



## Controlling Noise in Ensemble Data Assimilation Schemes

MALAQUIAS PEÑA

*NOAA/NCEP/EMC, Camp Springs, Maryland*

ZOLTAN TOTH

*NOAA/ESRL/GSD, Boulder, Colorado*

MOZHENG WEI

*NOAA/NCEP/EMC, Camp Springs, Maryland*

(Manuscript received 6 November 2008, in final form 24 November 2009)

### ABSTRACT

A variety of ad hoc procedures have been developed to prevent filter divergence in ensemble-based data assimilation schemes. These procedures are necessary to reduce the impacts of sampling errors in the background error covariance matrix derived from a limited-size ensemble. The procedures amount to the introduction of additional noise into the assimilation process, possibly reducing the accuracy of the resulting analyses. The effects of this noise on analysis and forecast performance are investigated in a perfect model scenario. Alternative schemes aimed at controlling the unintended injection of noise are proposed and compared. Improved analysis and forecast accuracy is observed in schemes with minimal alteration to the evolving ensemble-based covariance structure.

### 1. Introduction

The ensemble Kalman filter (EnKF) technique introduced by Evensen (1994) has inspired numerous studies on the development of flow-dependent data assimilation schemes (Evensen 2003). The technique uses short-range ensemble forecasts to provide time- and space-dependent error structures, resulting in potentially more accurate representations of the background error covariance. A fundamental difficulty in applying ensemble data assimilation techniques to complex systems such as the atmosphere is that practical ensemble sizes are insufficient to accurately estimate the elements of the background error covariance matrix  $\mathbf{P}^f$ . The size of an ensemble is limited by the computational cost involved in running a model multiple times. For comparison, a typical ensemble size is usually less than  $10^2$ , whereas the degrees of freedom of

a model state of a typical global atmospheric model, for example, is around  $10^{10}$ . Although in some instances the structure of forecast errors can be represented in a subspace of much smaller dimension than that of the model state (Toth and Kalnay 1997; Patil et al. 2001), typically the ensemble is not sufficiently large. The resulting undersampling of the background error covariance produces spurious long-distance correlations, limiting the accuracy of the analysis and forecast fields and can also lead to filter divergence (Houtekamer and Mitchell 1998), requiring the introduction of ad hoc procedures to deal with the problems.

Approaches designed to prevent filter divergence in ensemble data assimilation schemes focus primarily on solving an ill-conditioned matrix problem. Evensen and van Leeuwen (1996) used a singular value decomposition of the sum of background and observational covariance matrices to discard contributions from eigenvectors corresponding to the least significant eigenvalues. Anderson and Anderson (1999) multiply the background error covariance matrix in each analysis cycle by a constant “inflation” factor to improve the analysis. Corazza et al. (2002) add random values to a background error covariance

---

*Corresponding author address:* Malaquias Peña, Science Applications International Corporation, National Centers for Environmental Prediction, Room 807, 5200 Auth Rd., Camp Springs, MD 20746.

E-mail: malaquias.pena.mendez@noaa.gov

matrix spanned by bred vectors before integrating the model forward. This additive noise is interpreted by Corazza et al. (2002) as a procedure to refresh the bred vectors and prevent collapsing into one dominant direction. Corazza et al. (2007), Whitaker et al. (2008), and Yang et al. (2009) combine additive noise with a localization treatment for better analysis performance. Hamill and Snyder (2000) and Wang et al. (2007) combine climatological and flow-dependent covariance matrices in a hybrid scheme. Another covariance treatment scheme called covariance relaxation introduced by Zhang et al. (2004) relaxes the analysis error covariance matrix  $\mathbf{P}^a$  partially to  $\mathbf{P}^f$  without adding additional “noise,” which has been demonstrated to be effective in many subsequent studies at the mesoscales including real-time real-data applications (e.g., Meng and Zhang 2008). The covariation relaxation performs well with small ensemble sizes (Zhang et al. 2006).

To control spurious long-distance correlations associated with sampling errors of the background error covariance, some techniques use localization procedures. Houtekamer and Mitchell (1998) use a cutoff radius beyond which covariances between variables are assumed to be zero. Houtekamer and Mitchell (2001) and Whitaker and Hamill (2002) and several other authors thereafter use an element-wise multiplication of the background error covariance matrix to reflect the correlation decaying with distance (Gaspari and Cohn 1999). Ott et al. (2004) introduced the local ensemble Kalman filter technique that divides up the global surface into local regions and performs ensemble Kalman filter assimilation in each of them independently.

Common in these methods is that while attempting to improve performance and control filter divergence, they introduce methods that intentionally (e.g., by injecting additive noise) or inadvertently (e.g., by the use of a cutoff radius for correlations or by the use of flow-independent components in the background covariance matrix) introduce noise (i.e., components that are dynamically inconsistent with the flow) into the background covariance matrix. Theoretically, the addition of noise into a representation of a dynamical system is expected to destroy part of the information about the state of the system and may degrade analysis and forecast performance.

The purpose of this study is to assess the extent to which noise introduced intentionally or inadvertently in efforts to avoid filter divergence due to the use of limited size ensembles may degrade the performance of assimilation schemes. The hypothesis is that, in a perfect model scenario, more accurate analysis and forecast fields are obtained by limiting the extent to which noise can affect the dynamically constructed background covariance information derived through the ensemble first-guess fields.

To test the hypothesis, two of the cited approaches (adding random perturbations and combining a static background error covariance with a flow-dependent covariance in a hybrid) will be contrasted with alternative methods proposed here to more closely control noise in the generation of ensemble forecasts.

The paper is organized as follows. Section 2 presents the model used, the general data assimilation procedure, and the experimental set up. Section 3 gives a description of the procedures used to estimate the background and analysis error covariance matrices. Section 4 describes the data assimilation techniques evaluated in the study. Section 5 presents the results, and section 6 provides a summary and concluding remarks.

## 2. Experimental design

### a. The model

This study is based on the Lorenz (1996) model, which is a one-dimensional system where a variable is defined at  $N$  equally spaced points around a latitude circle. With the parameters given in Lorenz (2006), the model exhibits regime changes similar to those observed in the barotropic vorticity equation models in an elongated rectangular domain. Specifically, the model’s governing equations are

$$dx_n/dt = -x_{n-2}x_{n-1} + x_{n-1}x_{n+1} - x_n + F, \quad (1)$$

where  $n = 1, 2, \dots, N$ ,  $N = 21$  is the dimension of the model state;  $F = 5.1$  is the forcing term; and  $t$  is time. The variations of  $x_n$  are intended to simulate the behavior of some atmospheric quantity. The definition of  $x_n$  is extended to all values of  $n$  by letting  $x_{n+N} = x_n$ , so that the variables form a cyclic chain. The model is integrated using the fourth-order Runge–Kutta numerical scheme with a step size of 0.05 corresponding to a 6-h interval.

