A Comparison of Probabilistic Forecasts from Bred, Singular-Vector, and Perturbed Observation Ensembles

THOMAS M. HAMILL AND CHRIS SNYDER

National Center for Atmospheric Research, Boulder, Colorado*

REBECCA E. MORSS

Massachusetts Institute of Technology, Cambridge, Massachusetts

(Manuscript received 26 February 1999, in final form 16 August 1999)

ABSTRACT

The statistical properties of analysis and forecast errors from commonly used ensemble perturbation methodologies are explored. A quasigeostrophic channel model is used, coupled with a 3D-variational data assimilation scheme. A perfect model is assumed.

Three perturbation methodologies are considered. The breeding and singular-vector (SV) methods approximate the strategies currently used at operational centers in the United States and Europe, respectively. The perturbed observation (PO) methodology approximates a random sample from the analysis probability density function (pdf) and is similar to the method performed at the Canadian Meteorological Centre. Initial conditions for the PO ensemble are analyses from independent, parallel data assimilation cycles. Each assimilation cycle utilizes observations perturbed by random noise whose statistics are consistent with observational error covariances. Each member's assimilation/forecast cycle is also started from a distinct initial condition.

Relative to breeding and SV, the PO method here produced analyses and forecasts with desirable statistical characteristics. These include consistent rank histogram uniformity for all variables at all lead times, high spread/skill correlations, and calibrated, reduced-error probabilistic forecasts. It achieved these improvements primarily because 1) the ensemble mean of the PO initial conditions was more accurate than the mean of the bred or singular-vector ensembles, which were centered on a less-skilful control initial condition—much of the improvement was lost when PO initial conditions were recentered on the control analysis; and 2) by construction, the perturbed observation ensemble initial conditions permitted realistic variations in spread from day to day, while bred and singular-vector perturbations did not. These results suggest that in the absence of model error, an ensemble of initial conditions performs better when the initialization method is designed to produce random samples from the analysis pdf. The perturbed observation method did this much more satisfactorily than either the breeding or singular-vector methods.

The ability of the perturbed observation ensemble to sample randomly from the analysis pdf also suggests that such an ensemble can provide useful information on forecast covariances and hence improve future data assimilation techniques.

1. Introduction

Numerical weather forecasts exhibit deterministic chaos (Lorenz 1963); small errors in the initial condition can grow exponentially and eventually render a forecast useless. Since perfect weather forecasts are thus unattainable, forecast information is more appropriately expressed in a probabilistic framework, whereby the user is provided with information on the likelihood of a range of forecast events. Such probabilistic forecasts are increasingly desired by a wide range of forecast users (e.g., Fritsch et al. 1998).

Ideally, probabilistic forecasts could be generated by evolving the probability density function of the forecast [the ''Liouville'' equation; Ehrendorfer (1994a,b)]. Such integrations for low-order dynamical systems typically reveal the initially sharp probability density function becoming increasingly diffuse with time. For numerical weather prediction (NWP) models, however, this approach is computationally unfeasible.

A computationally tractable method to approximate the evolution of the probability density function (pdf) is through *ensemble forecasting.* Here, a limited number of forecasts are generated by integrating a numerical forecast model forward in time multiple times from dis-

* NCAR is supported by the National Science Foundation.

Corresponding author address: Dr. Thomas M. Hamill, NCAR/MMM, P.O. Box 3000, Boulder, CO 80301.
E-mail: hamill@ucar.edu

tinct and plausible initial conditions (Leith 1974). The mean of this ensemble of forecasts acts as a nonlinear filter, averaging out the nonpredictable aspects of the flow that vary from member to member and leaving the aspects that tend to agree. Further, ideally, the relative frequency of forecast model outcomes may be used to generate calibrated probabilistic weather forecasts.

The best method for specifying a set of initial conditions for ensembles of forecasts is still actively debated, and pioneering efforts have focused more on the dynamical characteristics of the initial condition than on the statistical aspects. These ''dynamically constrained'' techniques add to a control forecast perturbations that will grow or have grown rapidly. The presumption is that if a limited-size ensemble must be used, forecasts from these perturbations are already spanning the most important subspaces of the ensemble forecast. To this end, the European Centre for Medium-Range Weather Forecasts (ECMWF) has adopted a method dubbed the ''singular vector,'' or SV, approach (Buizza and Palmer 1995; Molteni et al. 1996). Singular-vector perturbations are designed to maximize growth over a finite time interval (typically, 2 days). The SV approach uses an adjoint (Errico 1997) and linear tangent of the forecast model to determine these growing directions, and the perturbations are designed in this subspace. The breeding technique, used at the National Centers for Environmental Prediction (NCEP; Toth and Kalnay 1993, 1997), generates perturbations in directions where past forecast errors have grown rapidly. This is achieved by periodically renormalizing differences between member forecasts, somewhat analogous to the procedure used to determine the Lyapunov exponents and vectors of a dynamical system (Wolf et al. 1985). Legras and Vautard (1996) show that the breeding and SV techniques are related through notions of ''backward'' and ''forward'' Lyapunov vectors.

Houtekamer and Derome (1995, hereafter HD95) introduced a Monte Carlo perturbation methodology that we shall refer to hereafter as the perturbed observation (PO) method. This method does not produce dynamically constrained perturbed initial conditions; rather, it is designed to approximate a random sample from the probability distribution for the true state at the same analysis time t_a, given[1] all available observations for $t < t_a$. We will refer to this distribution as the analysis pdf. To produce such samples, multiple, parallel data assimilation cycles are performed, and the method stochastically simulates errors in both the observations and the first guess. In the context of this perfect model experiment, for example, the first guess and the observations thus are treated probabilistically. The PO technique, using an ensemble Kalman filter for data assimilation, has been shown to produce a random sample from the correct distribution in the case that observational errors are Gaussian, dynamics are linear, and the ensemble size is large (Burgers et al. 1998). In the case examined here, dynamics are nonlinear, the PO technique is approximate (due to use of a three-dimensional variational assimilation scheme), and the ensemble size will be limited.

The relative merits of dynamically constrained versus Monte Carlo methods are still unclear. Comparisons of forecasts from the different operational forecast centers are not particularly illuminating, since in addition to different perturbation methodologies, the different centers use different analysis schemes, different forecast models running at different resolutions, and differently sized ensembles. Experiments using the same forecast model provide some perspective. A comparison of the use of dynamically constrained (SV and bred) versus unconstrained (e.g., Mullen and Baumhefner 1989) perturbations was explored by Anderson (1997), who found that more realistic forecasts could be obtained for the Lorenz (1963) model from random perturbations than from either of the dynamically constrained techniques. Evidence from operational numerical weather prediction models may suggest just the opposite (Toth and Kalnay 1993, 1997). HD95 found little difference in the skill of ensemble mean forecasts from bred, PO, and SV methodologies using a T21L3 quasigeostrophic (QG) model.

Our intent in this article is to extend the HD95 comparison of perturbation methodologies. As in HD95, we will compare ensemble-forecast characteristics produced from PO, bred, and (approximate) SV ensembles. Rather than considering the accuracy of the ensemble mean as in HD95, here we explore other aspects of quality, evaluating them using rank histograms (Anderson 1996; Hamill and Colucci 1997, 1998a), forecast dispersion, spread–skill relationships, and the accuracy of subsequent probabilistic forecasts. We will generate approximate analogs to current implementations of the breeding and SV techniques and compare them with the PO technique.

An important ancillary result of this paper is that the mean of the PO analyses may have lower rms error than the control analysis (i.e., ensemble averaging is beneficial, even at the analysis time). While we regard this property as a potentially important benefit of the PO technique, it is also of interest to understand whether the PO technique has other desirable properties beyond its better mean. Thus, we also construct an alternative version of the PO ensembles in which the PO ensemble initial conditions are recentered on a control initial condition. This also permits more ready comparison with the results in HD95.

All experiments here are conducted using a quasigeostrophic channel model in a perfect-model framework; the same model is used to generate both the reference, or ''true solution,'' and the forecasts. A three-

[1] In general, this distribution also depends on the forecast model, the analysis scheme, and their errors.

TABLE 1. Approximate model levels (mb).

Model level	Pressure (mb)
0 (θ_b)	1000
1	917
2	771
3	648
4	545
5	458
6	385
7	324
8	272
9 (θ_t)	250

TABLE 2. Model parameters: U is the maximum velocity of the jet in the zonal state toward which the solution is relaxed; N is the Brunt–Väisälä frequncy; f is the Coriolis parameter; β is the meridional gradient of f; ν is the coefficient for fourth-order horizontal diffusion; K is the vertical eddy viscosity assumed in the Ekman pumping; and τ is the relaxation time.

$U = 60$ m s^{-1}	$\nu = 1.24 \times 10^{15}$ m^4 s^{-1}
$N = 1.13 \times 10^{-2}$ s^{-1}	$K = 5$ m^2 s^{-1}
$f = 10 - 4$ s^{-1}	$\tau = 20$ days
$\beta = 1.6 \times 10^{-11}$ m^{-1} s^{-1}	

dimensional variational (3DVAR) data assimilation system is used (Parrish and Derber 1992, hereafter PD92), with simulated radiosondes assimilated every 12 h. A reference control forecast is generated, and the bred and SV perturbations are centered on this forecast. The PO forecasts are then generated, and analyses and forecasts are compared after an initial adjustment period.

Admittedly, the assumption of a perfect forecast model is not a realistic analog for actual numerical weather prediction, where model error may be significant or even dominant. However, the use of a perfect model permits examination of perturbation methodologies in a manner where all other complications are eliminated.

This paper is organized as follows: section 2 provides a brief review of the quasigeostrophic model, the data assimilation technique, and the observational network to be used. Section 3 describes the perturbation strategies as implemented in this model. Section 4 compares analyses and subsequent forecasts from the bred, SV, and PO techniques and explores some of the forecast problems. Section 5 provides a discussion of the results, and section 6 summarizes the results and implications.

