Ensemble Forecasting at NCEP and the Breeding Method

ZOLTAN TOTH* AND EUGENIA KALNAY

Environmental Modeling Center, National Centers for Environmental Prediction, Camp Springs, Maryland

(Manuscript received 24 April 1995, in final form 11 April 1997)

ABSTRACT

The breeding method has been used to generate perturbations for ensemble forecasting at the National Centers for Environmental Prediction (formerly known as the National Meteorological Center) since December 1992. At that time a single breeding cycle with a pair of bred forecasts was implemented. In March 1994, the ensemble was expanded to seven independent breeding cycles on the Cray C90 supercomputer, and the forecasts were extended to 16 days. This provides 17 independent global forecasts valid for two weeks every day.

For efficient ensemble forecasting, the initial perturbations to the control analysis should adequately sample the space of possible analysis errors. It is shown that the analysis cycle is like a breeding cycle: it acts as a nonlinear perturbation model upon the evolution of the real atmosphere. The perturbation (i.e., the analysis error), carried forward in the first-guess forecasts, is ''scaled down'' at regular intervals by the use of observations. Because of this, growing errors associated with the evolving state of the atmosphere develop within the analysis cycle and dominate subsequent forecast error growth.

The breeding method simulates the development of growing errors in the analysis cycle. A difference field between two nonlinear forecasts is carried forward (and scaled down at regular intervals) upon the evolving atmospheric analysis fields. By construction, the bred vectors are superpositions of the leading local (time-dependent) Lyapunov vectors (LLVs) of the atmosphere. An important property is that all random perturbations assume the structure of the leading LLVs after a transient period, which for large-scale atmospheric processes is about 3 days. When several independent breeding cycles are performed, the phases and amplitudes of individual (and regional) leading LLVs are random, which ensures quasi-orthogonality among the global bred vectors from independent breeding cycles.

Experimental runs with a 10-member ensemble (five independent breeding cycles) show that the ensemble mean is superior to an optimally smoothed control and to randomly generated ensemble forecasts, and compares favorably with the medium-range double horizontal resolution control. Moreover, a potentially useful relationship between ensemble spread and forecast error is also found both in the spatial and time domain. The improvement in skill of 0.04–0.11 in pattern anomaly correlation for forecasts at and beyond 7 days, together with the potential for estimation of the skill, indicate that this system is a useful operational forecast tool.

The two methods used so far to produce operational ensemble forecasts—that is, breeding and the adjoint (or ''optimal perturbations'') technique applied at the European Centre for Medium-Range Weather Forecasts—have several significant differences, but they both attempt to estimate the subspace of fast growing perturbations. The bred vectors provide estimates of fastest sustainable growth and thus represent probable growing analysis errors. The optimal perturbations, on the other hand, estimate vectors with fastest *transient growth* in the future. A practical difference between the two methods for ensemble forecasting is that breeding is simpler and less expensive than the adjoint technique.

1. Introduction

It has long been accepted that running an ensemble of numerical forecasts from slightly perturbed initial conditions can have a beneficial impact on the skill of the forecast by means of ensemble averaging (e.g., Leith 1974). Beyond providing a better estimate of the first moment of possible future states, the ensemble members also offer the possibility of estimating higher moments such as the forecast spread, which can be used as an indicator of expected skill, and, ultimately, the full probability distribution. Theoretically, the probability of future states can also be computed through the Liouville equations (e.g., Ehrendorfer 1994) if the initial probability distribution is assumed to be known. However, computational and other problems make the use of these equations unfeasible for numerical weather prediction in the foreseeable future. The only current practical solution to estimating forecast probabilities is through ensemble forecasting.

One of the crucial aspects of an ensemble strategy is the generation of initial perturbations. These perturba-

* Additional affiliation: General Sciences Corporation, Laurel, Maryland.

Corresponding author address: Dr. Zoltan Toth, Environmental Modeling Center, NCEP, 5200 Auth Rd., Rm. 204, Camp Springs, MD 20746.
E-mail: Zoltan.Toth@noaa.gov

tions should realistically represent the span of possible errors in our control analysis. But since the number of ensemble forecast members is strongly limited by computational costs, it is important that this limited number of perturbations optimally sample the initial error probability distribution. As Ehrendorfer and Tribbia (1997) showed for short-range forecasts, among all possible error patterns, the most important to sample are the fastest growing directions.

At the European Centre for Medium-Range Weather Forecasts (ECMWF), a combination of total energy-based singular vectors are used to sample analysis uncertainty for initial ensemble perturbations (Palmer et al. 1992; Molteni et al. 1996). At the National Centers for Environmental Prediction (NCEP, formerly known as the National Meteorological Center), the bred vectors, which represent a nonlinear extension of the Lyapunov vectors (the fastest growing perturbations on the attractor) are used for the same purpose (Toth and Kalnay 1993). In yet another approach, Houtekamer et al. (1996) use multiple analysis cycles (with perturbed observational data and different model formulations) for generating initial ensemble perturbations.

The breeding method has been used for generating initial perturbations at NCEP for operational ensemble forecasts since 7 December 1992. At that time a system with a single breeding cycle was introduced, and a combination of bred perturbations and control forecasts provided 5 global predictions (14 if lagged forecasts were also considered) valid to 10 days every day (Tracton and Kalnay 1993; Toth and Kalnay 1993). A further hypothesis was that multiple breeding cycles, that differ due to nonlinear interactions in the perturbations, can provide initial perturbations for larger ensembles as well. The results presented in the following sections were obtained in the process of investigating optimal strategies for the breeding method. Based on these experimental results, in March 1994 the operational NCEP ensemble was expanded to 7 independent breeding cycles on the new Cray C90 supercomputer, and the forecasts were extended to 16 days. This configuration now provides 17 independent global forecasts valid for more than two weeks every day.

The purpose of this paper is to document the research work that led to the above implementation, using results from off-line experiments. In particular, the major issues considered are the following: 1) the use of multiple breeding cycles to generate initial conditions for a large set of ensemble forecasts (as compared to using only one breeding cycle); 2) the introduction of regionally dependent rescaling in the breeding cycle to reflect the geographically varying uncertainty in the analysis; and 3) the impact of changing the perturbation size on the performance of the ensemble forecasts.