### b. Data assimilation process

An outline for the algorithm to carry out ensemble data assimilation is given next. An ensemble of initial atmospheric states, the ensemble analysis,<sup>1</sup> is integrated forward 6 h to the next analysis cycle. This short-range ensemble forecast, referred to as the background field, is used to compute the ensemble mean forecast  $\bar{\mathbf{x}}^f$  and  $\mathbf{P}^f$  using the following definitions:

$$\bar{\mathbf{x}}^f = \frac{1}{K} \sum_{k=1}^K \mathbf{x}_k^f, \quad (2)$$

<sup>1</sup> The first analysis ensemble is generated from normally distributed random departures from the observations.

$$\mathbf{X}^f = [\mathbf{x}_1^f - \overline{\mathbf{x}}^f, \mathbf{x}_2^f - \overline{\mathbf{x}}^f, \dots, \mathbf{x}_k^f - \overline{\mathbf{x}}^f], \quad \text{and} \quad (3)$$

$$\mathbf{P}^f = \mathbf{X}^f \mathbf{X}^{fT} (K - 1)^{-1}, \quad (4)$$

where  $\{\mathbf{x}_k^f\}$ , ( $k = 1, \dots, K$ ) is the ensemble of forecast states;  $\mathbf{X}^f$  is the forecast perturbation matrix whose columns are the normalized deviations from the ensemble mean; and  $K$  is the ensemble size.

Once  $\mathbf{P}^f$  is computed, it is used in two ways: to compute the new analysis state  $\mathbf{x}^a$  via a three-dimensional variational data assimilation (3DVAR) algorithm, and to estimate the next analysis perturbation matrix  $\mathbf{X}^a$ , which in analogy with (4) defines  $\mathbf{P}^a$ :

$$\mathbf{P}^a = \mathbf{X}^a \mathbf{X}^{aT} (K - 1)^{-1}. \quad (5)$$

The ensemble data assimilation methods used in this study (and described in section 3) generate  $\mathbf{X}^a$  centered on  $\mathbf{x}^a$ . To make the comparison among methods independent on the analysis method, a 3DVAR methodology (e.g., Kalnay 2003) is used to obtain the analysis state. In this study, the biconjugated gradient method is used to obtain the analysis from the following equation:

$$(\mathbf{I} + \mathbf{P}^f \mathbf{H}^T \mathbf{R}^{-1} \mathbf{H})(\mathbf{x}^a - \overline{\mathbf{x}}^f) = \mathbf{P}^f \mathbf{H}^T \mathbf{R}^{-1} (\mathbf{y} - \mathbf{H}^T \overline{\mathbf{x}}^f), \quad (6)$$

where  $\mathbf{y}$  is the observation vector,  $\mathbf{I}$  is the identity matrix of dimension  $N \times N$ ,  $\mathbf{R}$  is the observation error covariance, and  $\mathbf{H}$  is the operator that maps the model state onto the observations. In this study,  $\mathbf{R} = \mathbf{I}$  and  $\mathbf{H} = \mathbf{I}$ . The method for estimating the next ensemble analysis from  $\mathbf{P}^f$  is presented in section 3.

### c. Perfect model scenario and experiments

To simplify the study, the experiments are carried out in a perfect model scenario. That is, the truth is generated from the same model that produces the forecasts. Thus, filter divergence in the assimilation procedures, if it occurs, will be unrelated to model errors and will be attributed to sampling errors resulting from the calculation of the covariance matrices.

The model is integrated forward 200 time steps to remove any transient behavior. After this period, the model is integrated further to produce what is referred to as the true state of the system or truth. At each time step and at each grid point, a synthetic observation is created by adding normally distributed random values of unit amplitude to the truth and a mean of zero.

The simplicity of the model allows for experimentation with different ensemble sizes to depict the sampling error and assess the robustness of the results. The ensemble sizes analyzed are 20, 15, 12, and 9 members. For

each data assimilation cycle, the ensemble forecast is initialized from  $\mathbf{X}^a$  centered on  $\mathbf{x}^a$  and is evaluated for lead times out to 72 h, using the mean absolute error (MAE) as a performance metric. Averaged performance results are shown for four batches of 480 forecasts initialized every data assimilation cycle (6 h) for each method. Because the truth is known in these experiments, the analysis and forecast errors can be determined exactly.

## 3. Techniques to estimate error covariance matrices

The ensemble Kalman filter technique uses the Kalman filter equations to relate  $\mathbf{P}^f$  and  $\mathbf{P}^a$  matrices. A class of these techniques based on the Kalman square root filter algorithm (Tippett et al. 2003) avoids forming the full covariance matrices and instead uses a relationship of  $\mathbf{X}^a$  and  $\mathbf{X}^f$ . In this case,  $\mathbf{X}^a$  becomes the initial conditions for the ensemble in the next analysis cycle. Several of these approaches are shown to be equivalent given the same assumptions (Tippett et al. 2003). In this study, the focus is on the ensemble transform Kalman filter (ETKF; Bishop et al. 2001; Wang and Bishop 2003; Wei et al. 2006) given that a closely related approach for ensemble generation, the ensemble transform method (ET; Bishop and Toth 1999), is being used operationally to generate perturbations for the National Centers for Environmental Prediction (NCEP) global ensemble system (Wei et al. 2008) and the U.S. Navy Operational Global Atmospheric Prediction System (NOGAPS; McLay et al. 2008). The ET method does not assimilate observations but instead uses an externally derived analysis to center the initial ensemble perturbations. Bishop et al. (2001) introduced ETKF as a generalization of the ET method to assimilate observations and estimate their effects on forecast error covariance. Both the ET and ETKF methods have also been used to estimate the effects of potential adaptively corrected data on high-impact forecasts (Szunyogh et al. 2000; Majumdar et al. 2001, 2002). More recent advances of the ETKF have been made by Hunt et al. (2007) in an approach called the local ensemble transform Kalman filter.

### a. The National Meteorological Center method

For comparison with the ensemble approaches to computing error covariance matrices, the National Meteorological Center (NMC, now known as NCEP) method (Parrish and Derber 1992) is included in this study. In this method, the background error covariance is obtained by using the time average of the difference between 24- and 48-h single forecasts verifying at the same time. For this study, 2 months of forecasts were used to calculate the time average. The magnitude of the resulting covariance

is appropriately scaled to produce the best analysis. The optimal scaling factor was sought from several independent periods of 2 months and using MAE of the analysis as the minimization norm. The factor selected was 0.05.

