A Hybrid Ensemble Kalman Filter–3D Variational Analysis Scheme

THOMAS M. HAMILL AND CHRIS SNYDER

National Center for Atmospheric Research, Boulder, Colorado*

(Manuscript received 15 October 1999, in final form 27 January 2000)

ABSTRACT

A hybrid ensemble Kalman filter–three-dimensional variational (3DVAR) analysis scheme is demonstrated using a quasigeostrophic model under perfect-model assumptions. Four networks with differing observational densities are tested, including one network with a data void. The hybrid scheme operates by computing a set of parallel data assimilation cycles, with each member of the set receiving unique perturbed observations. The perturbed observations are generated by adding random noise consistent with observation error statistics to the control set of observations. Background error statistics for the data assimilation are estimated from a linear combination of time-invariant 3DVAR covariances and flow-dependent covariances developed from the ensemble of short-range forecasts. The hybrid scheme allows the user to weight the relative contributions of the 3DVAR and ensemble-based background covariances.

The analysis scheme was cycled for 90 days, with new observations assimilated every 12 h. Generally, it was found that the analysis performs best when background error covariances are estimated almost fully from the ensemble, especially when the ensemble size was large. When small-sized ensembles are used, some lessened weighting of ensemble-based covariances is desirable. The relative improvement over 3DVAR analyses was dependent upon the observational data density and norm; generally, there is less improvement for data-rich networks than for data-poor networks, with the largest improvement for the network with the data void. As expected, errors depend on the size of the ensemble, with errors decreasing as more ensemble members are added. The sets of initial conditions generated from the hybrid are generally well calibrated and provide an improved set of initial conditions for ensemble forecasts.

1. Introduction

Since Lorenz (1963, 1969) it has been recognized that perfect numerical weather forecasts will always be unattainable; even the smallest of errors in the initial condition will grow inexorably, eventually rendering any single deterministic forecast useless. Rather than pinning unrealistic hopes upon the accuracy of a single numerical forecast, Epstein (1969) suggested an alternative goal for numerical weather forecasting, namely, to estimate future states of the atmosphere's probability density function (pdf) given an estimate of the initial pdf. Typically, the pdf evolves from a relatively specific distribution of initial states through increasingly more diffuse states. At each forecast lead time, the user is provided the probability of each model state occurring.

Explicitly computing the evolution of a complex model's pdf is not presently feasible, so an alternative, ensemble forecasting, or EF (Leith 1974), has been embraced. In EF, multiple, individual numerical forecasts are generated from different sets of initial conditions and/or different numerical model configurations. The presumption underlying EF is that the subsequent sets of forecasts may be taken as a representative random sample from the evolved pdf. Ensemble forecasts have been produced operationally in the United States and Europe since late 1992 (Toth and Kalnay 1993, 1997; Molteni et al. 1996).

Ensemble forecasts are potentially useful for more than just making probabilistic weather forecasts; they may also provide information that may be used during data assimilation to generate more accurate initial conditions. A key avenue to improving data assimilation is accurate specification of the error statistics for the background forecast, also known as the ''prior'' or ''first guess'' (Schlatter et al. 1999). These background error statistics are used to determine the relative weighting in the analysis between the background and the observations and the influence of observations away from the observation locations. Many current and past operational data assimilation methods use long time series of previous forecasts to develop spatially homogeneous and temporally invariant approximations to background error statistics. Schemes that use such statistics include

* The National Center for Atmospheric Research is sponsored by the National Science Foundation.

Corresponding author address: Dr. Thomas M. Hamill, NCAR/MMM/ASP, P.O. Box 3000, Boulder, CO 80307-3000.
E-mail: hamill@ucar.edu

optimum interpolation (e.g., Gandin 1963; Schlatter 1975; Lorenc 1981) and three-dimensional variational data assimilation, or 3DVAR (Lorenc 1986; Parrish and Derber 1992, hereafter PD92). Such simplified error statistics are necessitated by the computational difficulty of specifying accurate flow-dependent error statistics. Increasing computational resources, however, have opened new possibilities, including the use of EFs for estimating background errors.

The ensemble Kalman filter (EnKF, hereafter; see Evensen 1994; Evensen and van Leeuwen 1996; Houtekamer and Mitchell 1998; Burgers et al. 1998; Mitchell and Houtekamer 2000) is one such possibility. The EnKF consists of a set (or ensemble) of parallel short-term forecasts and data assimilation cycles. This ensemble differs from present operational ensembles in that it is intended to provide a probablisitic forecast in the short term rather than at the medium range. Since, in the short term, forecast errors retain memory of the initial (analysis) errors, the EnKF must therefore incorporate probabilistic information on analysis errors in the generation of the ensemble. This is accomplished, at least approximately, by providing a set of ''perturbed'' observations to each of the member assimilation cycles. These perturbed observations consist of the actual observations plus distinct realizations of random noise, whose statistics are consistent with the assumed observational error covariances.

The EnKF is related to the classic Kalman filter (KF, hereafter; Kalman and Bucy 1961), which provides the optimal analysis in the case that the forecast dynamics are linear and both background and observation errors have normal distributions. The primary difference is that the KF explicitly forecasts the evolution of the complete forecast error covariance matrix using linear dynamics, while the EnKF estimates this matrix from a sample ensemble of fully nonlinear forecasts. The EnKF also addresses the computational difficulty of propagating or even storing the forecast error covariance matrix, whose elements equal in number the *square* of the dimension of the forecast model (number of grid points times number of variables). Under assumptions of linearity of error growth and normality of observation and forecast errors, it can be shown that in fact this scheme produces the correct background error covariances as the ensemble size increases (Burgers et al. 1998). For smaller ensembles, however, the EnKF is rank deficient and its background covariance estimates suffer from a variety of sampling errors, including spurious correlations at between widely separated locations. Alternative approaches to both the extended KF and covariance propagation are reviewed in Ghil (1997).

The EnKF and KF approaches are being considered along with other sophisticated and computationally intensive approaches to data assimilation. Many operational centers are using or experimenting with using four-dimensional variational data assimilation systems, or 4DVAR (Thompson 1969; Daley 1991; Thépaut et al. 1993; Courtier et al. 1994; Fisher and Courtier 1995; Talagrand 1997). This technique finds the trajectory that best fits past and present observations. Like the EnKF, 4DVAR is computationally intensive, requiring multiple integrations of tangent-linear and adjoint versions of the forecast model.

Houtekamer and Mitchell (1998, hereafter HM98) first demonstrated the EnKF in a meteorological context. They used a quasigeostrophic T21 spectral resolution, three-level global model (Marshall and Molteni 1993) under perfect-model assumptions. They demonstrated that analysis errors decreased significantly as ensemble size is increased. However, HM98 also noted problems with the EnKF; first, with small ensemble sizes, the approximation of the background error covariances from the ensemble was poor. There were also issues of rank deficiency, and background error covariances were biased when the member being updated was also used in the estimate of the background covariances (see also van Leeuwen 1999). To address this bias, HM98 proposed the use of a ''double EnKF,'' whereby the n-member ensemble was split up into two (n/2)-member subsets. The covariance information from one of the subsets was used in the data assimilation of the other subset. A disadvantage of this approach was that estimates of the background error statistics did not fully use all of the available information from the ensemble, but rather only half, exacerbating the sampling error and rank deficiency problems in the background error covariance estimate.

HM98 also found that it was desirable to exclude the effects of observations greatly separated from the analysis location. This was done to deal with the spurious correlations over long distances produced by a relatively small ensemble and to address rank deficiency problems. However, the use of a cutoff radius may introduce undesirable small-scale noise into the analysis, especially when observations are sparse. These discontinuities occur at points where adjacent grid points are updated using different sets of observations, that is, where one or more observations are included at one grid point and excluded at the next. More recently, P. L. Houtekamer and H. L. Mitchell (1999, personal communication) have attempted to minimize this effect by multiplying the EnKF-supplied background error covariances by a weighting factor that decreases smoothly to zero at finite distance.

Over the short term, limited computational resources may make it difficult or impossible to run an operational EnKF with a large number of members. If so, it would be appealing to have an algorithm that could still work with smaller-sized ensembles and that could benefit from whatever flow-dependent information this smaller ensemble provides. In this paper we demonstate how to construct a hybrid EnKF–3DVAR analysis scheme with these attractive properties. This scheme uses background errors from an ensemble only to an extent appropriate for the size of the ensemble by weighting flow-dependent background error statistics from the ensemble

together with the time-independent statistics from 3DVAR. The relative weighting can be adjusted to the observational network and the size of the ensemble.

An ancillary goal of this paper is to demonstrate that our hybrid scheme generates initial conditions that will produce a superior ensemble of forecasts at longer lead times. Currently, ensemble initial conditions for operational forecasts in the United States and Europe are created by adding ''dynamically constrained'' noise onto a control analysis (Toth and Kalnay 1993, 1997; Molteni et al. 1996). We have previously demonstrated that a perturbed observation (PO) ensemble, constructed in a manner similar to the system to be demonstrated here, may produce ensemble forecasts that are superior to the dynamically constrained methods (Hamill et al. 2000, hereafter HSM00). The only difference between the hybrid scheme described here and the previously demonstrated PO approach is that the the data assimilation method for the PO ensemble uses 3DVAR, so its estimate of background errors does not include information from the ensemble. For comparison, our tests here will be measured against 3DVAR and PO ensemble benchmarks.