In sections 2 and 3 we discuss basic questions related to ensemble forecasting. In sections 4 and 5 the characteristics and several technical aspects of the breeding method used at NCEP for generating initial ensemble perturbations are presented. Section 6 is devoted to experimental results. A short review about the operational implementation, further discussions, and conclusions are found in sections 7 and 8.

2. Ensemble forecasting and nonlinear filtering

Leith (1974) showed that averaging the ensemble forecasts yields a mean forecast superior to the control forecast, as long as the ensemble perturbations are *representative of the initial probability distribution* of the basic flow around the control analysis. Earlier studies (e.g., Houtekamer and Derome 1994; Toth and Kalnay 1993) using models of different sophistication confirmed Leith's results. In this section we illustrate why this is the case by means of a very simple error growth example. Though the model used below cannot describe the details of error growth arising in the atmosphere due to chaos, it can still offer some insight into the effect of ensemble averaging in an expected sense.

As an example, consider a traveling extratropical low. At the initial time, we assume that the center of the low is analyzed with a small error E_0. We assume that the error will grow exponentially at first, and that later nonlinear effects will lead to error saturation. We can therefore use Lorenz's (1982) simple error growth model:

$$\frac{dv}{dt} = av(1 - v), \tag{1}$$

where $v(t)$ is the algebraic forecast error measured at the center of the system at time t and a is the linear growth rate. We can create a simple ensemble by adding and subtracting a perturbation P from the control analysis. These perturbed analyses will have an error of $E_0 + P$ and $E_0 - P$, respectively. If the perturbation size is smaller than $2E_0$, one of these perturbed analyses will be closer to the true atmospheric solution than the control analysis, though we do not know a priori which one it is. If the perturbed initial conditions are plugged into the error Eq. (1), it is easy to see that the average of the two perturbed forecasts has a smaller error than the control at any forecast time t:

$$v_{\text{con}}(t) > \frac{1}{2}[v_{\text{pos}}(t) + v_{\text{neg}}(t)], \tag{2}$$

where $v_{\text{pos}}(t)$ and $v_{\text{neg}}(t)$ are the errors for the two perturbed forecasts. In Fig. 1 we show an example of the effect of ensemble averaging in this simple model.

We can generalize the above simple example by assuming that we measure the error $v(t)$ over the whole domain of a synoptic system. In this case, the initial error is a vector $\mathbf{E}_0$ of magnitude E_0, whose direction represents a particular spatial distribution pattern. Let us assume that the error growth with time is still given by (1). If the initial perturbation is chosen along the initial error pattern, that is, if $\mathbf{P}_0$ is parallel to $\mathbf{E}_0$, then Eq. (2) is still valid. Ensemble averaging again provides

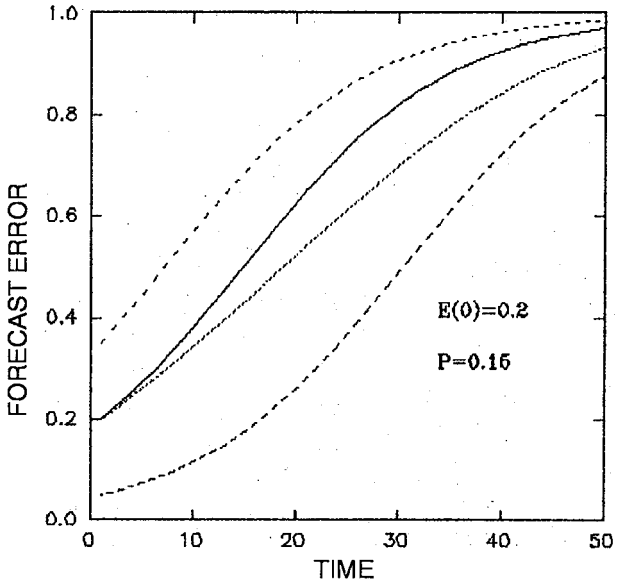

FIG. 1. Example indicating the gain from esemble averaging in a one-dimensional example using Lorenz's error growth equation. The solid curve is the error of the control forecast, the dashed curves are the error of the perturbed forecasts, and the dotted curve is the error of the ensemble mean. Here, $E(0)$ is the initial error of the control forecast and P is the amplitude of the twin perturbations.

a nonlinear filter that removes part of the growing error. As we will see later, much of this improvement is a characteristic of ensemble averaging and cannot be reproduced by simple spatial filtering.

On the other hand, if $\mathbf{P}_0$ is a growing perturbation *orthogonal* to $\mathbf{E}_0$, ensemble averaging will result in a worse forecast than the control, which has *no* error along $\mathbf{P}_0$. The ensemble average will diverge from the control forecast due to the different nonlinear evolution of the $+\mathbf{P}_0$ and $-\mathbf{P}_0$ perturbations, whose growth is also represented by Eqs. (1) and (2), and therefore it will have a larger total error than the control. This example, although admittedly very simplistic, suggests that if possible, one should try to use realistic analysis errors as initial ensemble perturbations, with an amplitude that corresponds with the magnitude of the estimated analysis error. In this ideal case, only the sign (and exact amplitude) of the analysis error would be unknown. Introducing growing perturbations that are not present in the analysis as errors may lessen the positive impact ensemble averaging can offer otherwise. In practice, of course, the analysis error pattern is not known. However, as we will see later, the important, growing part of the analysis error can be estimated through dynamical means with a small number of vectors.[1]

[1] The fact that in the atmosphere all perturbations (and errors in a perfect model environment) tend to turn toward the leading Lyapunov vectors within a couple of days (Lorenz 1965; Szunyogh et al. 1997) ensures that even suboptimal ensemble perturbations, that do not appreciably project onto the actual error field, will have a positive impact beyond a few days lead time.

Ideally one would want to use a large ensemble to represent all possible states of the atmosphere, given the control analysis. In this case the ensemble mean would provide at all lead times the best possible estimate for the future state of the atmosphere. In practice, however, only a small number of ensemble forecasts can be run. For a limited ensemble Leith (1974) showed that hedging the forecasts toward climatology can give an additional improvement in some measures of forecast skill. In this paper (except in section 6d where the effect of spatial smoothing is studied) we restrict our attention to the impact of ensemble averaging.