*b. An ensemble transform Kalman filter (ETKF)*

The ETKF technique converts forecast perturbations into analysis perturbations by a transformation matrix  $\mathbf{T}$ ; that is,

$$\mathbf{X}^a = \mathbf{X}^f \mathbf{T}. \tag{7}$$

Using (4), (5), and (7), and considering that  $\mathbf{P}^a$  satisfies

$$\mathbf{P}^a = \mathbf{P}^f - \mathbf{P}^f \mathbf{H}^T (\mathbf{H} \mathbf{P}^f \mathbf{H}^T + \mathbf{R})^{-1} \mathbf{H} \mathbf{P}^f, \tag{8}$$

the solution for  $\mathbf{T}$  is given (Bishop et al. 2001) as

$$\mathbf{T} = \mathbf{C} (\mathbf{\Gamma}_c + \mathbf{I})^{-1/2}, \tag{9}$$

where  $\mathbf{C}$  contains column orthogonal eigenvectors of  $\mathbf{X}^{fT} \mathbf{H}^T \mathbf{R}^{-1} \mathbf{H} \mathbf{X}^f$ , and  $\mathbf{\Gamma}_c$  is the diagonal matrix containing the corresponding eigenvalues. Since the ensemble mean is subtracted to obtain  $\mathbf{X}^f$ , by construction,  $\mathbf{P}^f$  will have  $K - 1$  independent columns. In large numerical systems,  $\mathbf{P}^f$  tends to be rank deficient. Note that  $\mathbf{R}$ , which is usually invertible, helps the sum in parenthesis in (8) to be invertible regardless on whether the inverse of  $\mathbf{P}^f$  exists. The direct application of the ETKF equations often leads to filter divergence after several assimilation cycles and a treatment is needed. Section 4 provides different types of approaches to prevent filter divergence.

*c. Hybrid ETKF–NMC*

Instead of adding random noise to the model integrations, it may be possible to avoid filter divergence by modifying directly the background error covariance matrix to eliminate the rank deficient problem in ETKF methods. Hamill and Snyder (2000) introduced the idea of enhancing the EnKF background error covariance with elements of a more stable covariance matrix. Their hybrid scheme linearly combines the background error statistics from a time-invariant 3DVAR covariance and an EnKF error covariance. In this study, the hybrid (HYB) is made by combining the control NMC method and the direct ETKF, that is, without any other treatment:

$$\mathbf{P}^{f-HYB} = \mathbf{P}^f + \alpha (\mathbf{P}^{f-NMC} - \mathbf{P}^f), \tag{10}$$

where the weighting parameter  $\alpha$  is optimized to minimize the analysis error. The time-invariant error covariance  $\mathbf{P}^{f-NMC}$  effectively removes the rank deficiency

problem of the flow-dependent EnKF error covariance. Here,  $\mathbf{P}^{f-NMC}$  is computed from sufficiently long sampling data, and it is basically a band-diagonal matrix. If one considers a rank-deficient  $\mathbf{P}^f$  from an EnKF, the second term in (10) fills in the gap in the band-diagonal portion of  $\mathbf{P}^f$ .

*d. Ensemble transform with regularization (ETR)*

In this section, a data assimilation scheme based on the ensemble perturbation ET method is introduced. In the ET method,  $\mathbf{X}^a$  is obtained from the background field span by the ensemble [(7)]. Here,  $\mathbf{T}$  is given as (see Bishop and Toth 1999)

$$\mathbf{T} = \mathbf{D} \mathbf{\Gamma}_D^{-1/2}, \tag{11}$$

where  $\mathbf{D}$  contains column of orthonormal eigenvectors of  $(K - 1)^{-1} \mathbf{X}^{fT} (\mathbf{P}^a)^{-1} \mathbf{X}^f$ , and  $\mathbf{\Gamma}_D$  is the diagonal matrix with the associated eigenvalues. In the new scheme,  $\mathbf{P}^f$  is obtained using 6-h ET ensemble forecasts, and the rank deficiency problem of  $\mathbf{P}^f$  is addressed using a ridging procedure (Tikhonov 1943; Ott et al. 2004):

$$\mathbf{P}^{f-RID} = \mathbf{P}^f + \lambda \mathbf{I}, \tag{12}$$

where the ridging parameter  $\lambda$  is a small fraction of the trace of  $\mathbf{P}^f$  divided by the dimension of the matrix. This approach uniformly modifies the variance structure of  $\mathbf{P}^f$  but not the covariance. The ridging parameter only influences the variance structure. Given that  $\lambda$  is constant, it will have more influence in the smaller than in the larger eigenvector directions. Ott et al. (2004) shows that for experiments based on the Lorenz (1996) model, the ridging approach in (12) performs better than the regular variance inflation (Anderson and Anderson 1999; Anderson 2001; Whitaker and Hamill 2002).

Once the background error covariance is estimated from (12),  $\mathbf{P}^a$  for the next data assimilation cycle can be obtained using the analysis precision equation (e.g., Kalnay 2003):

$$(\mathbf{P}^a)^{-1} = (\mathbf{P}^f)^{-1} + \mathbf{H}^T \mathbf{R}^{-1} \mathbf{H}. \tag{13}$$

From there,  $\mathbf{T}$  is computed using (11). To summarize, the new scheme, referred to here as ETR, will use the ET forecasts to model  $\mathbf{P}^f$  [(4)], will use a regularization method to prevent filter divergence [(12)], and will use the analysis precision [(13)] to estimate  $\mathbf{P}^a$  and, in turn, determine  $\mathbf{T}$  from where the next ensemble initial conditions will be generated using relation (7).

The similarity between (10) and (12) suggests that the ETR is a particular case of the hybrid approach, only that ETR does not perturb the off-diagonal elements of