2. Forecast model and data assimilation

a. The quasigeostrophic channel model

All experiments here were conducted in a perfect-model framework using the quasigeostrophic model used in Rotunno and Bao (1996) and Morss (1999). This is a midlatitude, beta-plane, finite-difference, channel model that is periodic in x (east–west), has impermeable walls on the north and south boundaries, and rigid lids at the top and bottom. 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. There is no stationary asymmetric forcing in the model such as land/sea contrasts or terrain. For these experiments, the domain is 16 000 $\times$ 8000 $\times$ 9 km; there are 129 grid points east–west, 65 north–south, and eight interior levels, with additional staggered top and bottom levels (at $z = 0$, 9 km) at which potential temperature is specified (Table 1). For these tests, the model performs 200 time steps per day. Additional model parameters are given in Table 2. Sample output of midtropospheric PV and geopotential height for three sequential days are illustrated in Fig. 1.

To measure the magnitude of perturbations or errors, three norms will be used here, the L^2 norm, the total energy norm, and the enstrophy norm. Given a geopotential perturbation Φ', PV perturbation q', and n model grid points, the L^2 norm is defined as

$$\|\cdot\|_{L^2} = g^{-1}n^{-1/2}\left[\sum_{j=1}^{n}(\Phi_j')^2\right]^{1/2}, \tag{1}$$

where g is the gravitational constant (9.8 m s^{-1}) and the subscript j denotes a gridpoint index. The energy norm and potential enstrophy norm (hereafter referred to as simply the ''enstrophy norm'') are defined as

$$\|\cdot\|_{\text{energy}} = f^{-1}n^{-1/2} \times \left\{\sum_{j=1}^{n}\left[\left(\frac{\partial\Phi'}{\partial x}\right)_j^2 + \left(\frac{\partial\Phi'}{\partial y}\right)_j^2 + \frac{f^2}{N^2}\left(\frac{\partial\Phi'}{\partial z}\right)_j^2\right]\right\}^{1/2} \tag{2}$$

and

$$\|\cdot\|_{\text{enstrophy}} = n^{-1/2}\left(\sum_{j=1}^{n} q_j'^2\right)^{1/2}. \tag{3}$$

Each norm emphasizes different scales of motion; the L^2 norm emphasizes errors in the larger scales, and the enstrophy norm the errors in the smaller scales.

b. Observational network

All forecast experiments were carried out using the observational network shown in Fig. 2. This network configuration was chosen to mimic roughly some of the characteristics of the current radiosonde network. Specifically, we introduced a data void for the eastern third of the domain to simulate a poorly observed oceanic region. Observation locations in the data-rich area were selected sequentially and randomly, using a one-dimensional Latin square algorithm (Press et al. 1992) that enforces a minimum distance between observations. The observational data density was specified so the error of the control analysis (see section 2d) was ~7% of the model climatological rms error, measured in the L^2 norm. This magnitude of analysis error broadly agrees with that of current analysis systems (e.g., Kalnay et al.

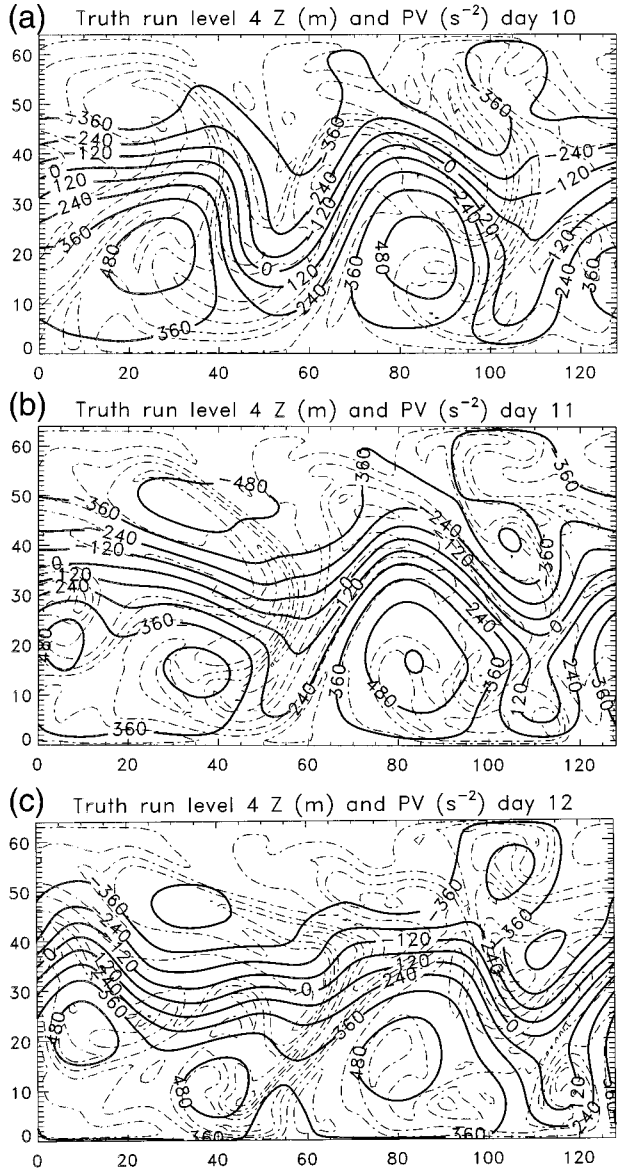

FIG. 1. Illustration of truth run fields of PV (dashed) and geopotential height (solid) at model level 4 on sequential days. Units are m for geopotential height (Z) and s^{-2} for PV: (a) day 10, (b) day 11, and (c) day 12.

1996). As will be shown, however, analysis errors for this observation network vary more in time than analysis errors in operational models.

In these experiments, all observations were presumed to be radiosondes (raob's), with observational error characteristics taken from PD92 and vertical observational error correlations from Bergman (1979). Further details and the specific covariance matrices used are shown in Morss (1999). For simplicity, observations were required to be located at model grid points, and representativeness error is subsumed into the observational error covariances.

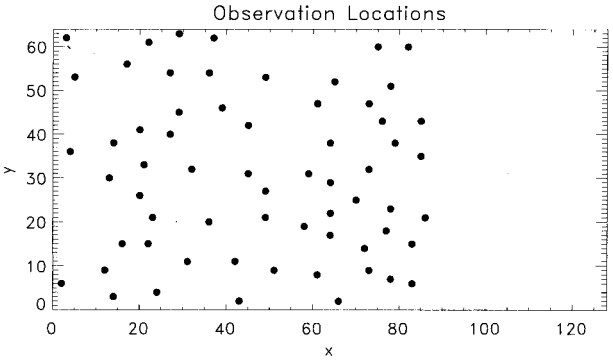

FIG. 2. Observational network used for forecast experiments.

c. Data assimilation methodology

All experiments employ a 3DVAR assimilation scheme described in detail in Morss (1999). This scheme assumes, following PD92, background error covariances that are diagonal in spectral space; more specifically, the scheme assumes that, if the background errors in PV were expanded in horizontal trigonometric series, the coefficients for each wavenumber pair and each model level would be independent. In practice, we tuned the covariances for each wavenumber to be consistent with the 12-h forecast errors by calculating forecast error statistics over a long analysis/forecast cycle, modifying the covariances and repeating the process until there was little change in the covariances.

d. The true solution and the control analyses

Our comparisons assume a perfect model; thus, the true solution was computed using the same model at the same resolution as is used for forecasts. The true solution began from a localized disturbance on the specified zonal state used for relaxation of PV. The QG model was then integrated for 300 days; the first 200 days of this integration, during which the solution is approaching a turbulent statistical equilibrium, were discarded, and the subsequent 90 days compose the true solution employed in our experiments.

Given the true solution, a series of control analyses and forecasts were made as follows: the first analysis was simply the true solution contaminated by random noise. Every 12 h thereafter, another analysis was made with the 3DVAR scheme by assimilating a set of soundings from the network of Fig. 2 and using the previous 12-h forecast as background. We call these (imperfect) soundings, which are incorporated into the control analysis, the control observations. These soundings were produced at each observation location by adding random error to soundings from the true solution at the appropriate time. The random observation errors were generated consistent with the observation error covariances given in section 2b; this was achieved by multiplying random, normally distributed $N(0, 1)$ numbers by the

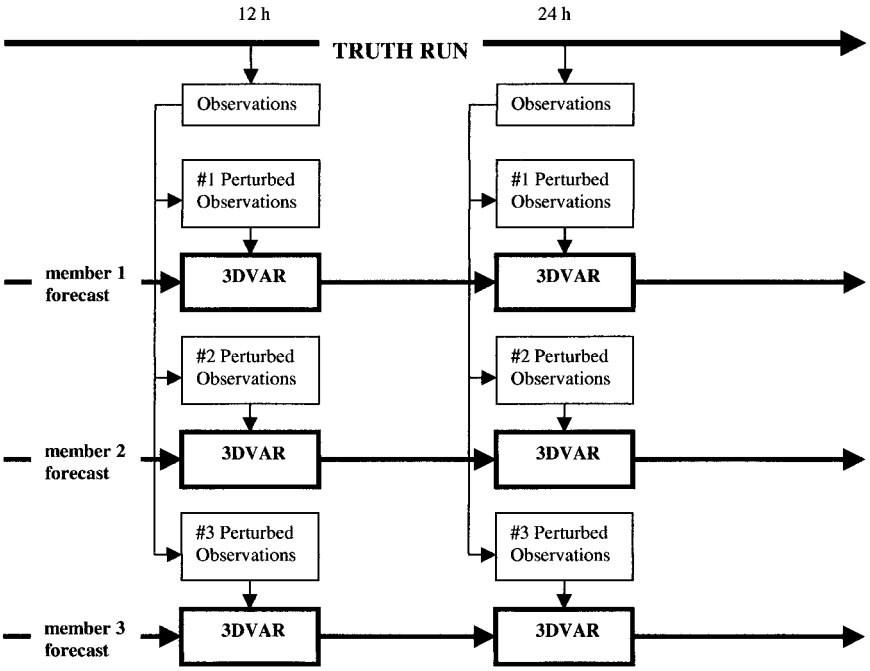

FIG. 3. Schematic illustrating the perturbed observation strategy as implemented here. Every 12 h, observations are generated from the truth run, and different sets of perturbed observations are assimilated into each ensemble member.

product of the square root of the eigenvalues and their respective eigenvectors of the observation error covariance matrix, as discussed in Houtekamer (1993). This cycle was then repeated; another 12-h forecast was generated from the control initial condition, and the data were assimilated using 3DVAR.