We conduct our experiments here under a ''perfect model'' assumption, whereby the same model used to generate a true solution is used to generate forecasts. We also use a quasigeostrophic channel model. Our results are thus not a direct analog to real-world numerical weather prediction, where error growth due to model deficiencies may be significant or greater than chaotic effects due to initial state deficiencies (Tribbia and Baumhefner 1988). Perfect-model experiments, however, do indicate how best to design an ensemble of initial conditions in the absence of model error. We presume then that model error will be (and should be) addressed as a further problem, be it through stochastic physics (e.g., Buizza et al. 1999), through modeling of errors in the data assimilation process (Mitchell and Houtekamer 2000), through the use of perturbations to model fixed fields and the land surface (Houtekamer and Derome 1995; Hamill and Colucci 1998a), or through other methods.

The rest of the article will be organized as follows. We start with a brief review of the experimental design and forecast model (section 2) and the simulation concepts and the hybrid analysis scheme design (section 3). We continue with an examination of the analysis error characteristics (section 4) and subsequent probabilistic forecast error characteristics (section 5). Section 6 concludes the paper.

2. Experimental design

Our experiments begin from the assumption of a perfect model. Thus, a long reference integration of the quasigeostrophic (QG) model provides the true state; the assimilation and forecast experiments then use that same model together with (imperfect) observations of the true state.

The quasigeostrophic model is the same one used in HSM00. It is a midlatitude, beta-plane, gridpoint channel model that is periodic in x (east–west), has impermeable walls on the north–south boundaries, and rigid lids at the top and bottom. There is no terrain, nor are there surface variations such as land and water. Pseudo–potential vorticity (PV) is conserved except for Ekman pumping at the surface, ∇^4 horizontal diffusion, and forcing by relaxation to a zonal mean state. The domain is 16 000 km $\times$ 8000 km $\times$ 9 km; there are 129 grid points east–west, 65 north–south, and eight model forecast levels, with additional staggered top and bottom levels at which potential temperature θ is specified. Forecast parameters are set as in HSM00.

All observations are presumed to be rawinsondes, with u and v wind components and θ observed at each of the eight model levels. Observations are imperfect, with errors drawn from the Gaussian distributions specified in HSM00. Moreover, the observation-error covariances are identical to those assumed by the data assimilation scheme. Observations and new analyses are generated every 12 h, followed by a 12-h forecast with the QG model that serves as background at the next analysis time.

The experiments are based on the four observational networks shown in Fig. 1: a network with a data void in the eastern third of the domain, a low-density network (observations $\sim$every 20^2 grid points), a moderate-density network ($\sim$every 10^2 grid points), and a high-density network ($\sim$every 5^2 grid points). Observations locations were selected sequentially and randomly, using a one-dimensional Latin square algorithm (Press et al. 1992), which enforces a minimum distance between observations. The moderate-density network is a superset of the low-density network, and the high-density network a superset of the moderate. For simplicity, observations are located at the model grid points.

We will focus primarily on the analysis characteristics during a 90-day interval, with a new analysis performed every 12 h. Many of the statistics will be derived from a subset of 20 of the times in this series, with the first sample analysis taken 10 days after the start of the cycle and with 4 days between each sample analysis. For these 20 initial times for the network with the data void, we also generated an ensemble forecast to 5-days lead time for purposes of evaluating the subsequent probabilistic forecasts. Most of the experiments were conducted with a 25-member ensemble, though 50- and 100-member ensembles were generated for the network with the data void in order to examine the influence of ensemble size.

3. The ensemble and hybrid assimilation schemes

An important property of this ensemble data assimilation system is that the ensemble approximates a random sample from the analysis pdf (HSM00). To this

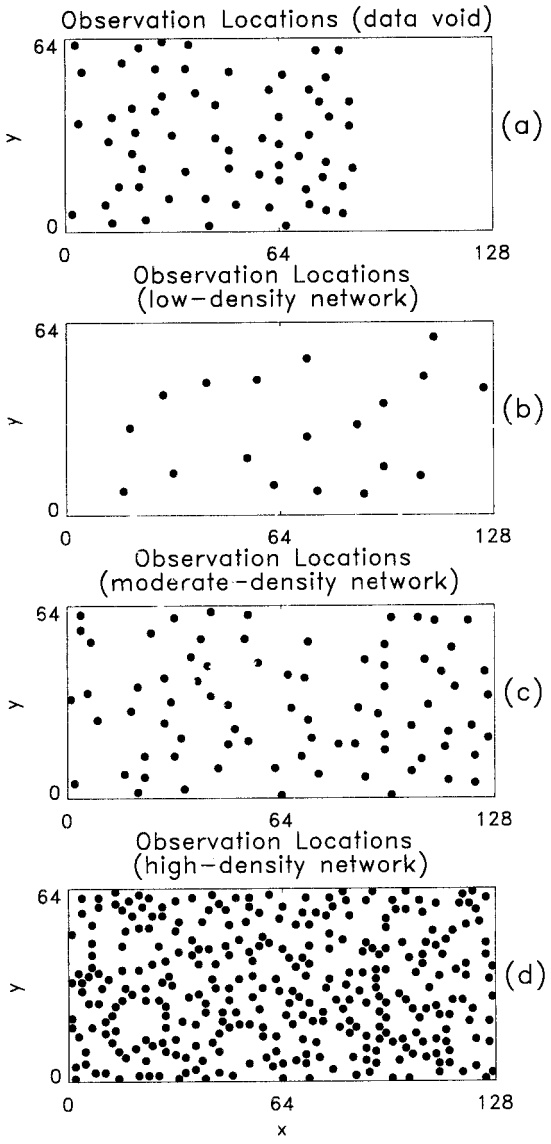

FIG. 1. Observation locations for four network configurations: (a) network with data void, (b) low-density network, (c) moderate-density network, and (d) high-density network.

end, we follow the Monte Carlo procedure pictured in Fig. 2; this procedure is identical to that of HM98 except for the differences in the assimilation method used for each member. We start with an ensemble of n analyses at some time t_0. These analyses were generated by adding perturbations to a control analysis. The specific method for how to generate this ensemble of initial conditions at the start of the cycle is not crucial for the long-term behavior; we chose to use perturbations that were constructed from scaled differences between random model states, following Schubert and Suarez (1989). We then repeat the following three-step process for each data assimilation cycle: 1) Make n forecasts to the next analysis time, here, 12 h hence. These forecasts will be used as background forecasts for n independent analyses. 2) Given the already imperfect observations at this next analysis time (hereafter called the ''control'' observations) generate n independent sets of *perturbed* observations by adding random noise to the control observations. The noise is drawn from the same distributions as the observation errors (see section 2). 3) Calculate n objective analyses, updating each of the n background forecasts using the associated set of perturbed observations. The data assimilation scheme is the hybrid EnKF–3DVAR, described below.

For comparison, we will also include a single 3DVAR control analysis receiving unperturbed observations.

Next, we describe the hybrid assimilation scheme used for each ensemble member, and whose background error covariances are a linear combination of the sample covariances from an ensemble of (12 h) forecasts and a stationary covariance matrix typically used in 3DVAR. We first briefly review the existing 3DVAR scheme for the QG model; our implementation of 3DVAR follows broadly PD92 and is described in more depth in Morss (1999). Notation below follows Ide et al. (1997).

Let $\mathbf{x}$ be the model state vector, whose elements, for the QG model, are the potential vorticity at each level and grid point, along with the potential temperature at each grid point of the top and bottom boundaries. Given a set of observations $\mathbf{y}^o$ and a background forecast $\mathbf{x}^b$, the analysis $\mathbf{x}^a$ under 3DVAR is that $\mathbf{x}$ which minimizes

$$J(\mathbf{x}) = \frac{1}{2}[(\mathbf{x} - \mathbf{x}^b)^{\mathrm{T}}\mathbf{B}^{-1}(\mathbf{x} - \mathbf{x}^b) + (\mathbf{y}^o - \mathbf{H}\mathbf{x})^{\mathrm{T}}\mathbf{R}^{-1}(\mathbf{y}^o - \mathbf{H}\mathbf{x})], \quad (1)$$

where $\mathbf{B}$ is an approximation of the background error covariances, $\mathbf{R}$ is the ''observation error'' covariance matrix, and $\mathbf{H}$ is an operator that maps the model state onto the observations (here assumed linear). In the perfect-model experiments conducted here, $\mathbf{R}$ is simply the measurement error covariance; that is, the observations are related to the true state $\mathbf{x}^t$ by $\mathbf{y}^o = \mathbf{H}\mathbf{x}^t + \boldsymbol{\varepsilon}$, where $\boldsymbol{\varepsilon}$ is a normally distributed, random vector with zero mean and covariance matrix $\mathbf{R}$.

The analysis increment $\mathbf{x}^a - \mathbf{x}^b$ satisfies

$$(\mathbf{I} + \mathbf{B}\mathbf{H}^{\mathrm{T}}\mathbf{R}^{-1}\mathbf{H})(\mathbf{x}^a - \mathbf{x}^b) = \mathbf{B}\mathbf{H}^{\mathrm{T}}\mathbf{R}^{-1}(\mathbf{y}^o - \mathbf{H}\mathbf{x}^b); \quad (2)$$

the derivation proceeds by differentiating J with respect to $\mathbf{x}$, setting the result equal to zero, and rearranging terms. At each assimilation time, our implementation of 3DVAR solves (2) for $\mathbf{x}^a - \mathbf{x}^b$ using a conjugate residual descent algorithm (Morss 1999; Smolarkiewicz and Margolin 1994). The analysis $\mathbf{x}^a$ is then used as the initial condition for a subsequent QG model forecast, and that forecast becomes the background $\mathbf{x}^b$ at the next assimilation time. Following Morss (1999), the iteration ends when the largest residual at any grid point is 5% of its maximum initial residual.