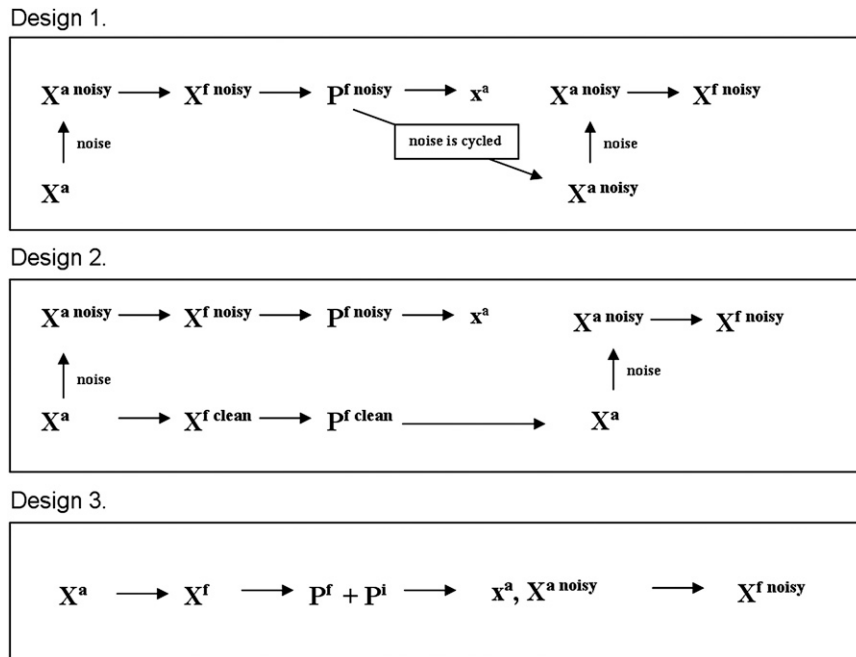


FIG. 1. Schematics of types of procedures to avoid filter divergence due to subsampled  $\mathbf{P}^f$ . In design 1, ad hoc noise is added to  $\mathbf{X}^a$  and is being cycled. In design 2, the ad hoc noise is added but is not cycled. In design 3, the background error covariance is modified directly by adding an invertible matrix.

$\mathbf{P}^f$ . This is an important feature of ETR that will reduce the introduction of noise into the dynamically generated error covariance matrices.

#### 4. Special designs to prevent filter divergence

Several procedures to deal with sampling errors and avoid filter divergence have been developed. This section describes three techniques that result in the introduction of noise into the ensemble data assimilation schemes. In design 1 of Fig. 1,  $\mathbf{X}^a$  is treated with an artificial noise by adding normally distributed random noise to form a new matrix  $\mathbf{X}^{a\_noisy}$  just before model integration. The ensemble forecast that comes out of these noisy perturbations  $\mathbf{X}^{f\_noisy}$  generates  $\mathbf{P}^{f\_noisy}$ , from which both the next analysis  $\mathbf{x}^a$  and the next analysis ensemble  $\mathbf{X}^{a\_noisy}$  are derived. After that, the cycle is repeated by further adding ad hoc noise, and so on. In this design the noise is propagated from one assimilation cycle to the next via the noisy analysis ensemble.

Design 2 is similar to the one just described but prevents cycling the ad hoc noise. Here, two ensembles are run in parallel: one initialized from  $\mathbf{X}^{a\_noisy}$  as in the procedure described above and the other initialized directly from  $\mathbf{X}^a$ . The ensemble from the noisy perturbations is used only to compute the new analysis  $\mathbf{x}^a$ , whereas the “clean” ensemble is used to compute  $\mathbf{X}^a$ ,

which provides initial perturbations for the next ensemble forecast. This design, which is computationally inefficient as it doubles the amount of computer requirements, is introduced here to analyze the effects of reduced cycling of noise.

In design 3,  $\mathbf{X}^a$  is integrated forward to produce  $\mathbf{X}^f$  and, subsequently,  $\mathbf{P}^f$ . To this resulting matrix, an invertible matrix  $\mathbf{P}^{f\_invertible}$  is added to produce  $\mathbf{x}^a$ . The hybrid method (Hamill and Snyder 2000) and the ETR belong to this class of schemes. In the hybrid scheme,  $\mathbf{P}^{f\_invertible} = \mathbf{P}^{f\_NMC}$ , whereas in the ETR scheme  $\mathbf{P}^{f\_invertible} = \mathbf{I}$ . In both methods the ad hoc term alters the dynamically produced background error covariance, but the latter affects the diagonal, thus preserving the background covariance structure.

In addition to a control scheme based on the NMC method, four ensemble data assimilation schemes are evaluated. The first is based on the ETKF scheme with additive random noise, or ETKFa. This treatment to prevent filter divergence corresponds to design 1 in Fig. 1. The noise is added on each analysis cycle and on each model grid point and consists of a random perturbation with a normal distribution of mean of zero and variance  $\sigma_r$ . This variance is optimized to render the best analysis using the MAE metric.

The second scheme is the ETKF without cycling noise, or ETKFn. This scheme has identical settings as

ETKFa except that the impacts of the noisy background error covariance are removed at each data assimilation cycle, as described in design 2 in Fig. 1. This scheme generates two background error covariances  $\mathbf{P}^{f\text{-noisy}}$  and  $\mathbf{P}^{f\text{-clean}}$  with the former used in the assimilation scheme to compute  $\mathbf{x}^a$  and the latter used to compute  $\mathbf{X}^a$  for the next analysis cycle. A comparison of ETKFa and ETKFn provides an assessment of the effects of cycling noise.

The third scheme evaluated, or hybrid, linearly combines the ETKF and the NMC methods to produce the background error covariance. Unlike ETKFa and ETKFn, the ETKF equations are applied without the additive noise. The last scheme, or ETR, combines a small fraction of  $\lambda\mathbf{I}$  and the ETKF matrix. Comparing the hybrid with the ETR will reveal the effects of adding the NMC structure, which contains sampling errors particularly reflected in the off-diagonal covariance matrix elements. The hybrid and the ETR belong to a class of schemes depicted in design 3 in Fig. 1. Notice that the hybrid and the ETR add a well-behaved invertible matrix (correlated noise) to the ETKF, whereas the ETKFa and ETKFn add random noise.

## 5. Results

### a. Snapshot of error covariances prior to filter divergence

This section describes how noise injected into the assimilation schemes modifies the physical and eigenmode structures of the background error covariance prior to filter divergence due to the direct application of the ETKF. Whether this is associated with a physical instability of the dynamical system or simply the effect of numerical (e.g., truncation) errors is not addressed in this study. The focus is on how the different schemes modify their background covariances to prevent filter divergence. It has been recognized (e.g., Houtekamer and Mitchell 1998; Anderson and Anderson 1999) that a rank-deficient  $\mathbf{P}^f$  induces an underestimation of forecast error variance, producing an analysis closer to the background, which, in turn, produces a  $\mathbf{P}^f$  with fewer degrees of freedom in the next data assimilation cycle, and so on, to the point where the analysis and the background field are indistinguishable and are both far away from the true state of the system.