3. Ensemble techniques

a. Perturbed observations

The goal of the PO method is to generate a set of initial conditions that approximate a random sample from the analysis pdf by stochastically simulating each source of error in the analysis. To this end, the PO technique generates an ensemble of parallel forecast–data assimilation cycles with each cycle receiving unique perturbed observations and unique initial conditions. 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 perturbations were constructed from scaled differences between random model states following Schubert and Suarez (1989). The PO method then iterates the following three-step procedure: 1) Make N forecasts to the next analysis time; for our implementation, the first step is to $t_0 + 12$ h. (2) For each of the N parallel cycles, generate N independent sets of perturbed observations valid at this analysis time by adding noise to the control observations, with the noise added consistent with observational error covariances. 3) Produce an objective analysis, updating each of the N first guess fields using the associated set of perturbed observations. Here, the data assimilation scheme is 3DVAR. This procedure is schematically illustrated in Fig. 3. A sample initial condition and its differences from the PO ensemble mean initial condition are shown in Fig. 4a.

The breeding and SV schemes construct initial conditions by adding perturbations to a control analysis. To provide a consistent benchmark for comparison, we also constructed an alternative version of the PO analyses. In this version of PO, the perturbation differences of individual PO analyses from the mean of all PO analyses were recentered on the control analysis (this is also what is done operationally at the Canadian Meteorological Centre). We shall refer to subsequent ensemble forecasts as PO/recenter.

A more quantitative explanation for the rationale of perturbing observations in the ensemble Kalman filter is provided by Burgers et al. (1998), and the reader is also referred to HD95 for other background on the PO methodology.

b. The breeding method of generating ensemble perturbations

The breeding method implemented here followed the methodology outlined in Toth and Kalnay (1993, 1997; Z. Toth 1998, personal communication). Perturbed initial conditions were generated in sets of ''positive'' and ''negative'' pairs around a control initial condition.

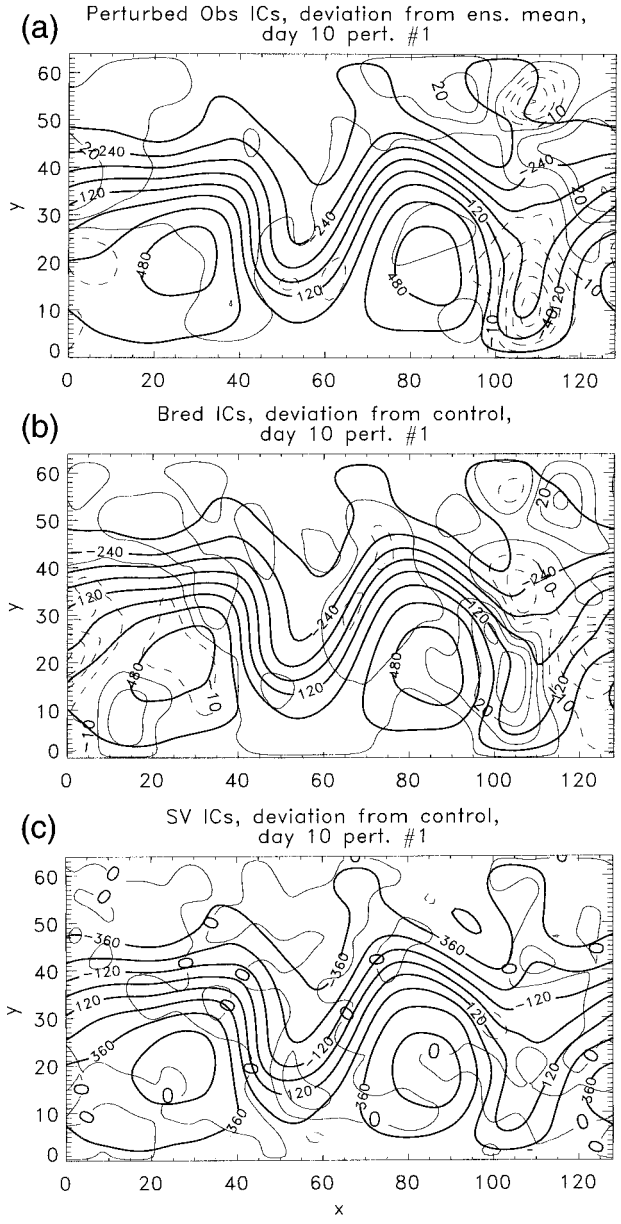

FIG. 4. (a) Sample PO perturbed initial condition (heavy solid) and difference from the ensemble mean initial condition (light solid and dashed) for geopotential height Z at level 4 and time $t = 10$ days into assimilation experiment. Contours for differences are every 10 m, with dashes indicating negative differences. (b) Bred geopotential height perturbed initial condition and deviation from the control analysis. (c) The SV perturbed initial condition and deviation from the control.

Starting with random perturbations, short-term forecasts were made (here, 12 h) for both members of a pair. The difference in model level 4 (~500 mb) geopotential height ($Z = \Phi/g$) between the two paired forecasts was smoothed with a Gaussian filter and the magnitude of the differences was compared to an estimate of the climatological analysis error. This estimate changed from region to region based on the local observational data density. If the difference exceeded twice the regional estimate of analysis error, the difference field was scaled back until the geopotential height difference was equal to two times the estimated analysis error; if not, the differences were unchanged. The perturbations were then centered around the control analysis, creating the positive and negative perturbation. The method was repeated for the remaining sets of pairs. Short-term forecasts were then generated for each ensemble member, and the breeding method was repeated at the next analysis cycle.

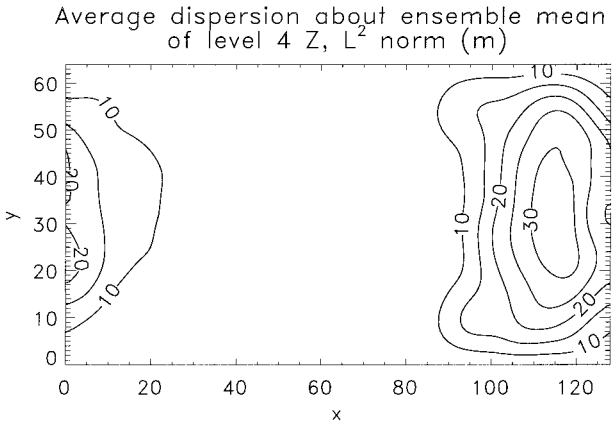

FIG. 5. Time-averaged standard deviation of the PO ensemble analysis about the ensemble mean geopotential height at model level 4.

As in the operational method, our implementation of the breeding technique requires a map of the spatial variation of typical analysis error (as in Fig. 6 of Toth and Kalnay 1997). We produced an estimate of this from the following: using the network from Fig. 2, the standard deviation of the PO ensemble about its mean measured in the L^2 norm was calculated at each model level 4 grid point. This was done for a set of 20 separate forecast case days four days apart from each other. Figure 5 illustrates this field of deviations, averaged over all case days. Differences between bred pairs were measured relative to this field.

A sample bred perturbed initial condition and its difference from the control analysis are shown in Fig. 4b.

c. *Approximate singular vectors*

The SVs are the directions in phase space where growth is maximized over a time interval $t_0 < t < t_1$. The first SV maximizes, over all possible perturbations to the control analysis at the initial time, amplification in a chosen norm between times t_0 and t_1. The second SV maximizes amplification in the subspace orthogonal to the first SV, the third maximizes amplification in the subspace orthogonal to that spanned by the first and second, and so on. Here, we choose a time interval $t_1 - t_0$ of 48 h and use the total energy norm.

We calculate perturbations that approximate the leading SVs by first constructing a larger set of perturbations

that are a random sample from a normal random vector with covariance proportional to $\mathbf{S}^{-1}$, where $\mathbf{S}$ is the matrix that defines the total energy inner product for the model; that is, $\mathbf{x}^T\mathbf{S}\mathbf{x}$ is the energy of a perturbation $\mathbf{x}$ [the relation between the statistics of initial perturbations and the norm used in calculating SVs is discussed further in Houtekamer (1995)]. Each scaled perturbation in this sample is then added to the control analysis at t_0 and integrated forward 48 h to t_1. The scaling is chosen small enough that the evolution of each perturbation is very nearly linear; in practice, the ratio of typical perturbation velocities to flow velocities is about 10^{-3}.

Next, we compute horizontal winds and temperatures from each perturbation at t_1 and calculate the eigenvectors of the resulting sample covariance matrix for the perturbations (i.e., we calculate the empirical orthogonal functions, or EOFs, in terms of winds and temperatures of the perturbations after 48 h). These eigenvectors approximate the evolved SVs at t_1, with errors that approach zero as the number of perturbations in the sample increases.

Of course, it is the SVs at initial time t_0 that have been suggested for use as ensemble perturbations. Each of the above eigenvectors at t_1 represents a linear combination of perturbations; these same linear combinations, but using the perturbations at t_0, approximate the initial SVs [related numerical techniques have been demonstrated in Lorenz (1965), Barkmeijer et al. (1998), and Bishop and Toth (1999)]. The SV perturbations were then generated as follows: the leading 12 singular vectors were selected to build 24 perturbations around the control (the 25th forecast was the control itself). Random rotations were generated, and the resulting rotated SV perturbations were then orthogonalized (under the energy norm) and normalized to a magnitude so their subsequent forecasts would have a domain-average energy equal to the PO domain- and time-average energy around day 2 of the forecast (T. Palmer 1998, personal communication). Positive and negative pairs of these SV perturbations were added to the control initial condition. An example of the resulting perturbations is shown in Fig. 4c.

We chose this approximate approach both because it was simple to implement and because we found it an intriguing application of ensemble techniques. Its obvious limitation is that reasonable approximation of the leading SVs may require the integration of an unfeasibly large set of perturbations. All results presented here use approximate SVs derived from samples of 200 perturbations at each t_0. An estimate of the quality of this approximation can be found in Fig. 6. This shows, for various sample sizes, the subspace similarity, that is, the projection of the leading 25 eigenvectors of the covariance matrix at $t_1 = 2$ days onto the subspace spanned by the leading 25 eigenvectors for a sample of size 400 (Buizza 1994). When 200 perturbations are used to construct the singular vectors, each of the first 10 eigenvectors have projections greater than 0.95 onto this subspace. Together these 10 vectors account for about 60% of the variance in the sample (not shown). At the final time t_1, there is thus little change in the subspace spanned by the leading 25 eigenvectors as the sample size increases beyond 200, and we conclude that the leading approximate SVs are nearly converged to the exact SVs.