3. Errors in the analysis

It is clear that with the initial ensemble perturbations we must represent accurately the probability distribution of the state of the atmosphere about our best estimate of the true state of the atmosphere, the latest control analysis. The shape of this probability distribution will depend on what kind of errors we may have in the control analysis. The more likely an error pattern, the higher probability we should assign to the control analysis plus and minus that particular error pattern. This calls for a careful examination of possible analysis errors.

a. Growing and nongrowing errors

A typical operational analysis performed with optimal interpolation or spectral statistical interpolation (see, e.g., Lorenc 1981; Parrish and Derber 1992) is a weighted average of 1) observational measurements and 2) a short-range dynamical forecast (first guess), started from the preceeding analysis. It has been long recognized that the resulting analysis is affected by random errors present in observations. Recently, it was also pointed out that the repeated use of a model forecast as a first guess has a profound dynamical effect on the errors in the analysis (Toth and Kalnay 1993; Kalnay and Toth 1994). The analysis cycle can be considered as the running of a nonlinear perturbation model upon the true state of the atmosphere. The perturbation amplitude (i.e., the analysis error) is kept small by periodic ''rescaling,'' performed at each analysis time, through the use of limited observational data.

In such a nonlinear perturbation setup, it is inevitable that the random errors introduced at each analysis time through the use of data will project into growing directions of the atmospheric flow at later times. This is because the growing components of the error, by definition, rapidly amplify while the decaying components quickly lose their amplitude in the short-range, first-guess forecast (see section 4a). And since the observations are not enough to determine the state of the atmosphere, these dynamically developing errors cannot be removed at the next analysis time: their amplitude can only be reduced (see, e.g., Fig. 9 of Bouttier 1994).

So at the start of the next short-range forecast in the analysis cycle, dynamically developed errors are already present in the initial conditions ready to amplify again. This is especially true over data-sparse regions like the oceans and on small spatial scales not resolved by current observing systems. The result is that the analysis contains both random errors introduced by the most recent observations, and growing errors associated with the instabilities of the evolving flow, dynamically generated (from random errors introduced in earlier analyses) by the repeated use of the forecast first guess. Adjoint-based sensitivity calculations (Langland and Rohaly 1996) and analysis error estimates (Iyengar et al. 1996) also point out the flow-dependent, dynamically conditioned nature of important analysis errors.

b. Which type of error is important?

If we could decompose and follow the development of the errors present in the analysis, we would see that random errors, introduced just at the latest analysis time by observational inaccuracies, will decay initially before projecting, after one or two days, onto growing directions of the evolving basic flow. Such an initial decay was apparent in early experiments in atmospheric predictability (see, e.g., Fig. 4 in Smagorinsky 1969). Even if the random errors are balanced, they will still initially grow very slowly or decay. By contrast, ''growing errors'' will, by definition, amplify, so that they are primarily responsible for short-range error growth. This means that even though the growing errors constitute only a portion of the total analysis error field, their contribution is relatively more important in the forecast error development. Therefore one might want to focus on the growing errors when creating ensemble perturbations. The benefit of this approach has been clearly demonstrated by Ehrendorfer and Tribbia (1997), who found that the fastest growing combinations of possible analysis errors give the best results as initial ensemble perturbations for at least the short range.

Another difference between nongrowing and growing errors is that the dimension of the subspace of possible neutral or decaying perturbations is large [at least of the order of the number of observations, $O(10^5)$], whereas the dimension of the subspace of fast growing perturbations is very much limited by the local (in phase space) dynamics of the atmosphere, $O(10^2)$. In other words, the occurence of any particular, realistic nongrowing perturbation pattern as analysis error is much [$O(10^3)$ times] less likely than that of a fast-growing perturbation. It follows that while the growing subspace can, the large-dimensional nongrowing subspace cannot be sampled well with a small ensemble. We note again that ensemble members/perturbations have to be weighted with their likelihood of being actual analysis errors. The introduction of nongrowing perturbations of realistic amplitudes into a relatively small ensemble will violate this rule, resulting in a gross oversampling of a few nongrowing perturbation patterns (while leaving the rest of the large-dimensional nongrowing subspace unsampled). In case of a small ensemble, the nongrowing perturbations would have to be weighted by the small likelihood of their occurence (as compared to that of the fast-growing perturbations)—otherwise they may have a slight negative impact on the quality of the ensemble. Oversampling a few nongrowing directions is equivalent to introducing perturbations that do not project onto actual analysis errors and, as seen in section 2, these perturbations have the potential for degrading the ensemble. The arguments in this paragraph are in line with our earlier experiments (Toth and Kalnay 1993; Kalnay and Toth 1994; Iyengar et al. 1996) and are related to the results of Ehrendorfer and Tribbia (1997), who showed that at least for the short range, the most efficient description of the forecast error covariance can be derived from an ensemble based on the fastest growing pertrubations (singular vectors). These arguments, however, are not universally accepted in the research community and thus should be considered only hypothetical.

Evidence suggests that the growing part of the error in the analysis can be described by a number of leading fast-growing directions (Toth et al. 1997). It is only in cases when one direction (connected, e.g., to explosive intensification of a weather system) exhibits much higher perturbation growth rate than any other, that a single perturbation may be adequate for representing initial uncertainty. Otherwise, a number of perturbations is necessary for a successful description of initial and forecast uncertainty. One should note, however, that when at later lead times the forecast error/perturbation development becomes highly nonlinear, the role of initial perturbations diminishes and the performance of different perturbation methodologies may become similar.

4. Local Lyapunov vectors and their estimation through breeding

Since the important, growing component of the analysis error occupies only a relatively small subdomain in the phase space, and it depends on the basic flow, it is possible to compute estimates of possible growing analysis errors through dynamical methods.

a. The breeding method

For this purpose, Toth and Kalnay (1993) proposed a method called *breeding of the growing vectors* of the atmosphere (BGV). This procedure consists of the following simple steps: (a) add a small, *arbitrary* perturbation to the atmospheric analysis (initial state) at a given day t_0 (or to any other basic state, such as a long model run); (b) integrate the model from both the perturbed and unperturbed initial conditions for a short period $t_1 - t_0$ (e.g., 1 day, as for all experiments in this paper); (c) subtract one forecast from the other, and (d)

scale down the difference field so that it has the same norm (e.g., rms amplitude or rotational kinetic energy) as the initial perturbation. This perturbation is now (e) added to the analysis corresponding to the following day t_1, and the process (b)–(e) is repeated forward in time. Note that once the initial perturbation is introduced in step (a), the development of the perturbation field is dynamically determined by the evolving atmospheric flow.