Figure 2 shows a snapshot of the background error covariance matrix and corresponding eigenvalue spectrum for each scheme tested. In Fig. 2a, the NMC method produces a matrix that is spatially smooth and symmetric with respect to the diagonal. The elements far away from the diagonal drop to zero, indicating that

there is little or no correlation among grid points that are far from each other. In this case, the sampling errors are evident in the high covariance for long distances in some points. Because the NMC method uses a climatological sample to produce  $\mathbf{P}^f$ , it is generally well behaved but does not contain flow-dependent features. The spectrum of eigenvalues in the bottom panel Fig. 2 shows nonzero values. Figure 2b shows the background error covariance matrix associated with the direct application of ETKF without the addition of noise. The ETKF method detects flow-dependent features that are displayed as organized areas with high-covariance values. Note, for example, the feature centered on grid points 14 and 10 with large correlations with the close-neighbor grid points. The spectrum of eigenvalues, shown at the bottom of Fig. 2b, indicates a large drop in amplitude from eigenvalue 4 to 5 and a zero amplitude of the last eigenvalue. Figure 2c shows the background error covariance and eigenvalues associated with the ETKFa scheme. For this example with  $\sigma_r = 0.25$ , the trace of the covariance increases with respect to the ETKF. The small amount of noise maintains some but not all features. The spectrum is now smooth but does not resemble the original spectrum of Fig. 2b. The spatial structure produced by the ETKFn method is shown in Fig. 2d and appears similar to that of Fig. 2c. The bottom panel in Fig. 2d shows  $\mathbf{P}^{f\text{-clean}}$  as filled bars, which should be the original spectrum of Fig. 2b, and  $\mathbf{P}^{f\text{-noisy}}$  as empty bars. Here,  $\mathbf{P}^{f\text{-clean}}$  is used to generate the initial conditions for the ensemble forecasts of the next analysis cycle.

Figure 2e shows that the spatial structure of the background error covariance associated with the hybrid scheme is smooth, inherited from the NMC background error matrix, and with superimposed flow-dependent features from the ETKF. Likewise, the eigenvalues of the hybrid covariance matrix, depicted by the empty bars, show an amplitude structure that is a combination of the NMC and ETKF with no treatment (filled bars in Fig. 2e). This amplitude varies depending on the  $\alpha$  in Fig. 2f; both the structure of the background error covariance and the spectrum resemble that of the ETKF with no treatment, except that the spectrum has nonzero amplitude for eigenvalue 20.

Figure 2 indicates, as expected, that the methods detect the same signal in the last ETKF cycle but act differently. The ETKFa method produced nonzero eigenvalues at the end of the spectrum. The hybrid approach produced a spectrum that is shaped by both the NMC method and the ETKF. The two approaches modify the spectrum of eigenvalues. On the other hand, the ETR preserves the ETKF structure and fills the nonzero values at the end of the spectrum.

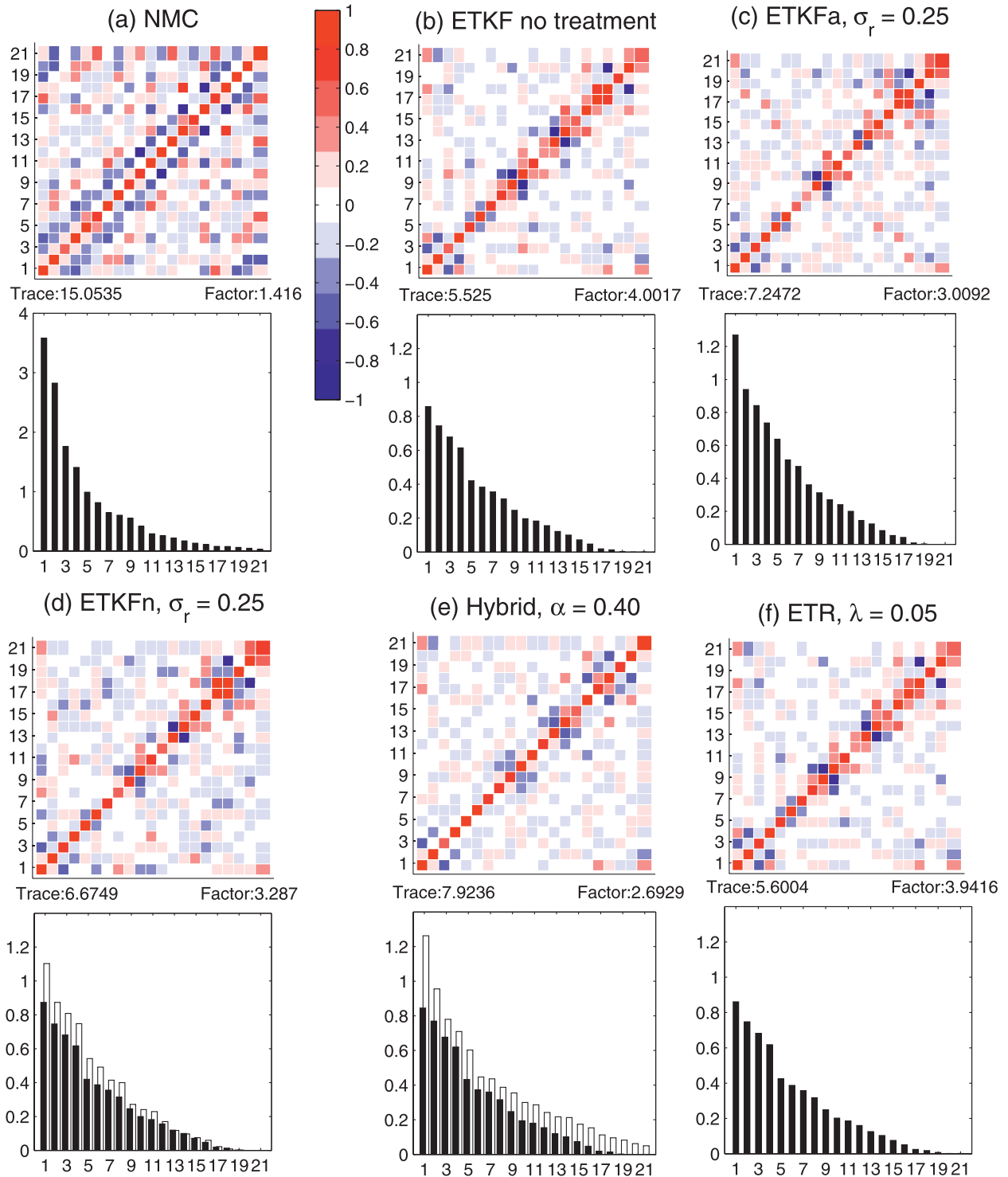


FIG. 2. The  $P^f$  and eigenvalues for the methods tested. Ensemble includes 20 members. All schemes have identical initial conditions and stop at the fourth analysis cycle. Parameter values are adjusted according to a trial period.