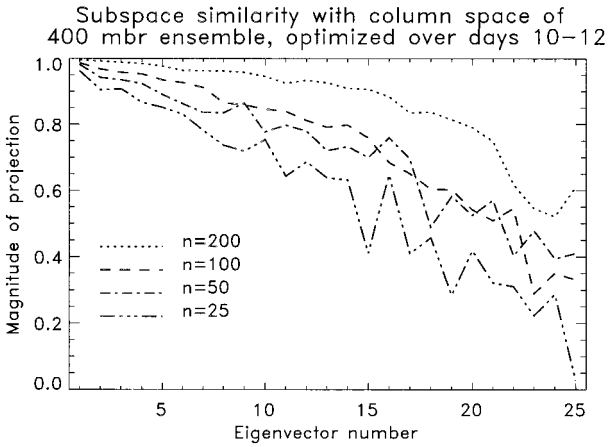

FIG. 6. Subspace similarity index indicating the amount of projection of eigenvectors of ensembles of size 25, 50, 100, and 200 project onto the subspace of the leading 25 singular vectors from a 400-member ensemble. Projection amounts are shown for a 2-day forecast starting at day 10.

Sampling problems are worse at t_0, as might be expected given that we seek to approximate perturbations that grow rapidly between t_0 and t_1. Indeed, the approximations to the SVs at t_0 are noisy (see Fig. 4c) compared to calculations of SVs using the tangent linear and adjoint. As we will show below, however, the approximate SVs still grow very rapidly (relative to the leading Lyapunov exponent, say) between t_0 and t_1, and must therefore have a strong projection on the leading subspace of exact SVs. This rapid growth, combined with the fact that the leading approximate SVs are, by 48 h, quite similar to the exact SVs, indicates that the approximate SVs should have comparable performance for ensemble forecasting, at least after 48 h.

4. Comparison of ensemble initial conditions and forecasts

We now evaluate the quality of ensemble analyses and forecasts from each perturbation methodology. We start with an examination of the characteristics of analysis error for the PO ensembles and the control. Next, we determine whether an ensemble of initial conditions for each of the three perturbation techniques has uniform rank histograms (Anderson 1996; Hamill and Colucci 1997, 1998a), which are also known as ''Talagrand diagrams.'' Uniformity of the rank histogram is a necessary (but not sufficient) condition for the analysis to be considered a random sample from the analysis pdf.

Thereafter, we probabilistically evaluate the ensembles of forecasts. We again use rank histograms but also explore the error growth characteristics of each forecast in the various norms and the accuracy of subsequent probabilistic forecasts. We will also present spread–skill relationships for the three methods.

Twenty separate forecast case days were used, starting 10 days into the assimilation and then producing forecasts every 4 days thereafter. Forecasts were evaluated to a lead time of 5 days.

a. Control analysis and PO ensemble mean analysis characteristics

We first document the characteristics of the control analyses relative to the PO ensemble. This control analysis should be more accurate on average than the individual PO initial conditions, since PO initial conditions receive observations with additional random errors added. To verify this, a 90-day PO assimilation cycle was carried out with 25 members using the observational network in Fig. 2. Similarly, the control initial condition was generated for the same 90-day cycle. Figures 7a–c show the analysis error measured for the PO ensemble (dots), the control (solid line), and the PO ensemble mean (dot–dash). As shown, the control initial condition usually has lower error than the majority of the PO initial conditions. Interestingly, though the control analysis is on average more accurate than individual ensemble members' analyses, the control analysis is also typically *higher* in error than the ensemble mean analysis. The improvement of the ensemble mean analysis over the control is most obvious in the enstrophy norm, which emphasizes the smaller, less predictable scales.

The difference between the ensemble mean and the control analyses may have several causes. First, it is possible that if tested over a longer test period, the differences would be diminished. Another possibility is that this improvement is due to the small but cumulative effects on nonlinearities that develop during each 12-h forecast between assimilation cycles. These cause the ensemble mean of first guess fields to have (on average) less error than the control first guess [a similar result was also suggested in Kalnay and Toth (1994)]. The differences may be due to forecast nonlinearities because the analysis operator is linear and cannot contribute to this effect. Given a positively and negatively perturbed pair of first guess fields centered on a control first guess, and given a positively and negatively perturbed pair of observations centered on a control set of observations, the average of the pair of analyses will be the same as if the control first guess were updated with control observations.

This result indicates that the PO ensemble started with an advantage over the other two perturbation methods, since the swarm of PO ensembles in this simulation was more optimally centered in phase space than the bred or SV techniques, which were centered around this higher-error control analysis. Note that the improvement of the PO ensemble mean analysis over the control analysis was a result that was not duplicated operationally at the Canadian Meteorological Centre (P. Houtekamer 1998, personal communication). There, an improved assimilation scheme at higher resolution was used to generate the control analysis, so it was typically lower in error.

b. Ensemble initial condition characteristics

We first examine rank histograms of the analysis error. These histograms were generated by determining the rank of the truth at a given point when pooled with an ensemble sorted from lowest to highest. If the ensemble is a random sample from the same distribution as the truth, then the truth will be equally probable to occur in each rank and, over many points and days, the rank histogram should be populated uniformly across ranks.

Figures 8a–l show rank histograms of PO, PO/recenter, SV, and bred level 4 Z, u, and θ. As shown, the PO and PO/recenter initial conditions were much closer to exhibiting the desired uniformity of rank. Thus, the bred and SV initial conditions do not meet the necessary test of uniformity to be considered random samples from the analysis pdf.

Bred and SV methods each have different problems that contribute to their initial lack of uniformity of rank. One common reason their extreme ranks of Z were unduly populated is that the swarm of bred, SV, and PO/recenter ensembles were less optimally centered in phase space, as was discussed in section 4a. A more important reason for nonuniformity of the bred and SV rank histograms is that the initial size of bred and SV were not constructed in a manner that permits them to estimate the actual uncertainty of the flow that day. Current operational bred and SV ensemble initial conditions were specifically designed to have a *fixed* initial spread (and in the case of SVs, a very small initial spread); hence they cannot possibly be samples from the true distribution, which varies in time, as shown in Fig. 7.[2] Conversely, PO perturbations appear to vary in size with analysis error. Evidence for this variation is provided in Fig. 9, a plot of the standard deviation of PO level 4 Z first guess fields at days 10, 20, and 30. In all cases there were larger deviations over the data void than over the data-rich area, but the domain-averaged dispersion also varied from day to day, and this amount of dispersion was roughly consistent with the analysis errors for these days (Fig. 7). The PO initial conditions showed significant spread–skill correlation, to be discussed later.

Another characteristic that appeared in the rank histograms of Fig. 8 were the lower populations at the extreme ranks for u and θ than for Z in the breeding

[2] It may be possible to design alternative breeding or SV methods where initial spread varies in time depending on recent error growth. Alternatives to the operational methods were not tested here.

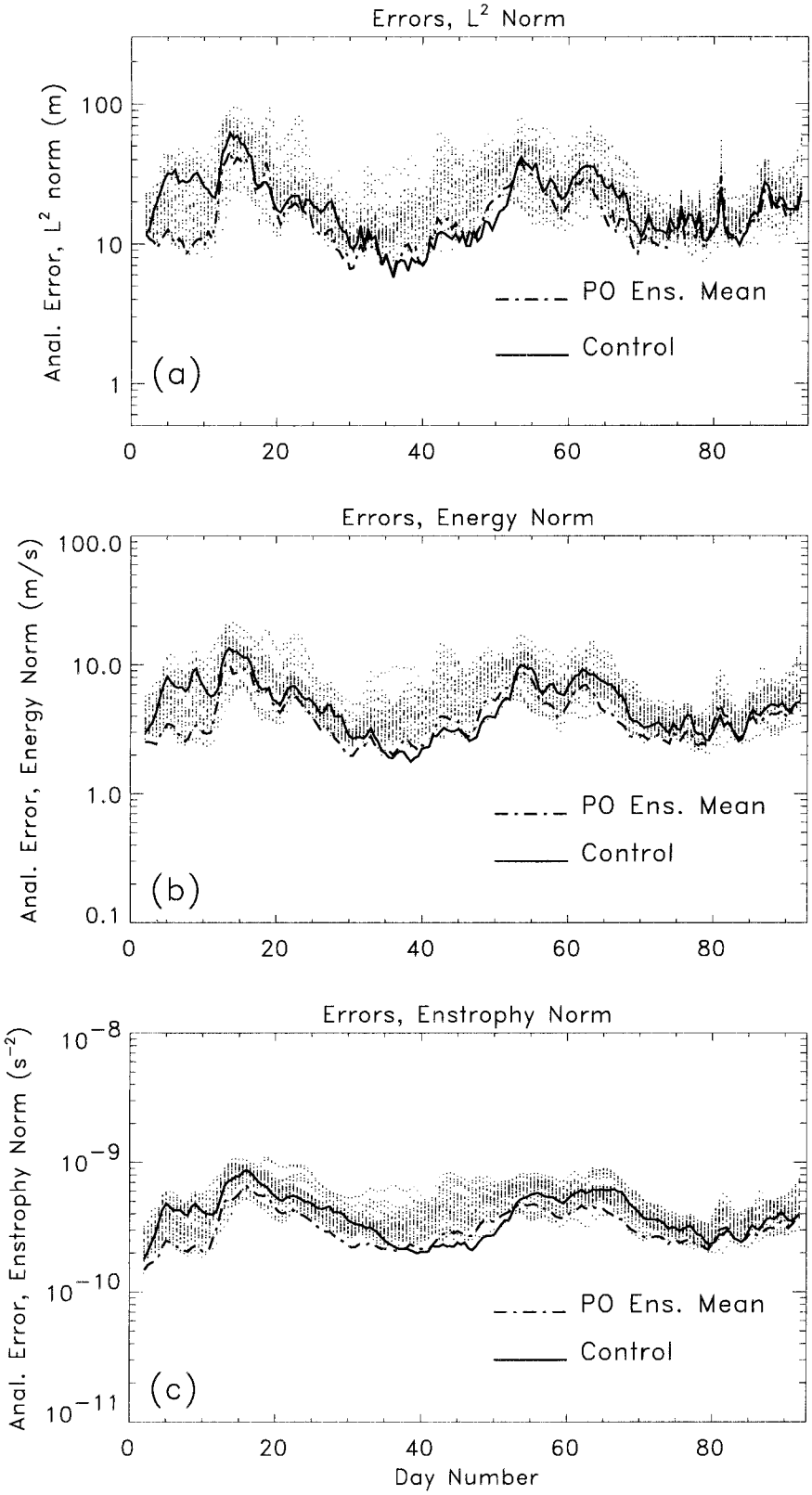

FIG. 7. Analysis error for PO ensemble (dots), the control for breeding and SV forecasts (solid), and the PO ensemble mean (dot–dash) measured in the (a) L^2 norm, (b) energy norm, and (c) enstrophy norm.

