSRF produces an analysis ensemble whose ensemble mean error is lower than the EnKF for the same ensemble size. The actual reduction in analysis error realized by implementing the EnSRF depends on the ratio of sampling error in the estimation of the background error covariance to sampling error in the estimation of the observational error covariance in the EnKF. Factors that can affect this ratio are the density of observations and the relative magnitudes of the background and observational error variance. We note that ampling error in the observational error covariance was relatively more important in the low-order model as compared to the global GCM experiments presented in WH01, so that the difference between the EnSRF and EnKF is larger in the low-order model. WH01 also describes how the errors associated with the noise added to the observations is relatively more severe when observations are processed serially in the EnKF.

The EnSRF as formulated here requires observations to be processed one at a time, which may pose quite a challenge in an operational setting where observations can number in the millions. It will be crucial to develop parallel algorithms (such as the one proposed by Houtekamer and Mitchell (2001)) to allow greatly separated observations to be processed independently. The treatment of model error, which we have not considered in this study, will likely be a crucial element in any future operational system. These will continue to be active

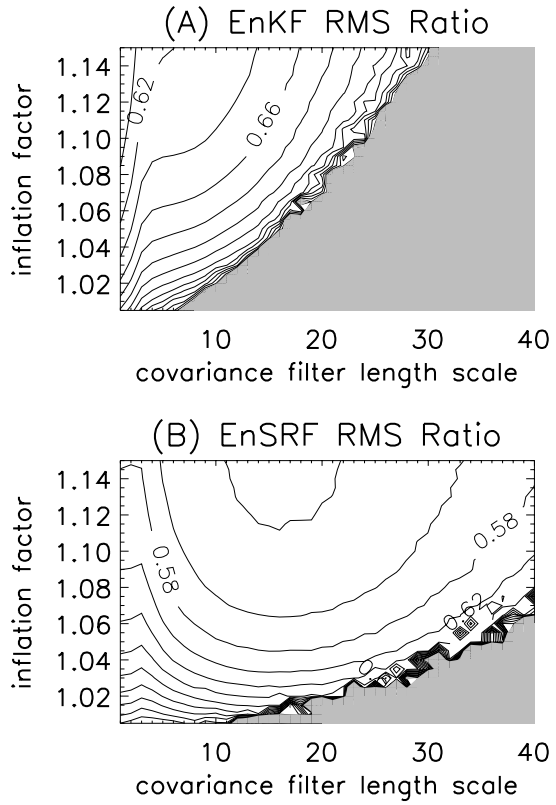

Figure 3. RMS ratio as a function of the distance at which the covariance filter goes to zero, and the covariance inflation factor, for the EnKF (A) and the EnSRF (B). Results are for a 10 member ensemble averaged over 50,000 assimilation cycles using the model of Lorenz and Emanuel (1998), with observations of every state variable. Observations have unit error variance. Shaded regions indicate regions in parameter space where the filter diverges.

areas of research as ensemble data assimilation methods are implemented in more complex and realistic systems.

REFERENCES

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

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.

Burgers, G., P. J. van Leeuwen, and G. Evensen, 1998: Analysis scheme in the ensemble Kalman filter. *Mon. Wea. Rev.*, **126**, 1719–1724.

Cohn, S. E., 1997: An introduction to estimation theory. *J. Meteor. Soc. Jap.*, **75(1B)**, 257-288.

Evensen, G., 1994: Sequential data assimilation with a nonlinear quasigeostrophic model using Monte Carlo methods to forecast error statistics. *J. Geophys. Res.*, **99 (C5)**, 10143–10162.

Gaspari, G. and S. E. Cohn, 1999: Construction of correlation functions in two and three dimensio ns. *Quart. J. Roy. Meteor. Soc.*, **125**, 723–757.

Hamill, T. M., J. S. Whitaker, and C. Snyder, 2001: Distance-dependent filtering of background error covariance estimates in an ensemble Kalman filter. *Mon. Wea. Rev.*, **129**, in press. Available at www.cdc.noaa.gov/∼hamill .

Held, I. M. and M. J. Suarez, 1994: A proposal for the intercomparison of the dynamical cores of atmospheric general circulation models. *Bull. Amer. Meteor. Soc.*, **75**, 1825–1830.

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

——, and ——, 2001: A sequential ensemble Kalman filter for atmospheric data assimilation. *Mon. Wea. Rev.*, **129**, 123–137.

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

Lermusiaux, P. F. J. and A. R. Robinson, 1999: Data assimilation via error subspace statistical estimation. Part I: Theory and schemes. *Mon. Wea. Rev.*, **127**, 1385–1407.

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

Lorenz, E. N. and K. A. Emanuel, 1998: Optimal sites for supplementary weather observations: Simulation with a small model. *J. Atmos. Sci.*, **55**, 399–414.

Maybeck, P. S., 1979: *Stochastic Models, Estimation and Control*, Academic Press, volume 1, chapter 7. 368–409.

Murphy, J. M., 1988: The impact of ensemble forecasts on predictability. *Quart. J. Roy. Meteor. Soc.*, **116**, 89–125.

Potter, J., 1964: W matrix augmentation. M.I.T. Instrumentation Laboratory Memo SGA 5-64, Massachusetts Institute of Technology, Cambridge, MA.

Whitaker, J. S., and T. M. Hamill, 2001: Ensemble data assimilation without perturbed observations. *Mon. Wea. Rev.*, **129**, submitted. Available from www.cdc.noaa.gov/∼jsw .