### b. Analysis performance

The analysis error of each scheme is evaluated for a range of values of their tuning parameters. For the ETKFa and the ETKFn, the analysis error is evaluated as a function of  $\sigma_r$ , of the random additive noise. The analysis error is evaluated as a function of the relative  $\alpha$  between the NMC and the ETKF for the hybrid and as a function of  $\lambda$  for the ETR. This procedure is repeated for different ensemble sizes. Figure 3 illustrates the typical behavior of analysis errors by plotting the mean as a function of the noise injected. The time mean average in Fig. 3a indicates that the ETKFa technique produces larger errors than the rest of the schemes but less than the NMC control for small values of  $\sigma_r$ . ETKFn shows better analysis performance than ETKFa but tends to be more prone to changes for very small values of  $\sigma_r$ , which is possibly caused by  $\mathbf{P}^{f\text{-clean}}$  becoming ill-conditioned when computing the next ensemble initialization. For a wide range of parameter values, the performance of hybrid, which uses climatological covariance estimates, is better than the other approaches that add random noise. The ETR, when all methods are tuned, outperforms the rest of the methods due to the minimal injection of noise into the background covariance matrix. The median error (Fig. 3b), which is less sensitive to outliers, suggests that the ETR has large outliers at  $\lambda \leq 0.025$ . Both the hybrid and ETR have a smooth increase in the median error as the noise increases. As expected, all methods coincide in the limit when all parameters are zero. The errors grow steadily (not shown) for larger parameter values. Overall, Fig. 3a shows that for each method there is a limit to the noise parameter values beyond which the MAE increases.

### c. Forecast performance

Figure 4 shows the forecast performance for each of the methods for the 12-member ensembles cases. For this ensemble size, the ensemble-based methods with proper treatment of their error covariance matrix generally outperform the reference NMC method. The forecast performance, as measured by the MAE, corresponds with the performance of the analysis. Because of better initial conditions, the ETR outperforms the hybrid technique, which in turn outperforms the ETKFa and ETKFn. Optimization ensembles of sizes 15 and 9 were also carried out and show similar performance patterns. The performance with 15 members is very similar to the performance with 12 members.

For ensemble sizes of 9 members, the ETR and the hybrid techniques require larger amounts of noise. For the hybrid, the optimal value for  $\alpha$  in the sample increases to 0.20; whereas for the ETR, the optimal value

for  $\lambda$  increased to 0.45. The four techniques still perform better or equal to the NMC method (not shown). For ensembles with less than 9 members, the performance deteriorates considerably. For large atmospheric models, where practical ensemble sizes are much smaller than the number of the system's degrees of freedom, additional procedures such as localization may be needed to cope with the limiting factor of small sample size in the estimation of the background error covariance. When implementing any technique, however, caution is necessary when limiting the extent to which the dynamically determined covariance information is adulterated.

## 6. Summary and conclusions

Flow-dependent data assimilation schemes have been developed to utilize information derived from ensemble forecasts. In these schemes, the background error covariance matrix is estimated explicitly by assuming that the structure of possible errors is equivalent to the structure of the deviations of ensemble members from the ensemble mean. For this to be true, it is necessary that the ensemble adequately samples the background error. For large systems, such as the ocean and the atmosphere, computational resources limit the size of practical ensembles, leading to sampling errors in the estimation of the covariance matrices. Ensembles with limited size may not sample well enough the error covariance, leading to spurious long-distance correlations and filter divergence. To avoid the subsampling problem that leads to filter divergence, some approaches use random noise, inflation, or other procedures to alter the background error covariance structure. These procedures, though they prevent filter divergence, have the unintentional effect of inserting artificial noise into the data assimilation process, which can reduce the accuracy of the data assimilation methods.

Two methods, the ETKF with additive noise (ETKFa) and a combination of the ETKF and NMC (hybrid), are compared with the corresponding new methods of ETKF with no cycled noise (ETKFn) and the ETR to assess the impacts of reducing or eliminating ad hoc noise in data assimilation. The ETKFn is a design that concurrently runs two ensembles: one that is and another that is not directly influenced by the addition of random noise. The ensemble that is initialized after noise is added is used to compute the background error covariance, after which this ensemble is discarded, whereas the other "noise free" ensemble is used as a first guess. The ETR is a method that treats the ET's background error covariance with a regularization procedure to directly address the problem of matrix rank deficiency. The regularization procedure enhances the variance (Ott