Some key elements of any 3DVAR scheme are the assumptions made to obtain the approximate back-

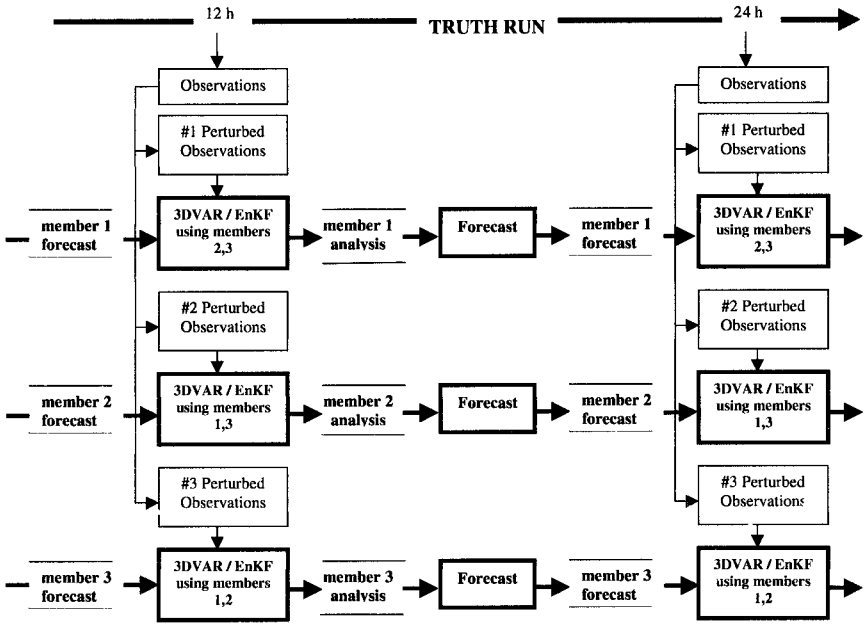

FIG. 2. Illustration of cycle used in generating hybrid analyses for a hypothetical three-member ensemble. Each member receives a different set of perturbed observations. Ensemble-based covariances are estimated using fellow members but excluding the member currently being processed.

ground covariances **B**. Here, **B** is assumed (a) to be fixed in time, (b) to be diagonal in horizontal spectral coefficients, and (c) to have separable horizontal and vertical structures with simple vertical correlations. Assumptions a and b follow PD92; for assumption b, **B** is defined by

$$\mathbf{B} = \mathbf{SCS}^{\mathrm{T}}, \tag{3}$$

where **S** is the transform from spectral coefficients to grid points and **C** is the diagonal matrix of variances of the spectral coefficients.

Though simple, this implementation of 3DVAR has characteristics that suggest its performance in the QG model is likely at least as good as the performance of operational 3DVAR scheme for primitive equation models. Most importantly, the simple covariance model (3) applies to potential vorticity. Because all other fields in the model may be derived from the potential voriticity through the usual quasigeostrophic relations, these simple covariances imply complex (and realistic) covariances for streamfunction, winds, and temperatures.

We now turn to our hybrid EnKF–3DVAR scheme. In this scheme, the approximate background covariances are not given by (3), but by a weighted mean of (3) and the sample covariance matrix $\mathbf{P}^b$ derived from the ensemble, which is fully time dependent and spatially inhomogeneous:

$$\mathbf{B} = (1 - \alpha)\mathbf{P}^b + \alpha\mathbf{SCS}^{\mathrm{T}}. \tag{4}$$

Details of the calculation of $\mathbf{P}^b$ for each ensemble member are given below. By changing α from 0.0 to 1.0, the analysis changes from using only flow-dependent, ensemble-based error covariances to using the original 3DVAR covariances, that is, a ''perturbed observation'' analysis. The analysis increments for the hybrid scheme are obtained through iterative solution of (2), just as done in our implementation of 3DVAR. In contrast, HM98 solve the Kalman gain equation directly. These procedures are equivalent mathematically, as both minimize (1).

This implementation has a number of potential advantages; first, it allows us to evaluate combinations of 3DVAR and ensemble-based background statistics, rather than relying strictly upon one or the other. Second, ensemble-based statistics alone will be rank deficient and subject to sampling errors. Blending in some 3DVAR statistics thus may ''fill out'' the covariance matrix and ameliorate some of sampling error problems encountered when using just ensemble-based statistics. The appropriate weighting to each can be adjusted through α.

Implementing this hybrid scheme requires an appropriate ensemble. We proceed as described in the beginning of section 3: Each ensemble member is used in turn as the background forecast in (2) [with **B** given by (4)] and is updated based on distinct ''perturbed'' observations (Fig. 2). Within this scheme, some subtlety is involved in the definition of the sample covariance matrix $\mathbf{P}^b$ for the ensemble. As discussed earlier, the simplest approach, in which $\mathbf{P}^b$ is the sample covariance matrix using the entire ensemble and each member is updated using the same background covariances, underestimates the variance of the background errors (HM98; van Leeuwen 1999). This underestimation can

in turn degrade the performance of the EnKF (HM98). To avoid this problem, HM98 used the double EnKF, in which the ensemble is split into two halves and the sample covariance from one half is used to update the members from the other half.

Here, as suggested in HM98, we adopt another approach and calculate $\mathbf{P}^b$ for the ith member from the sample that *excludes* the ith member, specifically,

$$\mathbf{P}_i^b = (n-2)^{-1} \sum_{j=1, j\neq i}^{n} (\mathbf{x}_j^b - \overline{\mathbf{x}}_i^b)(\mathbf{x}_j^b - \overline{\mathbf{x}}_i^b)^{\mathrm{T}}, \quad (5)$$

where subscripts refer to ensemble members, n is the number of members, and $\overline{\mathbf{x}}_i^b$ is the ensemble mean, again excluding the $\mathbf{x}_i^b$ from the sample. We have not attempted a comparison against the double EnKF, but we expect (5) to produce a more accurate background covariance estimate than the double EnKF since more ensemble members are used.

4. Analysis characteristics

a. Analysis errors and ensemble calibration as functions of α

In this section we evaluate the characteristics of analysis errors from ensembles for the different observational networks and for different values of α. We consider two aspects of the quality of the ensemble: the absolute magnitude of the error and the degree to which the ensemble approximates a random sample from the distribution of true states given past observations. We shall measure errors in the L^2 norm (streamfunction rms error), the total energy norm, and the pseudo–potential enstrophy norm, as in HSM00. These three norms emphasize different scales of motion, with the L^2 norm emphasizing the largest scales and enstrophy the smaller scales. The extent to which the ensemble is drawn from the correct distribution is evaluated using rank histograms (Anderson 1996; Hamill and Colucci 1997, 1998b; HSM00), also known as Talagrand diagrams.

Figure 3 presents the error statistics for the different networks using ensembles of size $n = 25$ and four choices of α. From Fig. 3, several points are notable. First, analysis errors generally decreased as α decreased. Generally the lowest errors were at $\alpha = 0.1$, though for the enstrophy norm in the moderate-density network, $\alpha = 0.4$ exhibited the lowest error of the α values tested. The relative decrease of error depended upon the density of observations and the norm; the ratio of analysis errors at $\alpha = 0.1$ to $\alpha = 1.0$ in the L^2 norm was around 40% for the low-density network, near 50% for the network with the data void, but less than 30% for the moderate- and high-density networks. The greater improvement for the sparse network and the network with the data void suggests that when there are fewer observations relative to the predominant wavelength (as is the case at the mesoscale), the greater the improvement of the hybrid over a standard 3DVAR.

Also note in Fig. 3 that the errors of the 3DVAR control analysis were generally about the same as the errors of the $\alpha = 1.0$ ensemble mean in the L^2 norm, but the ensemble mean errors were typically slightly lower than control errors in the enstrophy norm. With the exception of the low-density network, discussed below, the result that the ensemble mean analysis at $\alpha = 1.0$ is competitive with or better than the control 3DVAR analysis is similar to a result noted in HSM00. There we found that the ensemble mean of the perturbed observation ensemble analyses was generally better than the control analysis, especially in the enstrophy norm, due to the smoothing of smaller-scale, less predictable features. Here, the relative improvement over the 3DVAR control is perhaps less pronounced than that noted in HSM00.

For most simulations, the ensemble mean error at $\alpha = 1.0$ is less than the error of the 3DVAR control simulation (Fig. 3). However, for the low-density network, the error in the 3DVAR control appears to be substantially lower than the error of the ensemble mean simulations at $\alpha = 1.0$. This result is most likely due to an insufficiently short testing period; with so few observations in the low-density network, there were large low-frequency variations in the relative skill of the analyses, and the 90-day statistics may not have been representative of the long-term mean performance. When this comparison was repeated for a simulation almost three times as long, the relative results were reversed; ensemble mean error was much less than the 3DVAR control error. Ideally, it would be best to conduct a cycle of analyses over an even longer period of time to determine the statistical significance of any difference.

Figure 3 also suggests the use of flow-dependent covariances in the data assimilation changes analysis characteristics in other ways, in particular by reducing temporal variations in analysis error. The standard deviation of domain-average analysis error over the 90 days was typically greater when α was larger, especially for the network with the data void. This is further demonstrated in Figs. 4a,b, a time series of analysis errors for the data void at $\alpha = 0.1$ and $\alpha = 1.0$. As shown, one effect of using primarily flow-dependent covariances ($\alpha = 0.1$) was to reduce the errors during the long spells where 3DVAR errors ($\alpha = 1.0$) are especially high; that is, days with high errors were improved more than days with low errors. Also, the errors were reduced to a greater extent over the data void than over the data-rich region, as shown in Figs. 5a,b.