By construction, this method ''breeds'' the nonlinear perturbations that grow fastest on the trajectory taken by the evolving atmosphere in the phase space. One can decompose the initial perturbation $P(t_0)$ into growing and decaying components. Let us consider the development of a small perturbation on top of a nonlinear model trajectory (i.e., the difference between two nonlinear forecasts). At the end of a short-range integration, by definition, the relative contribution of the growing component will be larger, whereas that of the decaying component is smaller than at initial time. And after a few cycles, the decaying component will become negligible.

Note the similarity between the breeding method and the analysis cycle: in both cases, a nonlinear perturbation model is run with regular rescaling. In the case of breeding, the perturbation is run over the analyzed states. The perturbations are defined with respect to the analysis and then rescaling is done in a deterministic fashion, so that stochastic (or decaying) components are eliminated from the perturbations, as discussed above. The resulting bred perturbations are determined purely by the dynamics of the system. On the other hand, the analysis cycle is run based on observed data. The perturbations here can be defined as the difference (error) between the analysis/first guess and the true state of the atmosphere (which is unknown). In the first-guess short-range forecast, the growing components of this error will amplify. However, at the next analysis time observational data will be used to reduce the difference between the analysis and the true state of the atmosphere. The observed data contains random noise that will be periodically reintroduced into the analysis. Consequently, the errors present in the analysis, beyond the growing error connected to the use of short-range forecasts as first-guess fields, also contain a random or stochastic component.

The similarity between the analysis and breeding cycles is stronger on the small scales and over data-poor regions where the analysis is based primarily on the first-guess forecast (Daley 1991). Over data-rich regions and on the larger scales the observations can better remove the growing errors from the first guess but since current operational analysis techniques do not have knowledge about the flow-dependent part of these errors, this removal process cannot be complete (see Iyengar et al. 1996).

b. Lyapunov vectors

Theoretically, the bred perturbations are related to the local Lyapunov vectors of the atmosphere (LLVs, see Trevisan and Legnani 1995). The Lyapunov exponents (λ_i) have been widely used for characterizing the behavior of simple dynamical systems:

$$\lambda_i = \lim_{t\to\infty} \frac{1}{t} \log_2\left[\frac{p_i(t)}{p_i(0)}\right], \tag{3}$$

where p is a *linear* perturbation spanning the phase space of the system with orthogonal vectors. Note that while the first Lyapunov exponent is uniquely defined at least for Hamiltonian systems, the rest of the spectrum is derived via a periodic reorthogonalization of the perturbation vectors (see e.g., Benettin et al. 1980) and hence will depend on the frequency of reorthogonalization. The λ_i's can be computed either for the whole attractor (global Lyapunov exponents) or can be interpreted pointwise, where the growth ratio is evaluated for an infinitesimal time interval at t [local Lyapunov exponents, see, e.g., Trevisan and Legnani (1995)]. The leading Lyapunov exponents are associated with predictability properties of dynamical systems, namely how fast nearby trajectories diverge (or converge) on the attractor. Most importantly, if a system has at least one positive global Lyapunov exponent, its behavior is chaotic, that is, arbitrarily close points on the attractor will eventually separate into unrelated points (Wolf et al. 1985).

When the Lyapunov exponents are interpreted locally, each of them can be associated with a perturbation vector, **I**. The first of these vectors, with the largest exponent, can be uniquely determined: *any random perturbation* introduced an infinitely long time earlier develops linearly into the leading local Lyapunov vector. The importance of this property of LLVs in meteorology was first recognized by Lorenz (1965), who found in his experiments with a simple linear perturbation model that initially random perturbations had a strong similarity after 8 days of integration. Indeed, our breeding experiments with a state-of-the-art general circulation model indicate that one needs only a few days of integration (3–4 days) in order to get a good estimate of the leading local Lyapunov vectors of the atmosphere. These LLVs are the vectors that grew asymptotically fastest during a time period *leading* to the analysis. Hence they are likely to dominate growing analysis errors and, because of their sustainable growth, also the forecast errors.

c. Extension of Lyapunov characteristics into the nonlinear domain

There is an extensive body of literature on the global, and more recently, on the local Lyapunov exponents of simple dynamical models. These studies, however, use a *linear tangent* model approach and are concerned only

about error growth in a linear sense. In some studies, a regular rescaling of the perturbations, also used in the breeding method, has been applied. Rescaling in these linear methods, however, is used to avoid computer overflow, not to prevent/control nonlinear saturation (see, e.g., Benettin et al. 1976; Shimada and Nagashima 1979). New aspects of the breeding method proposed by Toth and Kalnay (1993) are that perturbations are developed for a 1) complex physical system, in a 2) nonlinear framework, at a 3) high horizontal and vertical resolution, and that it is 4) the perturbation vectors (and not only the exponents) that are studied and used for real world practical applications. With the breeding method it is possible to estimate the leading local Lyapunov vectors of the atmosphere with a comprehensive nonlinear perturbation model including all physical parameterizations.