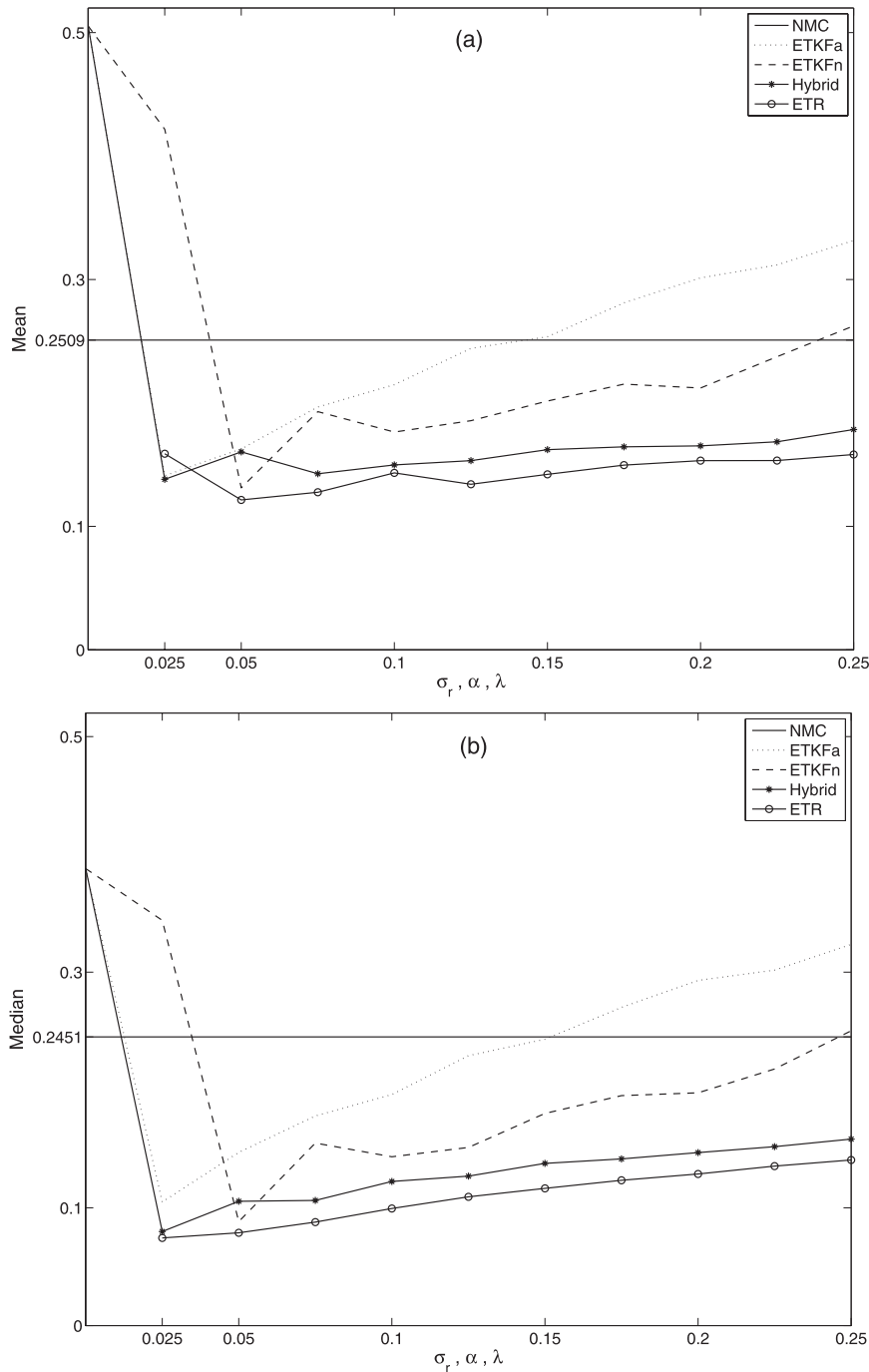


FIG. 3. (a) Mean and (b) median of time series of the absolute analysis error as a function of the corresponding noise parameter for each technique. The ensemble size is 12, and the sample size is 1920. The control method (NMC) is included.

et al. 2004) with minimal or no impact on the covariance structure. The ETR scheme has a formal similarity to the hybrid scheme of Hamill and Snyder (2000) as both methods alter the background error covariance matrix with an invertible matrix; however, the ETR does not

modify the off-diagonal elements of the matrix, which happens in the hybrid scheme. The regularization procedure could be applied also to the ETKF method to prevent filter divergence. In this study we found no significant improvement in the performance over the ETKFa.

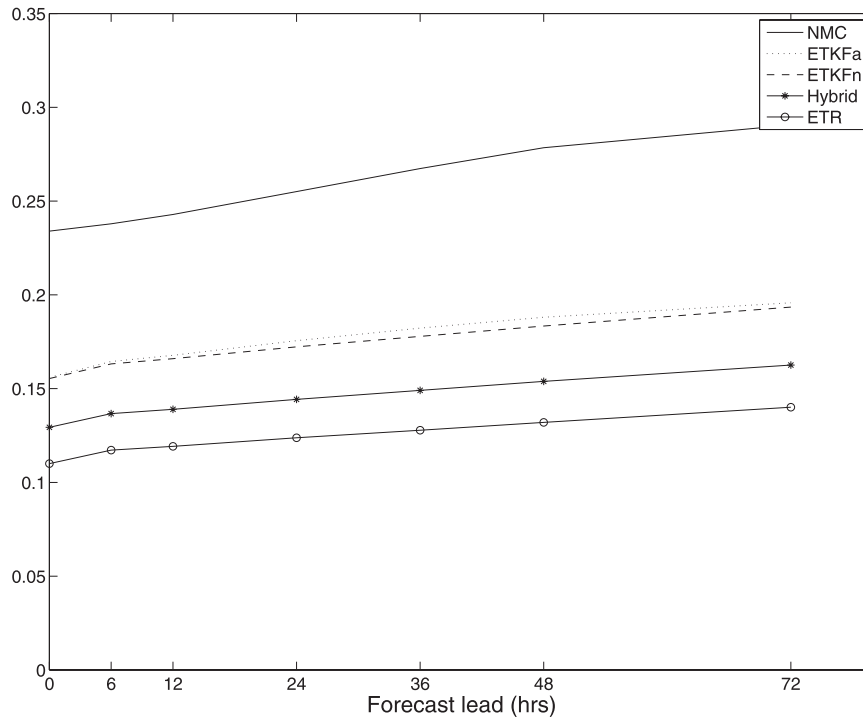


FIG. 4. Average forecast performance for 12-member ensembles where the parameters have been set to 0.05 in each of the schemes. The forecast of the NMC method is shown as a reference.

Anderson (2001) and Whitaker and Hamill (2002) showed that the addition of noise to the observations has the effect of both reducing the accuracy of the analysis-error covariance estimate by increasing the sampling error and increasing the likelihood that the analysis error covariance will be underestimated by the ensemble. The present study goes further by assessing how the addition of artificial (e.g., random) perturbations to the ensemble may increase the analysis and forecast errors under similar conditions and methods in a perfect model environment.

In this study, it was found that the performance of the ETKFa method can be marginally improved by the introduction of a second set of ensemble forecasts for cycling the background error covariance information (ETKFn). This second ensemble is not directly affected by the noise that is added only to the other ensemble for the computation of covariances. Further improvements are possible when the random noise is replaced by correlated noise in the hybrid method. The best performance is observed when the influence of noise on covariance propagation is further reduced with the use of a ridging method that affects only the diagonal part of the covariance matrix (ETR). Furthermore, evidence is given that for each method, there is a certain amount of noise beyond which the analysis and forecast performance deteriorates as the amount of induced noise increases.

The results presented in this study suggest that noise introduced intentionally or inadvertently into the generation of ensembles in attempts to reduce problems arising from sampling errors in the estimation of background covariances can deteriorate the quality of the analysis as it interferes with the dynamically based evolution of ensemble perturbations. The results suggest that ensemble perturbation approaches such as the breeding (Toth and Kalnay 1993, 1997) and ET (Wei et al. 2008; McLay et al. 2008) methods that are designed to carry nonlinear perturbations with no or limited interference from sources external to the dynamics of the system may offer valuable information on the background error covariance in evolving data assimilation schemes. Finally, note that localization procedures, usually carried out in conjunction with additive inflation noise, systematically limit the noise in the off-diagonal elements, resulting in improved performance over methods with the additive noise alone.

*Acknowledgments.* The authors wish to thank Deryl Kleist and David Parrish from EMC for their comments on an earlier version of the manuscript. Special thanks to Milija Zupanski for his insightful comments on the equivalence between ETR and ETKF under certain conditions. Shu-Chi Yang and an anonymous reviewer

are also thanked for their valuable comments that helped to further clarify the manuscript.