The effect of ensemble size on accuracy was examined only for the network with the data void. Ensembles of size $n = 50$ and $n = 100$ were also computed. Figure 6 plots the ensemble mean errors in the L^2 norm; other norms were similar. There is a marked reduction in error at $\alpha = 0.1$ for $n = 50$ compared to $n = 25$, and yet slightly lower errors with $n = 100$. For $\alpha = 0.4$, $n = 50$ has slightly higher errors than $n = 25$ members; again, we expect this result is due to testing the scheme

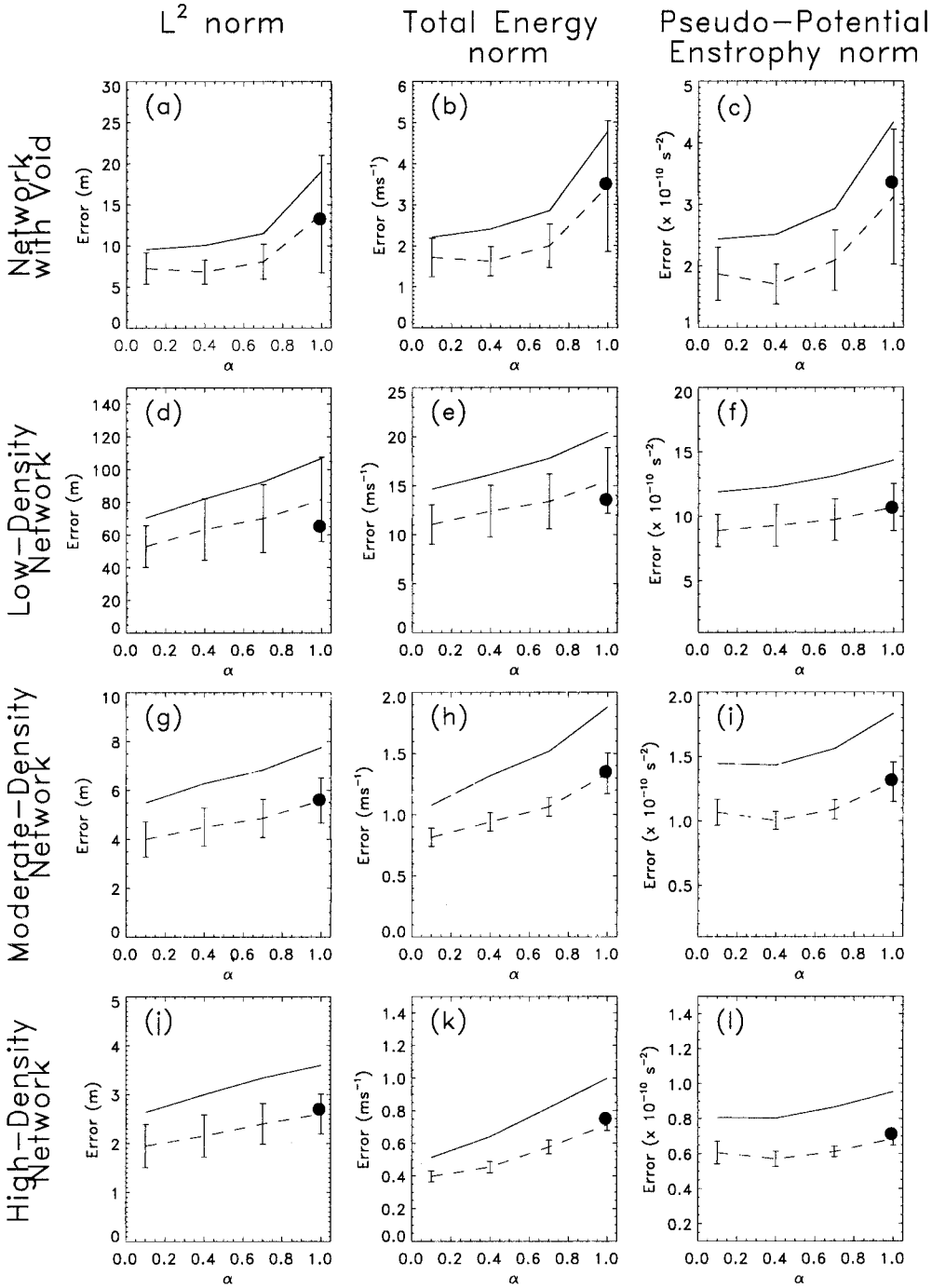

FIG. 3. Analysis errors as a function of norm, observational data density, and α. Solid lines denote average of individual member's errors over all $n = 25$ members and over time series, excluding initial adjustment period of 16 days. Dashed line indicates errors of ensemble mean, and error bars indicate ± 1 standard deviation from ensemble mean over the time series of ensemble mean errors. Errors of the 3DVAR control analysis are marked with a heavy dot.

over too short a period; in a longer integration, we expect $n = 50$ to have the same or smaller error than $n = 25$.

Strong spread–skill relationships were previously documented in HSM00 for a perturbed observation ensemble using a 3DVAR analysis scheme (equivalent to this experiment when $\alpha = 1.0$). Here, it was found that spread–skill correlations typically decreased as α decreased (not shown). We do not regard this as a problem with the EnKF approach, however. As indicated by the

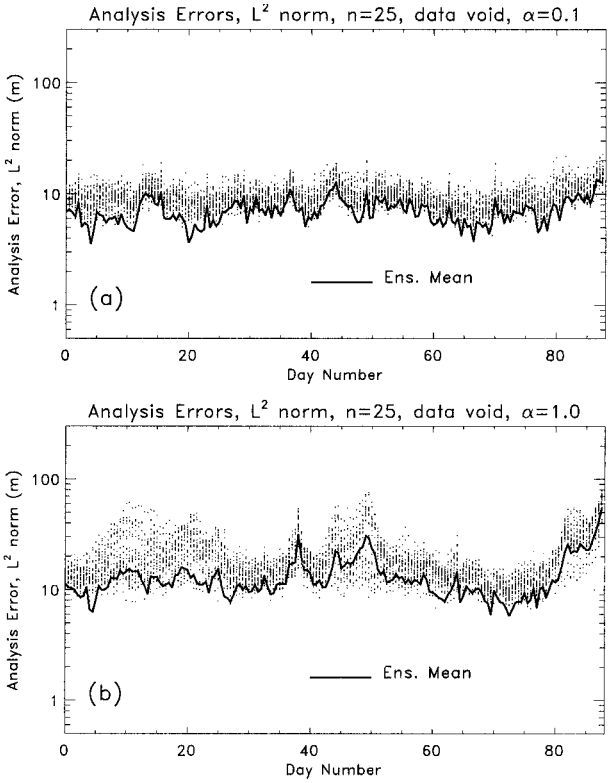

FIG. 4. Time series of analysis errors in the L^2 norm for the network with the data void ($n = 25$). Dots indicate errors of individual ensemble members, and the solid line indicates errors of the ensemble mean: (a) $\alpha = 0.1$; (b) $\alpha = 1.0$.

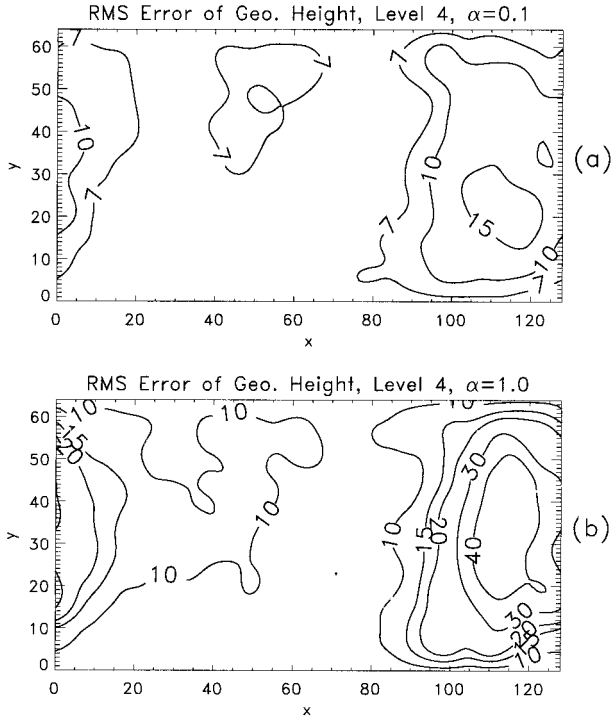

FIG. 5. Rms error of analyzed geopotential height (m) at level 4, averaged over all case days and ensemble members: (a) $\alpha = 0.1$; (b) $\alpha = 1.0$.

size of the error bars in Fig. 3, for $\alpha = 0.1$ there was much less variation with time in the analysis error characteristics. As indicated by Whitaker and Loughe (1998), the spread–skill relationship is a strong function of how much analysis errors vary with time. Hence, the decreased spread–skill relationship is a natural consequence of the EnKF reducing temporal variability in the magnitude of the analysis error.

We now consider a second metric of analysis quality, measuring the ability of the ensemble to sample from the distribution of plausible analysis states. For a properly constructed ensemble, low analysis error should be accompanied by uniformly distributed rank histograms (Hamill and Colucci 1997, 1998b; HSM00). The rank of an observation relative to a sorted n-member ensemble of forecasts should be equally likely to occur in any of the $n + 1$ possible ranks if the observation and ensemble sample the same underlying probability distribution. Hence, over many samples, a histogram of the ranks of observations relative to the sorted ensemble should be approximately uniform.