Nonlinearity plays a crucial role in complex systems where a host of different physical processes occur, associated with widely different growth rates and nonlinear saturation levels. A traditional linear approach may find the strongest instability of the system (such as convection) but this may be associated with processes with a very low nonlinear saturation level. For perturbation amplitudes larger than the saturation level, these perturbations will decay and are therefore irrelevant. Hence the bred vectors can be considered as an extension of the notion of LLVs into the nonlinear perturbation domain. Note that the perturbation amplitude is the only free parameter in the BGV method and that the bred vectors, just as the linear LLVs, are not sensitive to the type of norm used for rescaling.

d. Multiple breeding cycles

When a breeding cycle is started, an arbitrary initial perturbation field is added upon the control analysis. After three or four days of breeding, most of the originally decaying components in the perturbation disappear and the perturbation growth rate reaches an asymptotic value around 1.6 per day (with a perturbation amplitude of 1% in total climatological rms variance). After this time, the perturbations that remain are those that could produce the largest growth over the preceeding 3 days or so, given the initial perturbations. In Fig. 2 average growth rates are shown for a 5-day period for breeding cycles with different perturbation amplitudes. The independently run breeding cycles had a 24-h rescaling frequency and were started several days before the 5-day evaluation period. It is clear from Fig. 2 that the growth rate in a breeding cycle depends on the amplitude of perturbations but is always larger than that obtained with other perturbation methods such as scaled lagged averaged forecasting (Ebisuzaki and Kalnay 1991, not shown), difference fields between short-range forecasts verifying at the same time (Toth and Kalnay 1993), or Monte Carlo perturbations (which have a wider range of growth rate values than the dynamically conditioned perturbations).

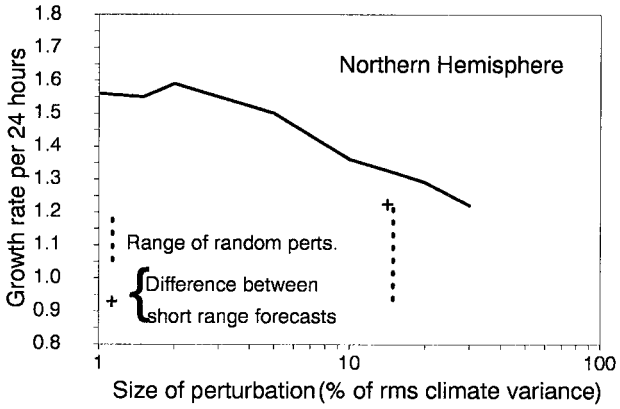

FIG. 2. Daily amplification of bred perturbations with different initial perturbation sizes over the Northern Hemisphere, computed for the period 23–27 February 1992. The range of amplification factors for different random (Monte Carlo) balanced perturbations is shown as a vertical dotted line. Average amplification factors for difference fields between short-range forecasts of different length verifying at the initial time of perturbed forecast integrations are also shown with a plus sign.

When the perturbation amplitude is in the range of 1%–10% of the natural variability, the perturbations are primarily associated with extratropical baroclinic instability. Within this amplitude range, the growth rate and shape of perturbations are largely independent of the perturbation amplitude. However, if the perturbation amplitude is reduced to less than 0.1% rms variance, then the growth rate increases enormously, with an amplification factor well above 5 per day. This is because the fastest growing perturbations in the model atmosphere are, in fact, related to convective and not baroclinic instability (see Fig. 3). The perturbations at this amplitude are highly nonlinear and are primarily associated with convection. The mostly tropical convective perturbations, however, saturate at less than 1% amplitudes, much smaller than the estimated size of the analysis errors (5%–10% of the rms of the natural variability). The patterns associated with convection are also present at larger perturbation amplitudes but are not detectable because they saturate at amplitudes much smaller than those of baroclinic instabilities. This also explains why convective instabilities do not produce dominant analysis errors.

Since breeding is a nonlinear process, the perturbations in the 1%–10% rms variance range, though primarily determined by the dynamics of the system, also depend on the perturbation at previous times, namely on how those perturbations project on certain growing directions, and on the small-scale forcing convection provides to the larger scales. This forcing (see Fig. 3) is largely stochastic with respect to the baroclinic processes that dominate perturbation development in the amplitude range of 1%–10% rms variance. If we start

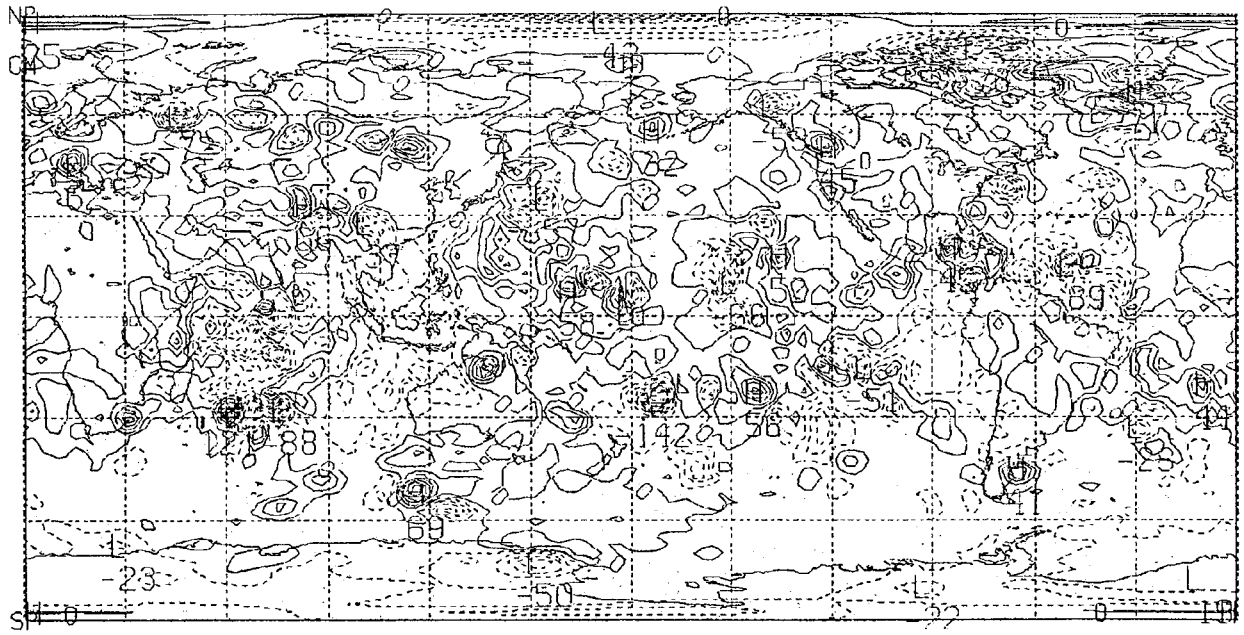

FIG. 3. An example of bred perturbations at relatively small amplitudes: 500-hPa streamfunction perturbation on 15 February 1992 with a perturbation amplitude of 0.015% total rms variance (equivalent to about 0.012 m in 500-hPa height). The contour interval is 0.001 m^2 s^{-1} and the labels are scaled by 10^4.