## REFERENCES

- Anderson, J. L., 2001: An ensemble adjustment Kalman filter for data assimilation. *Mon. Wea. Rev.*, **129**, 2884–2903.
- , 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.
- Bishop, C. H., and Z. Toth, 1999: Ensemble transformation and adaptive observations. *J. Atmos. Sci.*, **56**, 1748–1765.
- , B. J. Etherton, and S. J. Majumdar, 2001: Adaptive sampling with the ensemble transform Kalman filter. Part I: Theoretical aspects. *Mon. Wea. Rev.*, **129**, 420–436.
- Corazza, M., E. Kalnay, D. J. Patil, E. Ott, J. Yorke, I. Szunyogh, and M. Cai, 2002: Use of the breeding technique in the estimation of the background error covariance matrix for a quasigeostrophic model. Preprints, *Symp. on Observations, Data Assimilation and Probabilistic Prediction*, Orlando, FL, Amer. Meteor. Soc., 154–157.
- , —, and S. C. Yang, 2007: An implementation of the local ensemble Kalman filter in a quasi-geostrophic model and comparison with 3D-Var. *Nonlinear Processes Geophys.*, **14**, 89–101.
- 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.
- , 2003: The ensemble Kalman filter: Theoretical formulation and practical implementation. *Ocean Dyn.*, **53**, 343–367.
- , and P. J. van Leeuwen, 1996: Assimilation of Geosat altimeter data for the Agulhas current using the ensemble Kalman filter with a quasigeostrophic model. *Mon. Wea. Rev.*, **124**, 85–96.
- Gaspari, G., and S. E. Cohn, 1999: Construction of correlation functions in two and three dimensions. *Quart. J. Roy. Meteor. Soc.*, **125**, 723–757.
- Hamill, T. M., and C. Snyder, 2000: A hybrid ensemble Kalman filter 3D variational analysis scheme. *Mon. Wea. Rev.*, **128**, 2905–2919.
- Houtekamer, P. L., and H. L. Mitchell, 1998: Data assimilation using 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.
- Hunt, B., E. Kostelich, and I. Szunyogh, 2007: Efficient data assimilation for spatiotemporal chaos: A local ensemble transform Kalman filter. *Physica D*, **230**, 112–126.
- Kalnay, E., 2003: *Atmospheric Modeling, Data Assimilation and Predictability*. Cambridge University Press, 340 pp.
- Lorenz, E. N., 1996: Predictability: A problem partly solved. *Proc. Seminar on Predictability*, Vol. 1, Reading, United Kingdom, ECMWF, 1–18.
- , 2006: Regimes in simple models. *J. Atmos. Sci.*, **63**, 2056–2073.
- Majumdar, S. J., C. H. Bishop, B. J. Etherton, I. Szunyogh, and Z. Toth, 2001: Can an ensemble transform Kalman filter predict the reduction in forecast error variance produced by targeted observations? *Quart. J. Roy. Meteor. Soc.*, **127**, 2803–2820.
- , —, —, and —, 2002: Adaptive sampling with the ensemble transform Kalman filter. Part II: Field program implementation. *Mon. Wea. Rev.*, **130**, 1356–1369.
- McLay, J. G., C. H. Bishop, and C. A. Reynolds, 2008: Evaluation of the ensemble transform analysis perturbation scheme at NRL. *Mon. Wea. Rev.*, **136**, 1093–1108.
- Meng, Z., and F. Zhang, 2008: Test of an ensemble Kalman filter for mesoscale and regional-scale data assimilation. Part III: Comparison with 3DVAR in a real-data case study. *Mon. Wea. Rev.*, **136**, 522–540.
- Ott, E., and Coauthors, 2004: A local ensemble Kalman filter for atmospheric data assimilation. *Tellus*, **56A**, 415–428. [Available online at <http://arxiv.org/abs/physics/0203058v4>.]
- Parrish, D., and J. Derber, 1992: The National Meteorology Center's spectral statistical interpolation analysis system. *Mon. Wea. Rev.*, **120**, 1747–1763.
- Patil, D., B. R. Hunt, E. Kalnay, J. A. Yorke, and E. Ott, 2001: Local low dimensionality at atmospheric dynamics. *Phys. Rev. Lett.*, **86**, 5878–5881.
- Szunyogh, I., Z. Toth, A. Zimin, S. Majumdar, and A. Persson, 2000: The effect of targeted dropsonde observations during the 1999 Winter Storm Reconnaissance Program. *Mon. Wea. Rev.*, **128**, 3520–3537.
- Tikhonov, A. N., 1943: On the stability of inverse problems. *Dokl. Akad. Nauk SSSR*, **39**, 195–198.
- Tippett, M. K., J. L. Anderson, C. H. Bishop, T. M. Hamill, and J. S. Whitaker, 2003: Ensemble square root filters. *Mon. Wea. Rev.*, **131**, 1485–1490.
- Toth, Z., and E. Kalnay, 1993: Ensemble forecasting at NMC: The generation of perturbations. *Bull. Amer. Meteor. Soc.*, **74**, 2317–2330.
- , and —, 1997: Ensemble forecasting at NCEP: The breeding method. *Mon. Wea. Rev.*, **125**, 3297–3318.
- Wang, X., and C. H. Bishop, 2003: A comparison of breeding and ensemble transform Kalman filter ensemble forecast schemes. *J. Atmos. Sci.*, **60**, 1140–1158.
- , T. M. Hamill, J. S. Whitaker, and C. H. Bishop, 2007: A comparison of hybrid ensemble transform Kalman filter–optimum interpolation and ensemble square root filter analysis schemes. *Mon. Wea. Rev.*, **135**, 1055–1076.
- Wei, M., Z. Toth, R. Wobus, and Y. Zhu, 2006: Ensemble transform Kalman filter-based ensemble perturbations in an operational global prediction system at NCEP. *Tellus*, **58A**, 28–44.
- , —, —, and —, 2008: Initial perturbations based on the ensemble transform (ET) technique in the NCEP Global Operational Forecast System. *Tellus*, **60A**, 62–79.
- Whitaker, J. S., and T. M. Hamill, 2002: Ensemble data assimilation without perturbed observations. *Mon. Wea. Rev.*, **130**, 1913–1924.
- , —, X. Wei, Y. Song, and Z. Toth, 2008: Ensemble data assimilation with the NCEP Global Forecast System. *Mon. Wea. Rev.*, **136**, 463–482.
- Yang, S.-C., M. Corazza, F. Carrassi, E. Kalnay, and T. Miyoshi, 2009: Comparison of the local ensemble transform Kalman Filter, 3DVAR and 4DVAR in a quasigeostrophic model. *Mon. Wea. Rev.*, **137**, 693–709.
- Zhang, F., C. Snyder, and J. Sun, 2004: Impacts of initial estimate and observation availability on convective-scale data assimilation with an ensemble Kalman filter. *Mon. Wea. Rev.*, **132**, 1238–1253.
- , Z. Meng, and A. Aksoy, 2006: Test of an ensemble Kalman filter for mesoscale and regional-scale data assimilation. Part I: Perfect model experiments. *Mon. Wea. Rev.*, **134**, 722–736.