Rank histograms of potential temperature at the model top boundary (θ_T) are shown in Fig. 7 for $n = 25$. Histograms for other variables are not shown, since θ_T typically exhibited the worst nonuniformity. Notice the rank histograms in Fig. 7 are relatively uniform except

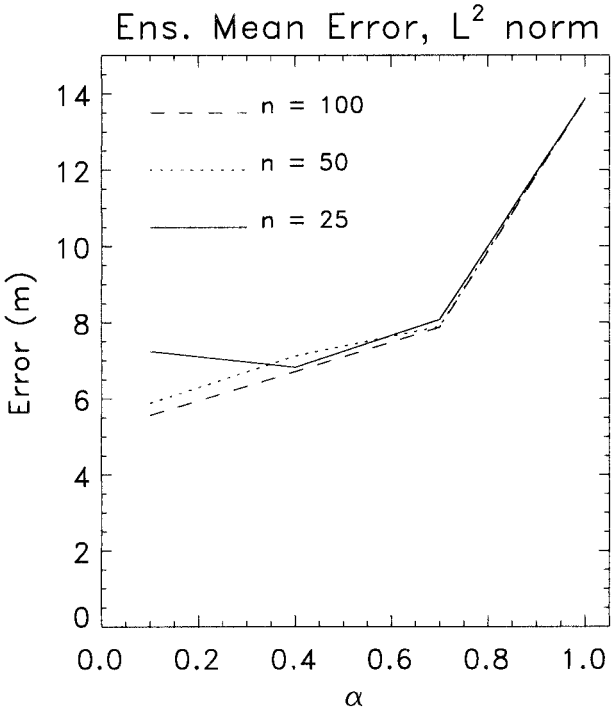

FIG. 6. Average ensemble mean analysis errors for network with data void for ensembles of size 25, 50, and 100 in the L^2 norm.

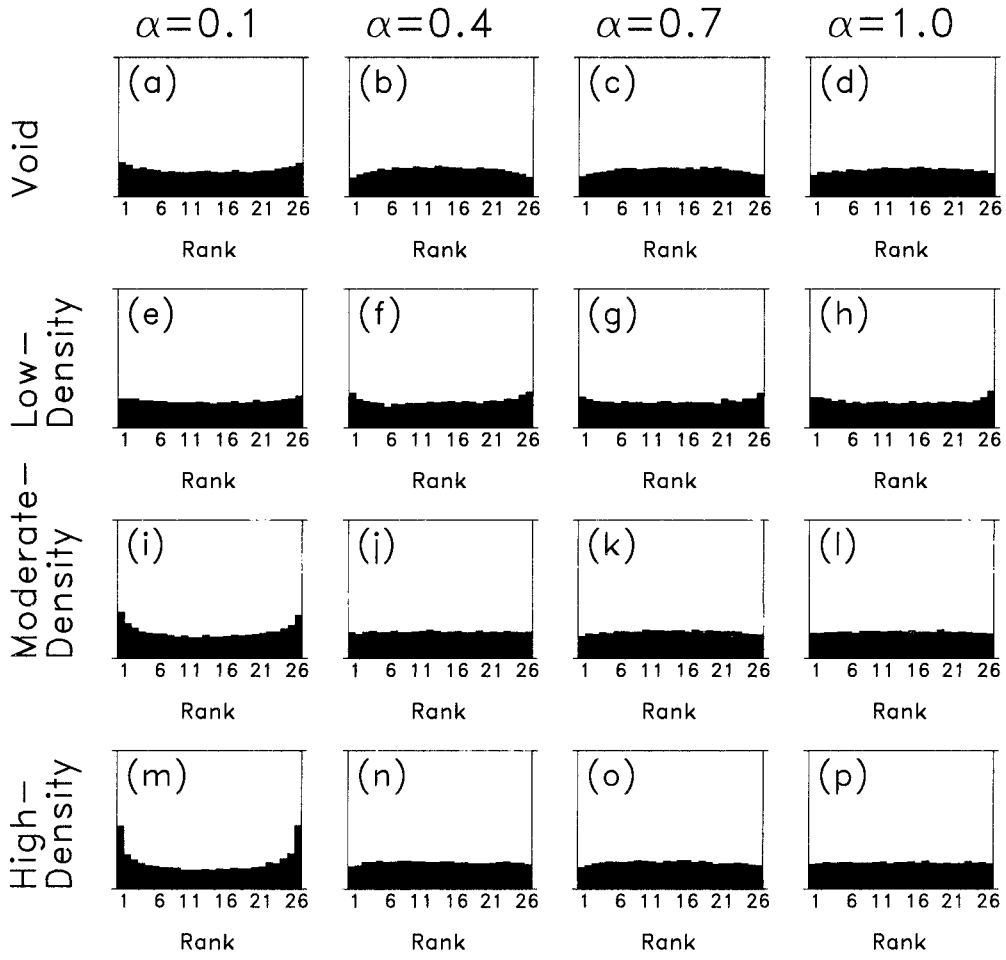

FIG. 7. Rank histograms for θ_T for various values of α and various observational networks; $n = 25$ members.

at $\alpha = 0.1$, where the extreme ranks are more highly populated. The nonuniformity of rank histograms for $\alpha = 0.1$ indicates there may be a problem with using small α in conjunction with smaller-sized ensembles. This is further illustrated in Fig. 8, which shows rank histograms for the network with the data void for ensembles of size $n = 50$ and $n = 100$. Problems with inappropriately high populations at the extreme ranks were corrected for these larger-sized ensembles. As is demonstrated below, with smaller ensembles, the ensemble-based covariances are more likely to produce spurious long distance correlations; these in turn induce inappropriately large analysis increments far from the original observation. We believe these occasional large corrections can produce a biased analysis that overpopulates the extreme ranks of the histogram.

Note also that the extent of nonuniformity of the rank histograms at $\alpha = 0.1$ are worse when there are more observations. We hypothesize that this is because the number of important degrees of freedom in the analysis increases when more observations are included. Hence, an ensemble of limited size does an increaingly poor job of sampling the relevant subspace as the density of observations is increased.

b. Single-observation analysis increments

Single-observation experiments are a typical benchmark for data assimilation schemes. Here, they are particularly useful for illustrating how analysis increments depend on α and n. Figures 9 and 10 show increments to u and v wind components induced by a single observation of the model level 7 ($\sim$320 hPa) zonal wind, which is 1 m s^{-1} higher than in the background ($n = 25$). To generate the increments, the data-void network's cycling was interrupted after 20 days and the single observation assimilated.

The resulting analysis increment for 3DVAR ($\alpha = 1.0$; Fig. 9) was relatively confined and produced u analysis increments in a dumbbell shape, similar to increments shown in PD92. As α decreased (Fig. 10), the location of the maximum increment was actually shifted to a few grid points east of the observation location, a negative u increment was found just south of the ob-

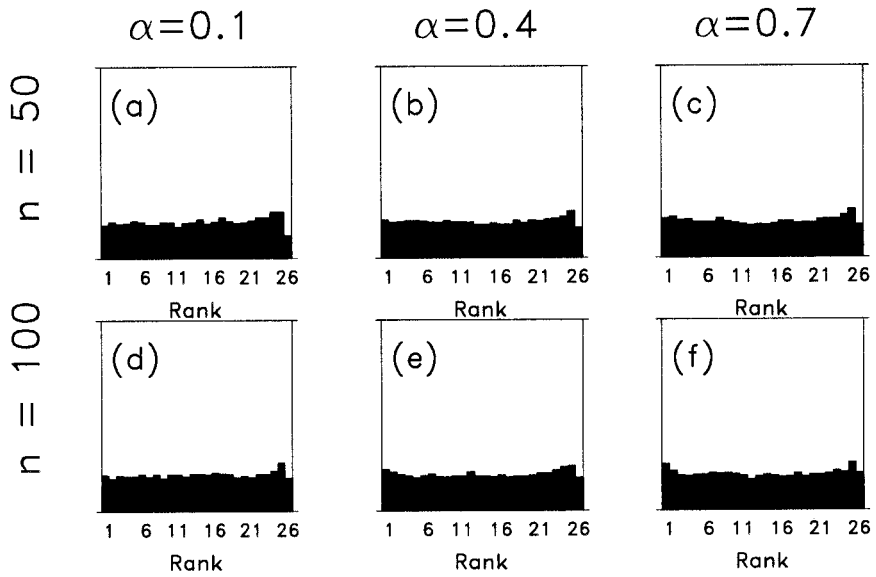

FIG. 8. Rank histograms for θ_T for various values of α using samples of size 25 from ensembles of size $n = 50$ and $n = 100$ for the network with the data void.

servation location, and a spatially complicated v increment was observed. The background error estimate generated from the ensemble correlation structure suggested that when stronger zonal winds were observed at that location, the entire core of the jet should be displaced farther north, weakening the zonal winds south of the observation.

There are also positive increments in Fig. 10 in the eastern part of the domain, far from the location of the observation. Increasing the ensemble size to 100 members (Fig. 11) demonstrates that these analysis increments far from the observation are largely spurious.