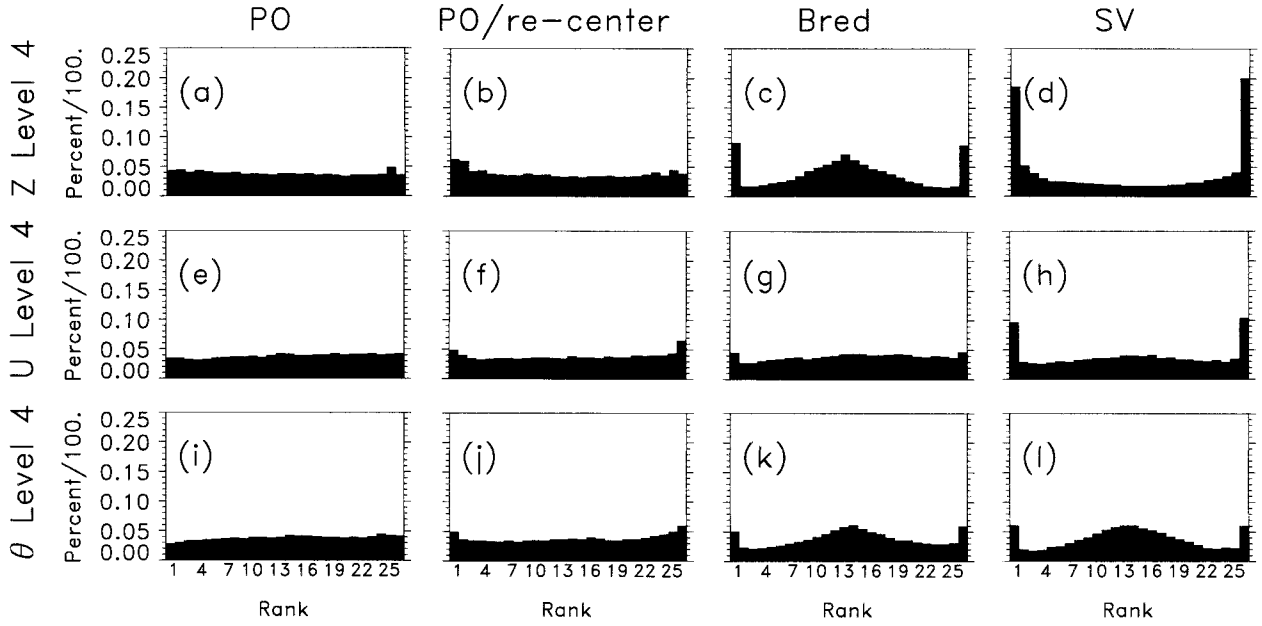

FIG. 8. Rank histograms of ensemble analyses: (a) PO level 4 *Z*, (b) PO/recenter level 4 *Z*, (c) bred level 4 *Z*, (d) SV level 4 *Z*, (e) PO level 4 *u*, (f) PO/recenter level 4 *u*, (g) bred level 4 *u*, (h) SV level 4 *u*, (i) PO level 4 θ, (j) PO/recenter level 4 θ, (k) bred level 4 θ, and (l) SV level 4 θ.

and SV techniques. We consider the breeding technique's reasons first. Here, the flatter rank histogram in *u* and θ was unfortunately not so much a sign of proper sampling of the pdf but rather a consequence of the *u* and θ fields having been noisier than the *Z* fields. The noise introduced to winds, potential temperatures, and PV perturbations was a result of the use of a regional rescaling process; transitions from regions where rescaling was performed to those where no rescaling was performed introduced kinks into the *Z* field. Taking spatial derivatives accentuated this noise, resulting in a larger spread of the ensemble. Hence, there was less probability the truth was beyond the span of the ensemble for these variables. Unfortunately, the larger spread was introduced arbitrarily at the transition zones from rescaling to no rescaling and was not necessarily associated with regions of enhanced uncertainty.

For the SV technique, we believe there were two primary causes for differently shaped *u*, θ, and *Z* rank histograms. First, again *u* and θ were obtained through spatial derivatives of *Z*, and the small-scale noise in the initial *Z* perturbations was accentuated by taking its derivative. Hence, *u* and θ perturbations were generally larger and noisier. The shapes of the rank histograms were also a consequence of initializing SVs with equally sized initial perturbations. A histogram of the SV perturbations (not shown) showed them to be more uniformly distributed around the control, but with smaller tails than the histogram of PO perturbations, which was approximately normally distributed. Assuming this normally shaped distribution was correct, this contributed to depleting the population near the extreme ranks.

c. Ensemble forecast characteristics

Figure 10 shows the rank histograms for the PO, PO/recenter, bred, and SV ensembles for level 4 *Z* at 1-, 3-, and 5-day forecast lead times. As shown, the rank histograms for the PO and PO/recenter ensembles started off relatively uniform (Fig. 8) and remained qualitatively near uniform throughout the forecast. This was one indication that these ensembles may have provided useful probability forecasts without requiring calibration for systematic deficiencies (Hamill and Colucci 1997, 1998a), at least in this perfect-model context. Conversely, the *Z* rank histograms for the bred and SV ensembles started off unduly populated at the extremes, showing that those ensembles still did not appropriately span the range of forecast possibilities. By day 5, all perturbation methods produced ensemble forecasts with relatively uniform rank histograms.

These forecast traits are better understood by considering the forecast dispersion (error growth) in various norms (Figs. 11a–c). By design, the PO, PO/recenter (not shown) and bred perturbations had approximately equal energy norms at the beginning of the forecast, averaged over many cases. Their growth rates in L^2 and energy were roughly somewhat comparable thereafter, suggesting that they are using qualitatively different perturbations than for the SV, which grow very rapidly in these norms during the first 2 days. After day 2, the growth rates were a bit more similar among all perturbation methods, suggesting all forecasts were now spanning the same dynamically important subspace. The SV and bred L^2 growth rates were larger than the PO growth rate.

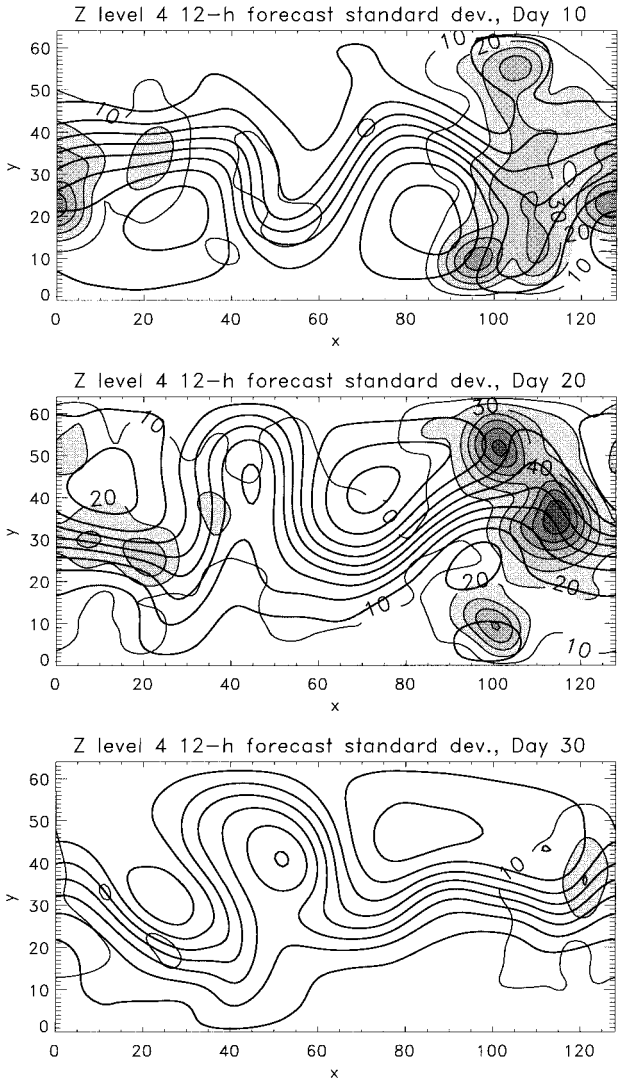

FIG. 9. Standard deviation of PO level 4 first guess Z about ensemble mean at (a) day 10, (b) day 20, and (c) day 30. Compare with Fig. 7.

Understanding the characteristics of dispersion in the enstrophy requires consideration of the algorithmic details. For the bred and especially for the SV ensembles, the perturbations were larger than those in the PO ensemble when measured in the enstrophy norm, indicating that the bred and SV perturbations had larger amplitudes at the smaller scales. For the bred ensemble, this was a result of the regional rescaling process; as mentioned earlier, transitions from regions of rescaling to those with no rescaling introduced kinks into the Z field, which produced discontinuities in PV. For the SV ensemble, the short-range forecasts inherited some unrealistic noise at small scales from use of the approximate initial SVs, as discussed in section 3c. The performance demonstrated here for the SV ensemble in the 0–2-day range is not necessarily indicative of what would be obtained with exact SVs constructed using the tangent-linear and adjoint models.

Next we consider the spread–skill relation. A properly calibrated ensemble should show some correlation of spread (deviation of the ensemble about its mean) and skill of the ensemble mean (Whitaker and Loughe 1998 and references therein). To test for possible spread–skill relationships here, we calculated the spatially averaged spread at level 4 on each case day and compared this to a spatially averaged rms error of the ensemble mean. We performed the spatial averaging separately for the data-rich eastern two-thirds of the domain (land, or ''l'') and the western third (ocean, or ''o''). Figure 12 shows the spread–skill relationship for each perturbation method at 0-, 1-, 3-, and 5-day lead times. As shown, the PO and PO/recenter ensembles started with a strong spread–skill relationship and maintain this during the forecast, while the bred and SV ensembles initially had no relevant spread–skill relationship but developed them by day 5. This demonstrates that the PO method was capturing the time-dependent initial condition uncertainty that the breeding and SV methods, by construction, did not. Note, however, that because of this chosen network design, the analysis errors here varied more widely with time than in operational analyses, permitting the stronger spread–skill relationships than would likely be observed operationally (Whitaker and Loughe 1998).