independent breeding cycles with different arbitrary initial perturbations, we find that after a transient period of about 3 days, the perturbations in the different cycles can be quite similar over any region (except for their phase, and to some extent, their amplitude, which are arbitrary) but only over roughly half of the global domain. Table 1 shows the results of a subjective intercomparison using 20 independent breeding cycles on a typical day. The local shape of the perturbations were compared to those observed in one perturbation (number 17) over three selected regions both over the Northern and the Southern Hemispheres (see Fig. 4). Perturbation 17 was chosen because it showed a pattern in the selected regions that many other perturbations reproduced. A plus, minus, or a blank indicate whether the same perturbation was observed with the same or opposite sign or whether a different perturbation was observed. When the same comparison is made with bred perturbations valid on different dates, even as close as 2 days apart, there is almost no correspondence among the perturbations, showing that the bred growing vectors crucially depend on the basic flow and its recent evolution.

TABLE 1. Subjective comparison of perturbations from 20 independent breeding cycles on 23 May 1992. Regional perturbations in three areas over both the Northern and Southern Hemispheres, marked with boxes in Fig. 4 and numbered correspondingly from left to right, are compared. If a perturbation in another cycle is very similar to that in cycle 17, a plus or minus sign appears, depending on the sign of the perturbation.

	NH 1	NH 2	NH 3	SH 4	SH 5	SH 6
br 1			+	−	−	
br 2	−	+	+	−	+	+
br 3	+	−	−			+
br 4		+	+		−	
br 5			−	+	−	
br 6	−					
br 7			−	+	+	
br 8	+		+	−	+	+
br 9		−	−	+	+	
br 10	−	−	−		−	
br 11	−	−	−	+	+	+
br 12	−	+	+			+
br 13	+		+			+
br 14				−		−
br 15	+	−		−		+
br 16	+				−	−
br 17	+	+	+	+	+	+
br 18	−			+	+	
br 19		+	+			
br 20		+		−		+
Total	12	11	14	11	12	11

From this experiment we note that the linear Lyapunov vectors [as well as the singular vetors, see, e.g., Molteni et al. (1996)] are regional in character and are associated with areas of high instability in the atmosphere, such as baroclinically unstable regions. In the areas where the perturbations are very similar, the largest Lyapunov exponent must have a value much larger than the succesive Lyapunov exponents. Over the rest of the domain, different perturbations appear in the independent breeding cycles, suggesting that the first few Lyapunov vectors associated with baroclinic instability have similar growth rates, and the appearance of one or another in any cycle depends on the details of perturbation evolution in that cycle a few days earlier and

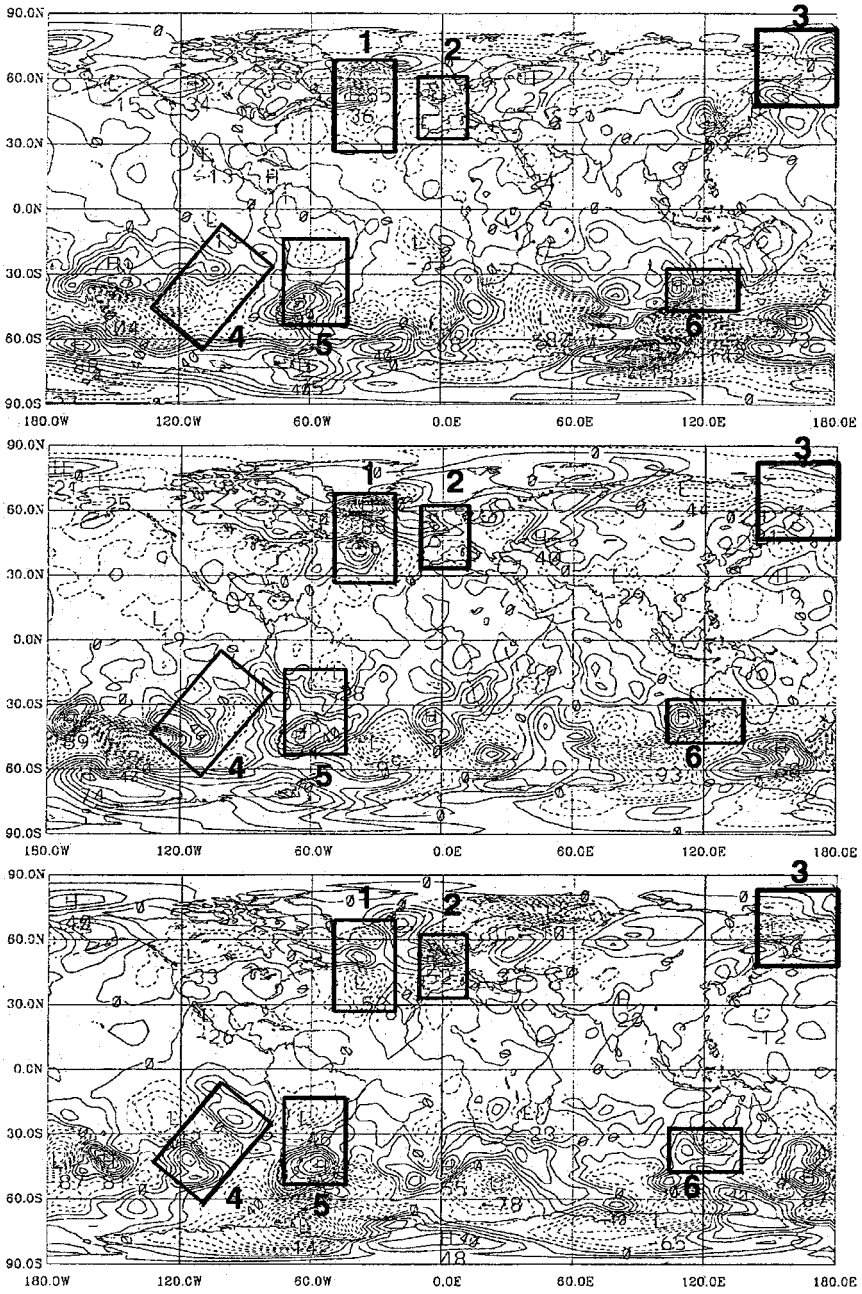

FIG. 4. The 500-hPa streamfunction perturbation fields from three independent breeding cycles (with hemispherically constant rescaling) for 23 May 1992. The three cycles were started with independent initial perturbations six days earlier. The six marked boxes correspond to the areas considered in Table 1. Panels (a), (b), (c) correspond to breeding cycles br8, br12, and br17 in Table 1, respectively.

also on the details of instantaneous stochastic forcing (convection).