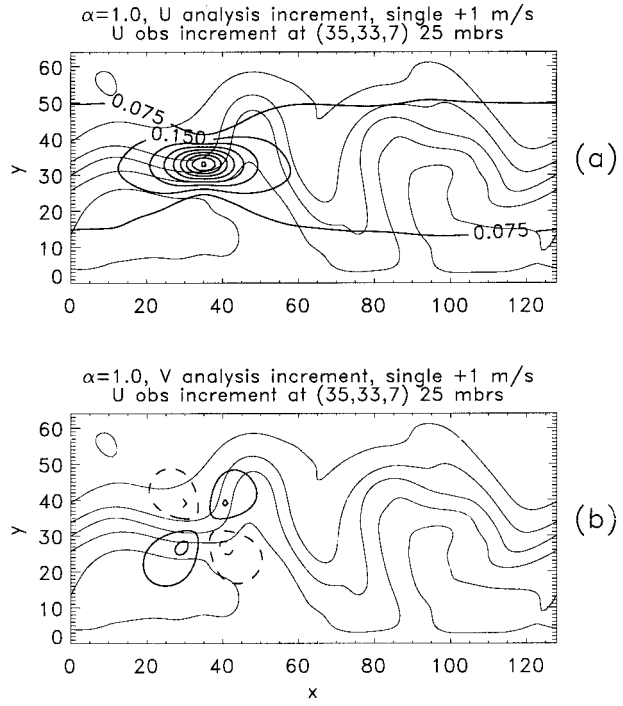

FIG. 9. The $\alpha = 1.0$ model level 7 analysis increments to (a) u wind component, and (b) v wind component induced by a $+1$ m s^{-1} single wind observation increment at model grid point (35, 33, 7) at day 62 of the experiment. Contours are every 0.075 m s^{-1} excluding 0.0 m s^{-1}. Dashed lines indicate negative values. Truth-run geopotential heights are overplotted in thin contours; $n = 25$ members.

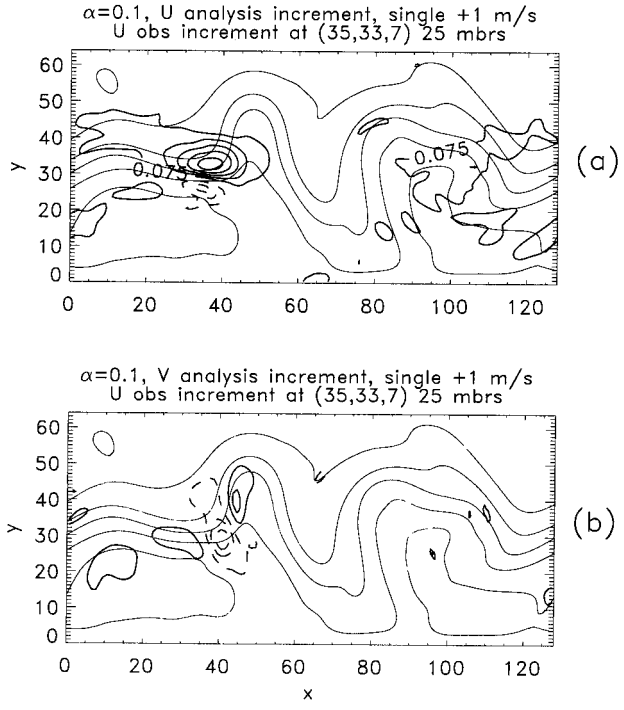

FIG. 10. As in Fig. 9 but for $\alpha = 0.1$.

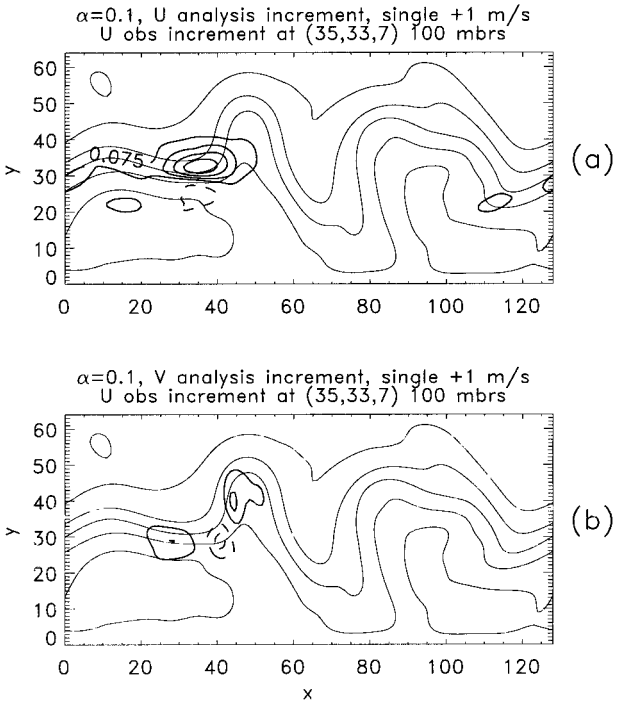

FIG. 11. As in Fig. 10 but for $n = 100$ members.

Additional ensemble members improve the background error covariance estimates and reduce the increments at distant points to near zero. Note also that the size of the maximum increment decreased as α was decreased and the ensemble size increased, consistent with having a better background error estimate.

Given that the accuracy of a small-sized ensemble may be degraded by spuriously large covariance estimates between far-distant locations, it may be desirable to filter the correlations generated by the ensemble, as previously suggested by P. L. Houtekamer (1999, personal communication). We are currently testing a digital smoothing filter with a Gaussian kernel designed by J. Purser, an extension of the filter described in Hayden and Purser (1995). We expect to describe the results of this filter in future work.

5. Ensemble forecast error characteristics

We briefly demonstrate that in this perfect-model scenario, an improved ensemble of analyses also results in improved probabilistic forecasts. Five-day forecasts were conducted from each of the $n = 25$ ensemble initial conditions for each of the 20 case days for the network with the data void. No filtering of covariances was performed here. Rank histograms for 1-, 3-, and 5-day forecasts are shown in Figs. 12a–p. The ensembles remain generally well calibrated throughout the subsequent forecast, though the extreme ranks are slightly more highly populated for $\alpha = 0.1$.

The quality of the forecasts were evaluated using Brier scores (Brier 1950; Wilks 1995), a measure of the mean-squared error of probabilistic forecasts. Here, we examine probabilities from the ensembles that the wind speed at model level 4 is greater than 60 m s^{-1}. The ensemble relative frequency is used to set probabilities; for example, if 5 of the 25 members indicate winds greater than 60 m s^{-1}, the probability is set to 20%. Table 1 shows Brier scores for 1-, 2-, 3-, and 5-day lead times for $\alpha = 0.1$, 0.4, 0.7, and 1.0. Brier scores are lowest (best) for $\alpha = 0.4$, with $\alpha = 0.1$ and $\alpha = 0.7$ nearly as accurate, and $\alpha = 1.0$ much worse in skill. The improvement of $\alpha = 0.4$, over $\alpha = 0.1$ and $\alpha = 0.7$, is not statistically significant using the test method outlined in Hamill (1999), but the improve over $\alpha = 1.0$ is significant (p value <0.001).

Another diagnostic for the skill of probabilistic weather forecasts is the relative operating characteristic, or ROC (Swets 1973; Mason 1982; Stanski et al. 1989). The application to ensemble forecasts is thoroughly described in HSM00. The ROC is a curve that indicates the trade-off between type I errors (incorrect rejection of null hypothesis) and type II (incorrect rejection of alternative hypothesis) as various sorted members of the ensemble are used as decision thresholds. Curves that are farther up and to the left (more hits, less false alarms) indicate better probabilistic forecasts. Figures 13a–d show the ROCs for various α's and lead times. The ROC scores are similar to the Brier scores, indicating a slight improvement of $\alpha = 0.4$ over $\alpha = 0.1$ and $\alpha = 0.7$, with $\alpha = 1.0$ much worse. Note that if measured in forecast lead time, the improvement of forecasts from $\alpha = 0.4$ over $\alpha = 1.0$ is approximately 1 day; that is, a day 2 probabilistic forecast at $\alpha = 0.4$ is as accurate as a day 1 forecast at $\alpha = 1.0$.

As demonstrated earlier, larger ensembles reduce analysis error. Does reduced error result in improved probabilistic forecasts? A larger ensemble can help in two ways; the larger sample naturally provides better estimates of forecast probabilities, and it enhances the accuracy of background error statisics estimated from the ensemble, thus producing a reduced-error ensemble of analyses. To isolate forecast improvements due only to the second aspect, probabilistic forecasts will be generated from 25-member subsets of the 50- and 100-member ensembles, so the ensemble size is equal in this comparison. Table 2 shows Brier scores as in Table 1 for sets of 25-member forecasts drawn from ensembles of size 25, 50, and 100. The reduced-error initial conditions in the larger ensembles result in reduced-error forecasts as well; the Brier scores for $n = 50$ are lower than for $n = 25$, and $n = 100$ is generally lower than for $n = 50$. ROC curves (not shown) are qualitatively similar to the Brier scores.