Another method for evaluating competing probabilistic forecasts is through the relative operating characteristic, or ROC (Swets 1973; Mason 1982; Stanski et al. 1989). This diagram evaluates type I (incorrect acceptance of the alternative hypothesis) and type II (incorrect acceptance of null hypothesis) statistical errors evaluated at various percentiles of a forecast probability distribution. In this diagram, the hit rate [$=1 - P$(type II error)] is plotted relative to the false alarm rate [$=P$(type I error)] at incremental percentiles of the forecast probability distribution. Details of the construction of the ROC are discussed in the appendix. If one forecast methodology has a ROC curve farther up and to the left on the diagram, it exhibits less of each type of error and may be considered the better forecast. Figure 13 shows ROC curves for P(level 4 wind speed $>$ 60 m s^{-1}). As shown, the PO ensemble has the highest ROC curve at all thresholds. PO/recenter, SV, and bred curves are all relatively similar.

We also compared the skill of probabilistic forecasts using the Brier score (Brier 1950; Wilks 1995). Table 3 provides Brier scores for PO, PO/recenter, SV, and bred forecasts for P(level 4 wind speed $>$ 60 m s^{-1}). Here, probabilities were calculated from the relative frequency of member forecasts exhibiting winds greater than 60 m s^{-1}. Using daily average Brier scores as samples, a paired t-test was conducted to determine whether or not the improvement of the PO ensemble over the SV and bred ensembles was statistically significant (Hamill 1999). These hypothesis tests indicate that the

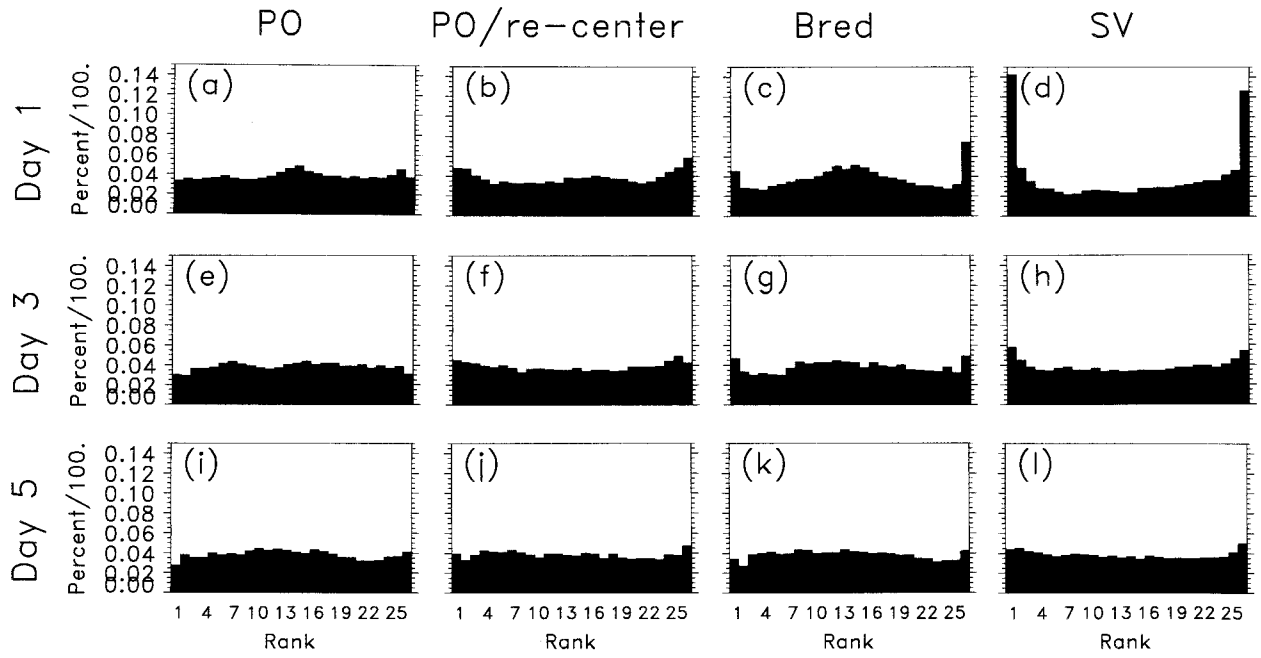

FIG. 10. Rank histograms of ensemble forecasts of the level 4*Z*: (a) PO, day 1 forecast; (b) PO/recenter, day 1 forecast; (c) bred, day 1 forecast; (d) SV, day 1 forecast; (e) PO, day 3 forecast; (f) PO/recenter, day 3 forecast; (g) bred, day 3 forecast; (h) SV, day 3 forecast; (i) PO, day 5 forecast; (i) PO/recenter, day 5 forecast; (k) bred, day 5 forecast; and (l) SV, day 5 forecast.

improvement of the PO over bred and SV was statistically significant at all three lead times ($p < 0.001$ for all tests). The improvement of PO/recenter over the bred and SV forecasts was not statistically significant.

5. Discussion

For this simulation, the PO method (without recentering) provided a reduced-error set of ensemble initial conditions and forecasts relative to the breeding and SV techniques. The PO initial conditions and forecasts were better calibrated, showed a stronger spread–skill relationship, and produced probabilistic forecasts with reduced errors. Only the PO method captured the time-varying uncertainty in the initial condition. The initial spread of the PO ensemble varied in time, apparently to an extent consistent with the analysis uncertainty. Some of the beneficial characteristics, such as spread–skill correlations and relatively flat rank histograms were retained after recentering PO perturbations on the control initial condition; other characteristics, such as ROC curves and Brier scores were degraded to a level no more skillful than bred or SV forecasts.

These illustrate the theoretical benefits of the PO method in the under a specific set of conditions. The relative performance of the perturbation methods in this perfect-model context and under our network design does not directly indicate the expected performance in an operational numerical weather prediction setting. The design of this experiment made the unrealistic assumptions of a perfect model, assumed full knowledge of the observational error characteristics, and used a network with an accentuated data void. In operational numerical weather prediction, model error growth can be insidious and as large or larger than error growth due to initial condition uncertainty. Also, for the observations, much information on the observational error characteristics is available (and is required for data assimilation) but that information is of course imperfect. In particular, the possibly non-Gaussian representativeness errors are not accounted for in this experiment; these are the errors that arise from representing the atmosphere with the finite basis of a numerical weather prediction model.

We also note that the PO method is not the only plausible method for designing sets of initial conditions that sample the analysis pdf. Though we tested the singular vector method here using an initial energy norm, the singular vector method need not use this norm. In principle, the initial norm should be based upon the analysis error covariances (Houtekamer 1995; Ehrendorfer and Tribbia 1997). When using an exact analysis

TABLE 3. Brier scores for P (level 4 wind > 60 m s^{-1}) tallied over all case days and for each individual case days, for 1-, 3-, and 5-day forecasts.

	Brier score day 1	Brier score day 3	Brier score day 5
PO	0.0215	0.0502	0.0920
PO/recenter	0.0280	0.0603	0.0102
Bred	0.0290	0.0615	0.0104
SV	0.0274	0.0636	0.0104

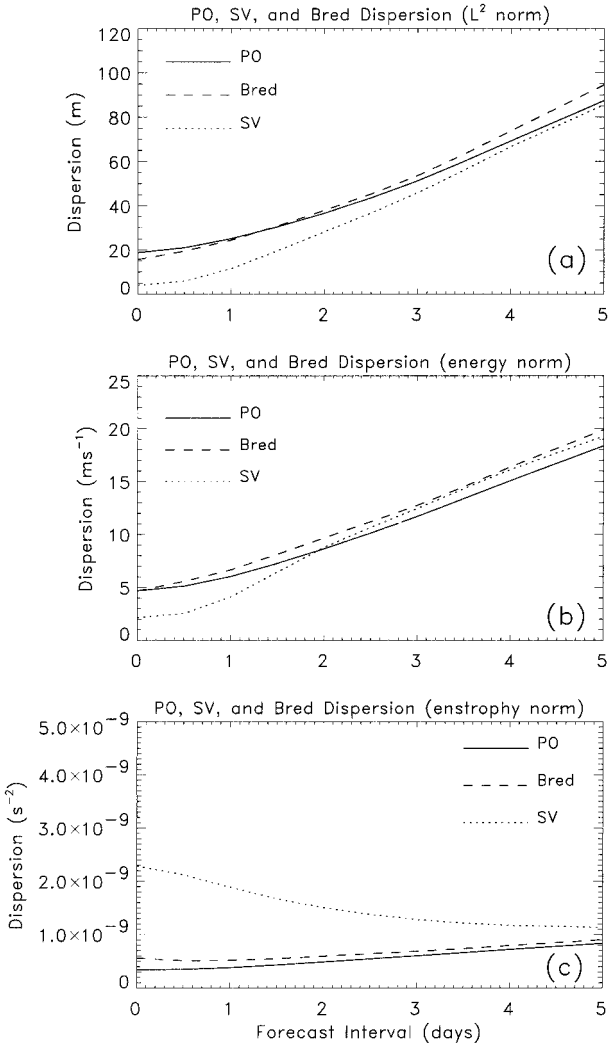

FIG. 11. Dispersion of the PO, SV, and bred ensemble as a function of forecast lead time in (a) L^2 norm, (b) energy norm, and (c) enstrophy norm.

error covariance norm, forecasts from the resulting singular vectors are ''optimal'' in the sense that they account for the maximum percentage of error variance possible for a given sized ensemble (under assumptions of linear error growth). These singular vectors may also account for day-to-day variations in analysis uncertainty, and hence we would expect their performance to be similar to that demonstrated here by the PO technique. The relative computational cost of the PO and ''analysis error covariance singular vectors'' is still unquantified. Also, growth of errors for analysis error covariance singular vectors may not be linear, as assumed. Still, based on their obvious theoretical appeal, ECMWF is vigorously exploring the use of analysis error covariance singular vectors for ensemble perturbations and for improving data assimilation (Barkmeijer et al. 1998, 1999; M. Ehrendorfer 1998, personal communication).