We can conclude from the above experiments that each global perturbation pattern is a superposition of a number of regional features, perhaps of the order of 10–20 in each hemisphere, which, in turn, are primarily associated with baroclinically unstable regions of the evolving basic flow. And as the basic flow has many degrees of freedom, so does the global perturbation field. It follows that the phase and amplitude of the regional patterns (and in the case of competing LLVs with similar growth rates, the patterns themselves) in one area are independent of those in remote areas. This ensures that the bred perturbations from independent cycles are *quasi-orthogonal* over the global domain, without imposing any constraints.

The subjective evaluation presented above is confirmed by numerical calculations done during July–September 1995 using five vectors, each defined as the difference between 24-h pairs of ensemble forecasts from five pairs of operational breeding cycles (see section 7 for details). On average, the absolute value of the correlation between two bred vectors over the globe during this period is 0.27—that is, only about 7% of the variance in any of the vectors can be explained by another vector. The low average correlation value indicates that there are approximately 13 independent degrees of freedom in the space spanned by the five bred vectors.[2] Again, over smaller areas the correlation can temporarily reach values close to one, indicating that a particular perturbation is able to grow much faster than any other one.

In summary, a bred global perturbation is a superposition of regional patterns, each of which is a combination of the leading local (in phase space) Lyapunov vectors (interpreted at a certain nonlinear perturbation amplitude) in that area of the atmosphere. The weights on the individual local Lyapunov vectors are randomly assigned by the arbitrary initial perturbation and the stochastic small-scale forcing but are, in a statistical (ensemble average) sense, proportional to the Lyapunov exponents themselves. The bred perturbations are therefore not unique in a strict sense but only in a statistical, ensemble average sense. And the more independent breeding cycles we have, the better we can span the space of possible fast-growing analysis errors. Nonlinear breeding hence can be considered as a generalization of the notion of Lyapunov vectors for complex nonlinear systems. Because of nonlinear interactions and stochastic forcing by convection, and because of the existence of many regional features, different breeding cycles do not converge to a single leading LLV but rather span the subspace of the fastest growing perturbations that can occur at the chosen level of perturbation amplitudes.

e. Optimal perturbations and Lyapunov vectors

There is another method to determine fast-growing perturbations of dynamical systems. This linear method uses the linear tangent and adjoint of a full model to compute the initial perturbations that grow fastest over a specified period, measured with a given norm (Lorenz 1965). In its application to ensemble forecasting at ECMWF (see, e.g., Molteni and Palmer 1993; Buizza and Palmer 1995; Molteni et al. 1996), the fastest growing perturbations are determined for a 48-h forecast trajectory created by the full model and the chosen norm is based on total energy. The optimal vectors (which are also called the singular vectors of the linear propagator, SVs) are those that amplify most over the optimization period, given the norm and other possible constraints. A more detailed comparison of the Lyapunov versus optimal vectors appears in Szunyogh et al. (1997) and Toth et al. (1996) so only a brief discussion of the subject is given here.

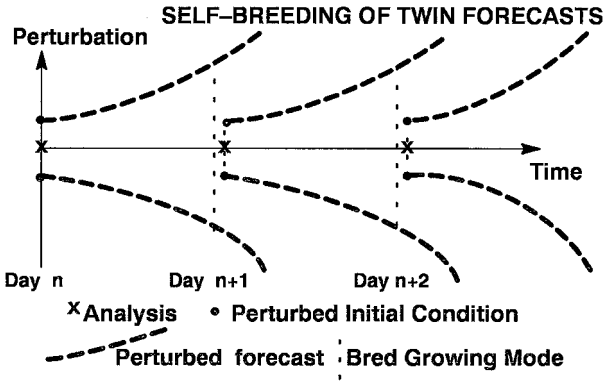

FIG. 5. Schematic of a self-contained breeding pair of ensemble forecasts.

Ehrendorfer and Tribbia's (1997) results strongly suggest that the best initial ensemble perturbations are the SVs, when they are defined given the likelihood of different analysis error patterns. In its current ensemble application at ECMWF, the SVs are computed without information on the likelihood of different analysis error patterns. The breeding method, on the other hand, is designed to estimate the fastest growing errors that can be present in the analysis.

In terms of practical considerations, the breeding and the optimal (or singular) vector methods used at NCEP and ECMWF, respectively, have several differences that may or may not be important for certain applications. 1) Computational efficiency: the adjoint technique is expensive whereas breeding is cost-free, apart from running the ensemble forecasts themselves (see Fig. 5). As a result, the breeding can be performed at full spatial resolution while for the optimal perturbations technique this is computationally impractical. 2) Breeding is performed with the full nonlinear model with physics while the optimization is currently done with a tangent linear system with limited physical parameterizations. 3) Localization: In contrast to the breeding method, the SV methodology can provide information on where a particular linear perturbation originates from. This may be especially valuable for adaptive observation strategies that aim at reducing certain aspects of the forecast error by taking extra upstream observations in sensitive areas.[3] On the negative side of the SV methodology is

[2] Correlation is the cosine of the angle between two directions in phase space. In two dimensions, the average correlation is $1/(2)^{1/2}$; in three, it is $1/(3)^{1/2}$, and in n dimensions it generalizes to $1/(n)^{1/2}$. The best estimate for dot is 13 given a correlation value of 0.27 [$1/(13)^{1/2} = 0.27$].

[3] A cheap alternative to using a traditional linear adjoint algorithm for targeting is a singular value decomposition in the subspace of an already existing ensemble (Bishop and Toth 1996).