6. Conclusions

A prototype hybrid 3DVAR–ensemble Kalman filter analysis scheme was demonstrated here. This hybrid

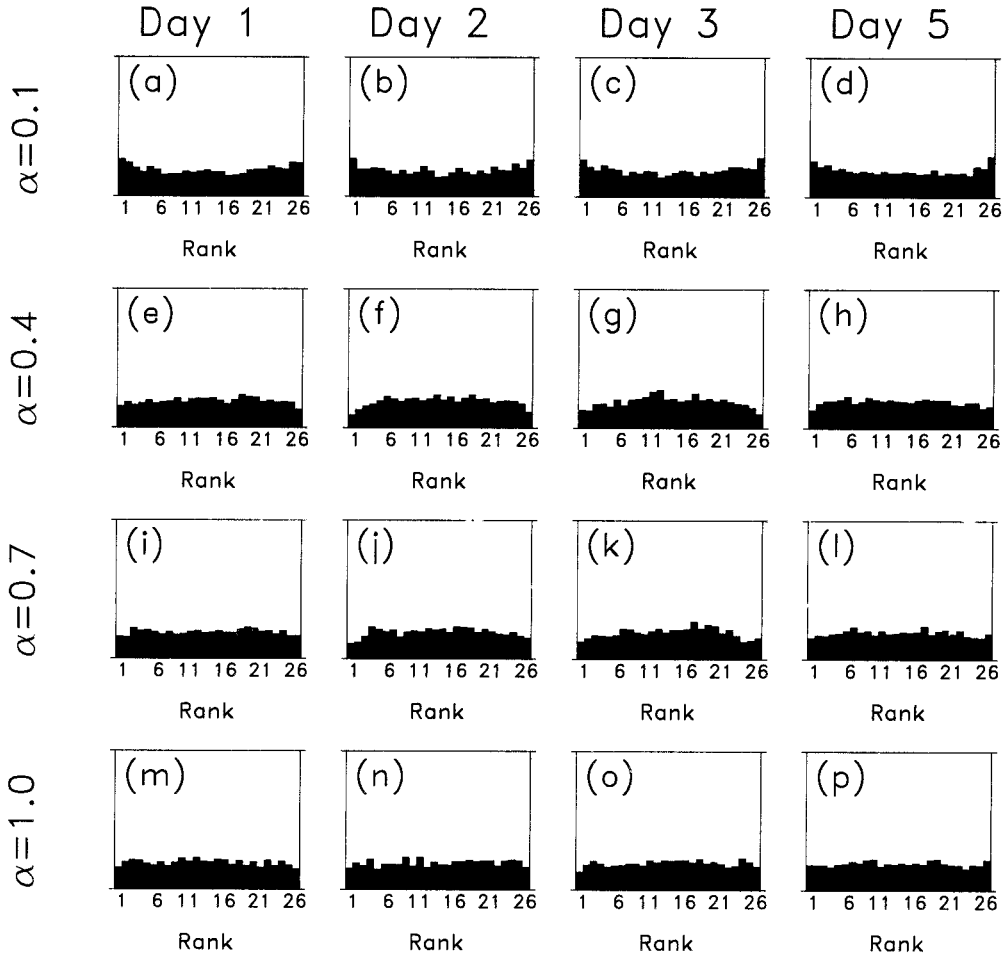

FIG. 12. Rank histograms for θ_T for forecasts using network with data void.

scheme was shown to produce a reduced-error set of analyses relative to using 3DVAR. This scheme allowed the relative weighting α between ensemble-based covariances and 3DVAR covariances to be adjusted according to the size of the ensemble to minimize the subsequent error characteristics, to control problems with spurious analysis changes far from the observation location, and to preserve uniformity of rank. Generally, low to moderate α's (relatively equal weightings of ensemble-based and 3DVAR background error covariances) were shown to be optimal for small ensembles, and small α's (primarily ensemble-based covariances) are optimal for larger ensembles. Using higher α apparently reduces both appropriate and inappropriate long distance correlations but forces correlations nearby the observations to be inappropriately isotropic. Use of $\alpha < 0.1$ was not described here, but caused serious problems for small ensembles.

TABLE 1. Brier scores for Pr(wind speed >60 m s^{-1}), for $n = 25$ members and various values of α.

Day	$\alpha = 0.1$	$\alpha = 0.4$	$\alpha = 0.7$	$\alpha = 1.0$
1	0.0080	0.0081	0.0094	0.0174
2	0.0151	0.0122	0.0147	0.0258
3	0.0285	0.0246	0.0282	0.0414
5	0.0558	0.0518	0.0601	0.0813

Networks with fewer observations showed more improvement over the 3DVAR control than networks with higher densities of observations. More ensemble members generally reduced the analysis errors by providing a better estimate of flow-dependent background error covariances.

The improved set of analyses generated an improved ensemble of weather forecasts. When tested on the network with the data void, use of improved covariances resulted in about a 1-day gain in forecast lead time (e.g., optimized day 3 hybrid probabilisitic forecasts were as skillful as a perturbed observation day 2 forecast).

Another appealing trait is that the algorithm is highly parallelizable; in principle, individual member's analyses and forecasts may be computed in parallel on individual processors. If the computer architecture and speed prohibit the use of a large ensemble, some incre-

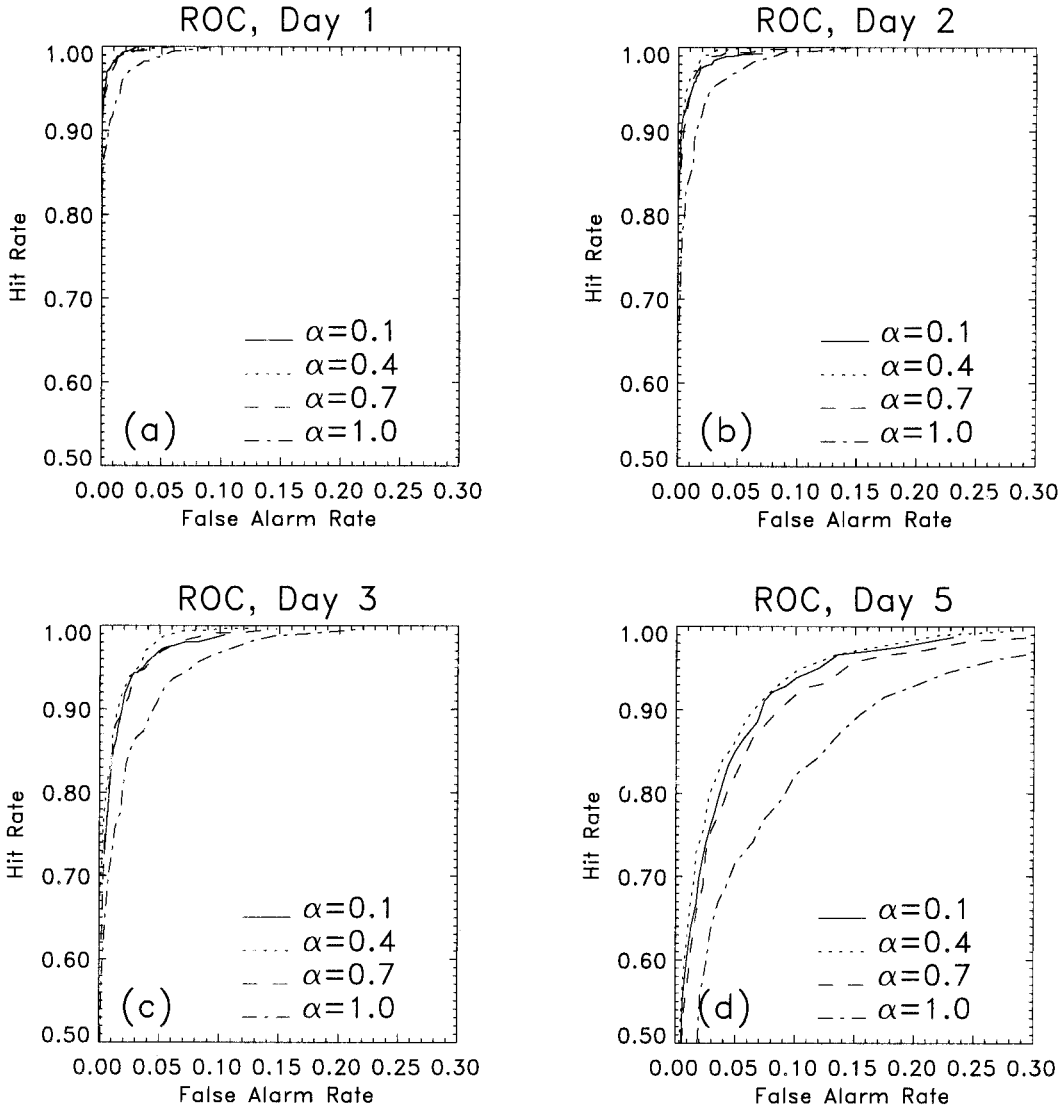

FIG. 13. Relative operating characteristic curves for P (wind speed > 60 m s^{-1}) for various values of α: (a) 1-, (b) 2-, (c) 3-, and (d) 5-day forecasts.

mental improvement over 3DVAR may still be possible with a smaller ensemble since the flow-dependent covariances may still be used, but weighted less heavily.

This study was conducted under perfect-model assumptions with a simple quasigeostrophic channel model. The quantitative improvements noted here may not be realizable with a primitive equation model or under circumstances where model error is nonnegligible. However, this hybrid scheme should be an appealing choice for further testing.

TABLE 2. Brier scores for Pr(wind velocity >60 m s^{-1}) for $\alpha = 0.1$ and 25-member ensembles taken from ensembles of $n = 25$, 50, and 100.