Though our assumption of no model error raises questions about operational validity, as previously discussed, such an experimental design does present advantages. Specifically, initial condition uncertainty may be considered in isolation from other effects, and our results suggest something different from the current conventional wisdom. For example, the rank histograms for real-world ensemble forecasts (e.g., Hamill and Colucci 1997) are more highly populated at the extreme ranks than for these perfect-model forecasts. Since the true solution frequently lies outside the swarm of ensemble forecasts, this sort of evidence is occasionally cited as a reason why ensemble forecasts should use a perturbation methodology that produces more dispersive forecasts. Perturbations that grow rapidly should produce forecasts that are more likely to encompass the truth. However, in this experiment, PO perturbations grew much more slowly than the singular vectors, yet insufficient error growth was not a problem; the PO method produced calibrated, reduced-error probabilistic forecasts.

Based on this, we suggest that the design of initial perturbations should address initial condition uncertainty alone and not also attempt to compensate for problems in the ensemble forecast associated with model errors. *The perturbation methodology should be designed to produce realistic random samples of the analysis pdf,* and (at least in the perfect-model context) the appropriate amount of forecast dispersion should naturally result.

What if forecasts conducted with the PO strategy do not encompass the truth often enough when implemented operationally? This then suggests that either model error is nonnegligible and/or there are aspects of the initial condition that are not properly perturbed. These issues are just beginning to be explored. If model error cannot be rendered unimportant, then perhaps stochastic forcing will need to be added to the model equations (e.g., Buizza et al. 1999), or different plausible model configurations used among the ensemble members (Stensrud and Fritsch 1994; Houtekamer et al. 1996), and/or model error accounted for during the data assimilation (Mitchell and Houtekamer 2000). Another possibility is that our definition of the ''initial condition'' should be expanded to include the state of the land surface and of poorly defined fixed parameters such as roughness length (Houtekamer et al. 1996; Hamill 1997; Hamill and Colucci 1998b). Numerical forecasts are often sensitive to the state of these variables, yet they are not currently perturbed in the forecasts from NCEP or ECMWF.

The evidence provided here also reinforces the growing supposition that judiciously designed Monte Carlo ensembles may be useful in defining background error statistics to be used during the data assimilation. Current operational assimilation schemes such as the 3DVAR scheme used here assume background errors are isotropic and stationary, whereas in truth they may vary sub-

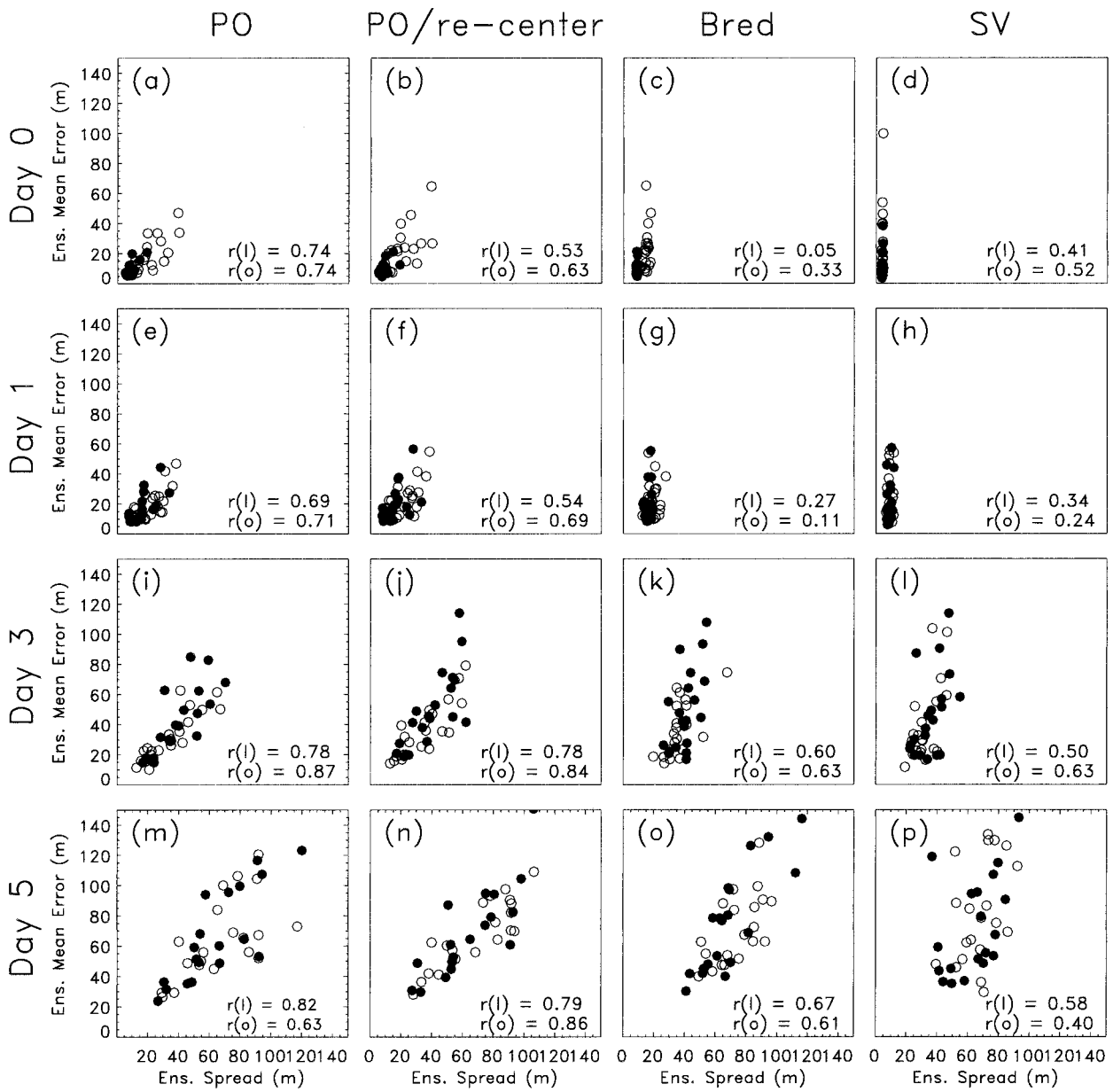

FIG. 12. Spread–skill relationships, comparing level 4 spatially averaged standard deviations of *Z* against spatially averaged rms errors of the ensemble mean *Z*. Darkened circles represent one case day's sample spread–skill over the data-rich western two-thirds of the domain (land). Unfilled circles represent spread–skill over the data void (ocean). Correlations of spread and skill also plotted for land (l) and ocean (o): (a) day 0, PO; (b) day O, PO/recenter; (c) day 0, bred; (d) day 0, SV; (e) day 1, PO; (f) day 1, PO/recenter; (g) day 1, bred; (h) day 1, SV; (i) day 3, PO; (j) day 3, PO/recenter; (k) day 3, bred; (l) day 3, SV; (m) day 5, PO; (n) day 5, PO/recenter; (o) day 5, bred; and (p) day 5, SV.

stantially in space and time. Figure 5 showed that when the observational data density is nonuniform, then local analysis errors vary with location. As was shown in Figs. 9 and 12, the standard deviations of the PO ensemble first guess fields on three different days exhibit temporal as well as spatial variation, apparently in a realistic manner. There is growing interest in Kalman filtering (e.g., Daley 1991), a procedure that generates not only numerical forecasts but estimates of background error covariances. The filter then uses these error statistics to improve data assimilation. Also, recent research by Evensen and van Leeuwen (1996), Houtekamer and Mitchell (1998), and Mitchell and Houtekamer (2000) have shown the potential of an approach called the ''ensemble Kalman filter.'' This technique uses sets of ensembles generated in a manner similar to the PO method to estimate the background error covariance statistics. Others have proposed alternative approaches to

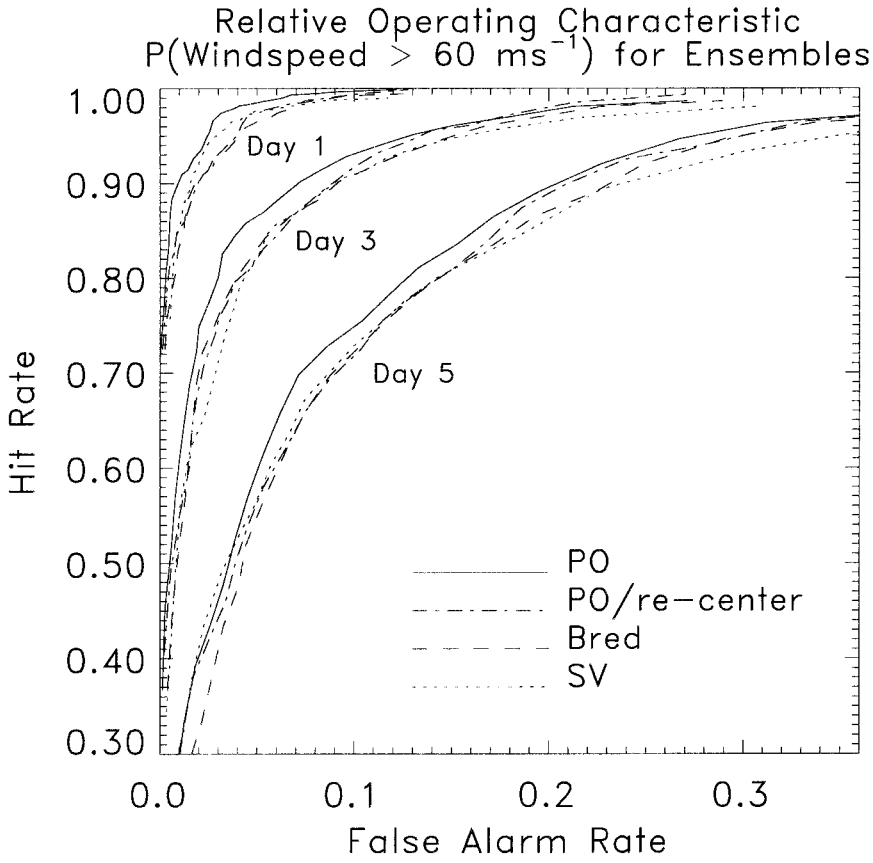

FIG. 13. Relative operating characteristic for P(level 4 wind speed $>$ 60 m s^{-1}) for PO, PO/recenter, bred, and SV ensembles at 1-, 3-, and 5-day lead times.

judiciously using ensembles to improve data assimilation (e.g., Anderson and Anderson 1999).

Implementation of the PO method operationally requires more computer resources than the breeding method since the 3DVAR data assimilation procedure must be repeated for each ensemble member. This may be no more expensive, however, than the computation of singular vectors used to design perturbations for the ECMWF ensemble. Also, the PO method is fully parallel and in principle, on the right computer, should require no more time than a single control analysis and forecast.

6. Summary

Characteristics of ensemble analysis and forecast errors were explored using a quasigeostrophic channel model in a perfect-model framework. Though the assumption of a perfect model is clearly an inappropriate analog to operational numerical weather prediction models, this simplification isolates the effects of predictability error growth and permits the exploration of ensemble characteristics without influence of model error.

The perturbed observation method in this simulation was shown to generate an improved set of initial conditions for ensemble forecasts. The PO perturbation methodology started from a reduced-error set of initial states and thus resulted in reduced-error probabilistic forecasts relative to the bred and approximate singular-vector methodologies.

There were a number of reasons why PO forecasts appeared to be improved. First, the PO forecast initial conditions were more accurate as a group. This could be seen by comparing the errors of the ensemble mean initial condition for the PO ensemble to the control initial condition. The PO ensemble mean was typically lower in error, indicating the PO ensemble was more appropriately centered in phase space. Much of the advantage of the PO ensemble was lost when PO initial conditions were recentered on the same control initial condition as the bred and SV initial conditions. The PO ensemble initial conditions also exhibited uniformity of verification rank and a positive spread–skill relationship, suggesting that the ensemble was capturing the flow- and location-dependent nature of the uncertainty in the analysis.

The PO technique also produced relatively uniform rank histograms, improved probabilistic forecasts, and substantially higher spread–skill correlations throughout

the subsequent forecast. This occurred even though the growth of perturbations was slower than that produced by the SV ensembles.

We suggest testing of the PO methodology in a more operationally relevant situation when computer resources permit.

Finally, the PO ensemble produces forecast information that may be useful for defining time- and flow-dependent background error statistics used during the data assimilation. Use of these statistics may reduce errors in the initial condition and in the subsequent numerical weather forecasts. Such ensemble-based strategies should be explored more fully in the near future.

Acknowledgments. The authors would like to thank Tim Palmer (EMCWF), Zoltan Toth (NCEP), Peter Houtekamer (CMC), and Jeff Anderson (GFDL) for their consultation and advice during the preparation of this manuscript. Eugenia Kalnay and two anonymous reviewers are thanked for their substantive formal reviews. This research was funded by NCAR's U.S. Weather Research Program.

APPENDIX

Generation of ROC from Ensembles

The ROC is a plot of ''hit rate'' versus ''false alarm rate'' calculated using forecasts at various quantiles of a probability distribution as decision thresholds. General use and interpretation of the ROC is explained in more depth in Swets (1973) and Stanski et al. (1989). We focus here on how to use ensemble data to calculate the ROC. The ROC is calculated for a specific forecast parameter, such as the probability that the wind exceeds a certain threshold T. Assume for a given sample location the ensemble forecast $\mathbf{x}$ of this parameter is a vector of N forecasts sorted from lowest to highest. Also assume an observation O is available at this location. Each sorted ensemble member from lowest to highest may be considered a potential decision threshold. A particular 2×2 contingency table $\mathbf{C}_i$, $i = 1, \ldots, N$ is associated with each of the N sorted members; hence, there will be N contingency tables calculated (Table A1). For a given sorted ensemble member x_i and an observation y, an element of a 2×2 contingency table $\mathbf{C}_i$ is populated, the element of the table depending on whether the observation and member forecast were each above or below the threshold T. For example, element b of $\mathbf{C}_1$ is incremented by 1 if the $x_1 > T$ and the $O < T$. The rest of the N contingency tables are incremented using the rest of the N ensemble members as forecasts. This process is repeated using (presumably independent) samples from different observation locations and different case days. The hit rate $[=a/(a + c)]$ and false alarm rate $[=b/(b + d)]$ are then calculated for each of the N contingency tables and plotted to generate the ROC.

TABLE A1. Contingency table of possible events in generating ROC.

		Observed $> T$?	
		Yes	No
Member Forecast $> T$?	Yes	a	b
	No	c	d

REFERENCES

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

——, 1997: Impact of dynamical constraints on the selection of initial conditions for ensemble predictions: Low-order perfect model results. *Mon. Wea. Rev.,* **125,** 2969–2983.

——, 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.

Barkmeijer, J., M., van Gijzen, and F. Bouttier, 1998: Singular vectors and estimates of the analysis error covariance metric. *Quart. J. Roy. Meteor. Soc.,* **124,** 1695–1713.

——, R. Buizza, and T. N. Palmer, 1999: 3D-Var Hessian singular vectors and their use in the ECMWF ensemble prediction system. *Quart. J. Roy. Meteor. Soc.,* **125,** 2333–2351.

Bergman, K. H., 1979: Multivariate analysis of temperatures and winds using optimum interpolation. *Mon. Wea. Rev.,* **107,** 1423–1444.

Bishop, C. H., and Z. Toth, 1999: Ensemble transformation and adaptive observations. *J. Atmos. Sci.,* **56,** 1748–1765.

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

Buizza, R., 1994: Sensitivity of optimal unstable structures. *Quart. J. Roy. Meteor. Soc.,* **120,** 429–451.

——, and T. N. Palmer, 1995: The singular vector structure of the atmospheric general circulation. *J. Atmos. Sci.,* **52,** 1434–1456.

——, M. Miller, and T. N. Palmer, 1999: Stochastic representation of model uncertainties 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.

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

Ehrendorfer, M., 1994a: The Liouville equation and its potential usefulness for the prediction of forecast skill. Part I: Theory. *Mon. Wea. Rev.,* **122,** 703–713.

——, 1994b: The Liouville equation and its potential usefulness for the prediction of forecast skill. Part II: Applications. *Mon. Wea. Rev.,* **122,** 714–728.

——, and J. J. Tribbia, 1997: Optimal prediction of forecast error covariances through singular vectors. *J. Atmos. Sci.,* **54,** 286–311.

Errico, R. M., 1997: What is an adjoint model? *Bull. Amer. Meteor. Soc.,* **78,** 2577–2592.

Evensen, G., 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.

Fritsch, J. M., and Coauthors, 1998: Quantitative precipitation forecasts: Report of the eighth prospectus development team, U.S. Weather Research Program. *Bull. Amer. Meteor. Soc.,* **79,** 285–299.

Hamill, T. M., 1997: Short-range ensemble forecasting using the Eta/RSM models. Ph.D. dissertation, Cornell University, Ithaca, NY, 216 pp. [Available from UMI Dissertation Services, 300 N. Zeeb Rd., P.O. Box 1346, Ann Arbor, MI 48106-1346.]

——, 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: Evaluation of Eta-RSM ensemble probabilistic precipitation forecasts. *Mon. Wea. Rev.,* **126,** 711–724.
——, and ——, 1998b: 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.
Houtekamer, P. L., 1993: Global and local skill forecasts. *Mon. Wea. Rev.,* **121,** 1834–1846.
——, 1995: The construction of optimal perturbations. *Mon. Wea. Rev.,* **123,** 2888–2898.
——, 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.
——, and Coauthors, 1996: A system simulation approach to ensemble prediction. *Mon. Wea. Rev.,* **124,** 1225–1242.
Kalnay, E., and Z. Toth, 1994: Removing growing errors from the analysis cycle. Preprints, *10th Conf. on Numerical Weather Prediction,* Portland, OR, Amer. Meteor. Soc., 212–215.
——, and Coauthors, 1996: The NCEP/NCAR 40-Year Reanalysis Project. *Bull. Amer. Meteor. Soc.,* **77,** 437–471.
Legras, B., and R. Vautard, 1996: A guide to Liapunov vectors. Proc. *ECMWF Seminar on Predictablility,* Vol. I, Reading, United Kingdom, ECMWF, 143–156.
Leith, C. E., 1974: Theoretical skill of Monte Carlo forecasts. *Mon. Wea. Rev.,* **102,** 409–418.
Lorenz, E. N., 1963: Deterministic nonperiodic flow. *J. Atmos. Sci.,* **20,** 131–140.
——, 1965: A study of the predictability of a 28-variable atmospheric model. *Tellus,* **17,** 321–333.
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. Royal Meteor. Soc.,* **122,** 73–119.
Morss, R. E., 1999: Adaptive observations: Idealized sampling strategies for improving numerical weather prediction. Ph.D. thesis, Massachusetts Institute of Technology, Cambridge, MA, 225 pp. [Available from UMI Dissertation Services, 300 N. Zeeb Rd., P.O. Box 1346, Ann Arbor, MI 48106-1346.]
Mullen, S. L., and D. P. Baumhefner, 1989: The impact of initial condition uncertainty on numerical simulations of large-scale explosive cyclogenesis. *Mon. Wea. Rev.,* **117,** 2800–2821.
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.
Rotunno, R., and J.-W. Bao, 1996: A case study of cyclogenesis using a model hierarchy. *Mon. Wea. Rev.,* **124,** 1051–1066.
Schubert, S. D., and M. Suarez, 1989: Dynamical predictability in a simple general circulation model: Average error growth. *J. Atmos. Sci.,* **46,** 353–370.
Stanski, H. R., L. J. Wilson, and W. R. Burrows, 1989: Survey of common verification methods in meteorology. Environment Canada Research Rep. 89-5, 114 pp. [Available from Atmospheric Environment Service, Forecast Research Division, 4905 Dufferin St., Downsview, ON M3H 5T4, Canada.]
Stensrud, D. J., and J. M. Fritsch, 1994: Mesoscale convective systems in weakly forced large-scale environments. Part III: Numerical simulations and implications for weather forecasting. *Mon. Wea. Rev.,* **122,** 2084–2104.
Swets, J. A., 1973: The relative operating characteristic in psychology. *Science,* **182,** 990–999.
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.
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.
Wolf, A., J. B. Swift, H. L. Swinney, and J. A. Vastano, 1985: Determining Lyapunov exponents from a time series. *Physica D,* **16,** 285–317.