Day	$n = 25$	$n = 50$	$n = 100$
1	0.0080	0.00636	0.00658
2	0.0151	0.0112	0.0107
3	0.0285	0.0203	0.0189
5	0.0558	0.0426	0.0392

It is worth considering what would be required for operational implementation of this scheme. Because the cost function of the hybrid scheme differs from that of 3DVAR only in its treatment of the background covariance matrix $\mathbf{B}$, much of the machinery of an existing operational 3DVAR scheme would carry over unchanged to the hybrid. In fact, the only modification necessary in the 3DVAR scheme used here is to include the sample covariance $\mathbf{P}^b$ in the subroutine that calculates the product $\mathbf{Bx}$ for arbitrary $\mathbf{x}$; this is straightforward and computationally inexpensive. Operational 3DVAR schemes, however, use various preconditioners, such as those based on $\mathbf{B}^{1/2}$, to accelerate the minimi-

zation. When $\mathbf{P}^b$ is included in $\mathbf{B}$, these preconditioners are not available and futher research will be required to find suitable (and feasible) preconditing techniques.

An ensemble that contains useful information on the statistics of very short-range (6h) forecasts is also clearly necessary. The best approach to constructing such an ensemble remains a point of active debate, but, based on experiments with the QG model (HSM00), we view the PO ensemble framework as a promising candidate. A very similar methodology, including some allowance for model errors, is operational at the Canadian Meteorological Centre (Houtekamer et al. 1996).

Computationally, the hybrid scheme amounts to performing n 3DVAR analyses, if we assume that condition number for hybrid scheme can be reduced to that for 3DVAR. One potential mitigating factor is the ease with which these n analyses could be parallelized. Moreover, the n analyses need not be completed in the short window usually allowed for the calculation of the operation 3DVAR analysis; instead, a single hybrid analysis could be performed in that window (to provide initial conditions for a single ''control'' forecast), while the other $n - 1$ were spread over the interval between analysis times.

We also recommend comparisons against other data assimilation techniques such as 4DVAR and other implementations of the EnKF. For comparisons with other EnKF schemes, we can make some informed guesses about when our hybrid may be more or less appropriate than other EnKF approaches. If a center continues to use an intermittent data assimilation approach and the cost of code modification is a primary concern, this hybrid approach is worth consideration, since modifications to 3DVAR are minimal. On the other hand, this scheme is probably less suitable for continuous assimilation of asynoptic observations. There, perhaps a scheme like HM98 will prove to be more computationally efficient, since it is designed to update the analysis only in the region of new observations.

Acknowledgments. We thank Peter Houtekamer, Craig Bishop, and Francois Van den Berghe for their informal reviews of this manuscript, and Dale Barker, Jim Purser, and an anonymous reviewer for their formal reviews.

REFERENCES

Anderson, J. L., 1996: A method for producing and evaluating probabilistic forecasts from ensemble model integrations. *J. Climate,* **9,** 1518–1530.

Brier, G. W., 1950: Verification of forecasts expressed in terms of probabilities. *Mon. Wea. Rev.,* **78,** 1–3.

Buizza, R., M. Miller, and T. N. Palmer, 1999: Stochastic simulation of model uncertainty in the ECMWF ensemble prediction system. *Quart. J. Roy. Meteor. Soc.,* **125,** 2887–2908.

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

Courtier, P., J. N. Thépaut, and A. Hollingsworth, 1994: A strategy for operational implementation of 4DVAR, using an incremental approach. *Quart. J. Roy. Meteor. Soc.,* **120,** 1367–1387.

Daley, R., 1991: *Atmospheric Data Analysis.* Cambridge University Press, 457 pp.

Epstein, E. S., 1969: Stochastic dynamic prediction. *Tellus,* **21,** 739–759.

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

——, 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.

Fisher, M., and P. Courtier, 1995: Estimating the covariance matrices of analysis and forecast error in variational data assimilation. Research Department Tech. Memo. 220, ECMWF, 28 pp. [Available from ECMWF, Shinfield Park, Reading, Berkshire RG2 9AX, United Kingdom.]

Gandin, L. S., 1963: *Objective analysis of meteorological fields.* Gidrometeorologicheskoe Izdatelstvo; English translation, Israel Program for Scientific Translations, 242 pp.

Ghil, M., 1997: Advances in sequential estimation for atmospheric and oceanic flows. *J. Meteor. Soc. Japan,* **75,** 289–304.

Hamill, T. M., 1999: Hypothesis tests for evaluating numerical precipitation forecasts. *Wea. Forecasting,* **14,** 155–167.

——, and S. J. Colucci, 1997: Verification of Eta–RSM short-range ensemble forecasts. *Mon. Wea. Rev.,* **125,** 1312–1327.

——, and ——, 1998a: Perturbations to the land surface condition in short-range ensemble forecasts. Preprints, *12th Conf. on Numerical Weather Prediction,* Phoenix, AZ, Amer. Meteor. Soc., 273–276.

——, and ——, 1998b: Evaluation of Eta–RSM ensemble probabilistic precipitation forecasts. *Mon. Wea. Rev.,* **126,** 711–724.

——, C. Snyder, and R. E. Morss, 2000: A comparison of probabilistic forecasts from bred, singular-vector, and perturbed observation ensembles. *Mon. Wea. Rev.,* **128,** 1835–1851.

Hayden, C. M., and R. J. Purser, 1995: Recursive filter objective analysis of meteorological fields: Applications to NESDIS operational processing. *J. Appl. Meteor.,* **34,** 3–15.

Houtekamer, P. L., and J. Derome, 1995: Methods for ensemble prediction. *Mon. Wea. Rev.,* **123,** 2181–2196.

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

——, J. Derome, H. Ritchie, and H. L. Mitchell, 1996: A system simulation approach to ensemble prediction. *Mon. Wea. Rev.,* **124,** 1225–1242.

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,** 181–189.

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

Leith, C. E., 1974: Theoretical skill of Monte Carlo forecasts. *Mon. Wea. Rev.,* **102,** 409–418.

Lorenc, A. C., 1981: A global three-dimensional multivariate statistical interpolation system. *Mon. Wea. Rev.,* **109,** 701–721.

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

Lorenz, E. N., 1963: Deterministic nonperiodic flow. *J. Atmos. Sci.,* **20,** 130–141.

——, 1969: The predictability of a flow which possesses many scales of motion. *Tellus,* **21,** 289–307.

Marshall, J., and F. Molteni, 1993: Toward a dynamical understanding of planetary flow regimes. *J. Atmos. Sci.,* **50,** 1792–1818.

Mason, I., 1982: A model for assessment of weather forecasts. *Aust. Meteor. Mag.,* **30,** 291–303.

Mitchell, H. L., and P. L. Houtekamer, 2000: An adaptive ensemble Kalman filter. *Mon. Wea. Rev.,* **128,** 416–433.

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.

Morss, R. E., 1999: Adaptive observations: Idealized sampling strat-

egies for improving numerical weather prediction. Ph.D dissertation, Massachusetts Institute of Technology, 225 pp. [Available from UMI Dissertation Services, P.O. Box 1346, 300 N. Zeeb Rd., Ann Arbor, MI 48106-1346.]

Parrish, D. F., and J. C. Derber, 1992: The National Meteorological Center's spectral statistical-interpolation analysis system. *Mon. Wea. Rev.,* **120,** 1747–1763.

Press, W. H., S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery, 1992: *Numerical Recipes in Fortran.* 2d ed. Cambridge University Press, 963 pp.

Schlatter, T. W., 1975: Some experiments with a multivariate statistical objective analysis scheme. *Mon. Wea. Rev.,* **103,** 246–257.

——, F. H. Carr, R. H. Langland, R. E. Carbone, N. A. Crook, R. W. Daley, J. C. Derber, and S. L. Mullen, 1999: A five-year plan for research related to the assimilation of meteorological data. NCAR Tech Note 443, 45 pp. [Available from UCAR Communications, P.O. Box 3000, Boulder, CO 80307-3000.]

Schubert, S. D., and M. Suarez, 1989: Dynamical predictability in a simple general circulation model: Average error growth. *J. Atmos. Sci.,* **46,** 353–370.

Smolarkiewicz, P. K., and L. G. Margolin, 1994: Variational solver for elliptic problems in atmospheric flows. *Appl. Math Comput. Sci.,* **4,** 527–551.

Stanski, H. R., L. J. Wilson, and W. R. Burrows, 1989: Survey of common verification methods in meteorology. Research Rep. 89-5, Environment Canada, 114, pp. [Available from Atmospheric Environment Service, Forecast Research Division, 4905 Dufferin St., Downsview, ON M3H 5T4, Canada.]

Swets, J. A., 1973: The relative operating characteristic in psychology. *Science,* **182,** 990–999.

Talagrand, O., 1997: Assimilation of observations: An introduction. *J. Meteor. Soc. Japan,* **75,** 191–209.

Thépaut, J.-N., R. N. Hoffman, and P. Courtier, 1993: Interactions of dynamics and observations in a four-dimensional variational assimilation. *Mon. Wea. Rev.*, **121,** 3393–3414.

Thompson, P., 1969: Reduction of analysis error through constraints of dynamical consistency. *J. Appl. Meteor.,* **8,** 739–742.

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 and the breeding method. *Mon. Wea. Rev.,* **125,** 3297–3319.

Tribbia, J. J., and D. P. Baumhefner, 1988: The reliability of improvements in deterministic short-range forecasts in the presence of initial state and modeling deficiencies. *Mon. Wea. Rev.,* **116,** 2276–2288.

van Leeuwen, P. J., 1999: Comment on ''Data assimilation using an ensemble Kalman filter technique.'' *Mon. Wea. Rev.,* **127,** 1374–1377.

Whitaker, J. S., and A. F. Loughe, 1998: The relationship between ensemble spread and ensemble mean skill. *Mon. Wea. Rev.,* **126,** 3292–3302.

Wilks, D. S., 1995: *Statistical Methods in the Atmospheric Sciences: An Introduction.* Academic Press, 467 pp.