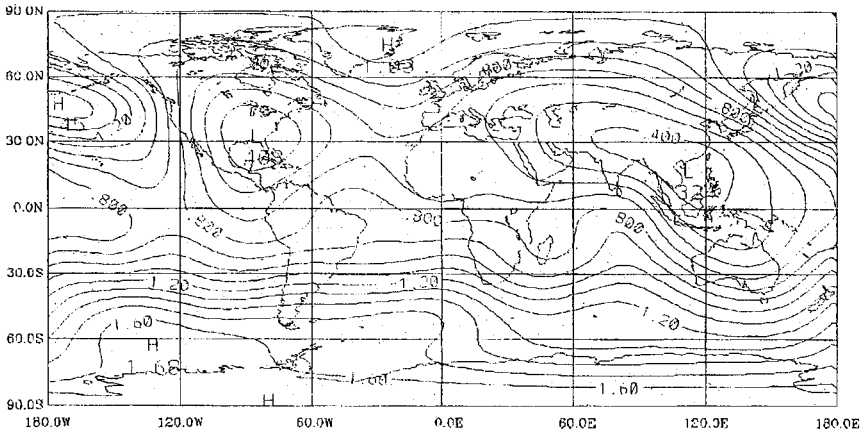

FIG. 6. Relative regional uncertainty (for 500-hPa streamfunction) present in the control analysis as determined from the rms difference between two analyses from independently run NCEP analysis cycles in April–May 1992. The analysis cycles were practically identical except that the initial first guesses differed slightly. The values shown are smoothed and the overall global mean is scaled to one.

that the fastest optimal perturbations cover only a small fraction of the geographical domain with relatively large amplitudes and only over the extratropics (cf. Figs. 5 and 6 in Buizza 1995), whereas with breeding, the fastest growing regional patterns are automatically determined for the whole globe (including the Tropics), and not only for those regions with highest growth rate. 4) The performance of the SV-based ECMWF ensemble and the breeding-based NCEP esnsemble was compared by Zhu et al. (1996), who found that the NCEP ensemble verified somewhat better in terms of probabilistic and other skill measures.

5. Ensemble perturbations

From the discussion about LLVs above, one could draw the conclusion that it does not really matter what initial perturbations are used for medium- or extended-range predictions since all linear perturbations turn into very similar vectors after a few days of integration. However, one should keep in mind that ensemble forecasts, just as the control forecast, are *nonlinear* integrations. With a perturbation size similar to the estimated size of errors in the analysis, nonlinearity becomes important after about two days, and earlier than that in fast developing synoptic systems. And since nonlinearity prevents the actual forecast errors from fully converging to the leading LLV, it is necessary that the ensemble perturbations realistically represent the initial uncertainty in the analysis, otherwise, as discussed in section 2, our ensemble will be suboptimal.

In this section we discuss several additional technical points about the breeding method that were investigated in the process of implementing operational ensemble forecasting at NCEP.

a. Regional rescaling

The breeding method was originally used at NCEP with hemispherically determined rescaling factors (Toth and Kalnay 1993). Depending on the hemispheric rms magnitude of the perturbation, a constant factor was applied over each hemisphere and a linearly interpolated value was used in the Tropics in the rescaling. While this method is adequate for studying the instabilities of the atmosphere as they are represented in our numerical models, it may not be optimal for ensemble forecasting. The perturbations should reflect not only the shape, but also the size of analysis errors. Consequently, we want to have larger regional perturbation amplitudes in regions sparsely observed, and vice versa. With hemispherically fixed rescaling, the perturbation amplitudes will be largest in the areas of strongest instabilities. While these areas are generally over the poorly observed oceans, they do not necessarily correspond to the regionally dependent uncertainty in the analysis.

To estimate the geographically dependent uncertainty in the analysis, we used a technique similar to that of Augustine et al. (1992). Two independent analysis cycles were run for a 30-day period in April–May 1992. The cycles were identical except that in one of them the first guess field was an ensemble average of two first guesses, perturbed by bred vectors with positive and negative signs. The two analyses gradually diverged from each other until, a few days later, the difference saturated. Beyond this time, we took the average of rms difference fields between corresponding pairs of analyses. Figure 6 shows the average difference field in the streamfunction at a model level approxiamtely at 500 hPa, scaled so that the global average is one, and smoothed with a Gaussian filter on a sphere (J. Purser 1993, personal communication). By using this spectral

filter, equivalent to T6–T7 (about 2000 km) resolution, we avoid the aliasing problem associated with simple truncation in wavenumber space. Different filtering characteristics are described in terms of ''equivalent'' triangular truncation. Over the Northern Hemisphere, the dominant features of the analysis uncertainty field are the minima over North America and Eurasia, especially over the eastern part of the continents, and the high values over the Pacific Ocean. This corresponds well to the good rawinsonde coverage over the continents. Due to the use of dynamical first guess, the information from the observations is ''transported'' eastward, resulting in minima over the eastern part of the continents.

While there is a hint of a similar behavior in the Southern Hemisphere east of Africa and over eastern Australia, there is more zonal symmetry, and the amplitude of the uncertainty increases poleward. Such behavior is also consistent with the uniform observational coverage provided by satellite temperatures and by the fast growth of perturbations in the strongly baroclinic southern high latitudes.

Note that with the above procedure, we can estimate the amplitude of growing errors in the analysis,[4] which, as we discussed in section 3b, are assumed to be more important in ensemble forecasting. Optimal interpolation (OI) could also be used to estimate the distribution of the analysis errors (Gandin 1965), but such an estimate is very dependent on the assumed error covariances for the forecast and the observations. In addition, the OI estimate would not properly account for the growing component of the error. Therefore we believe the OI estimate would be less reliable than the empirical procedure we have used here.

In a breeding cycle specifically modified for ensemble perturbations, we determine the scaling factor as a function of horizontal location. The perturbation amplitude is measured and rescaled regionally in a smooth fashion, to a level corresponding to the values shown in Fig. 6. At points where the perturbation amplitude (globally scaled to 1) is below that in Fig. 6, no rescaling is applied. So a perturbation traveling into a poorly observed oceanic area is allowed to grow freely, while those reaching a well-observed area are scaled back to the size of the estimated analysis error. Since the regional rescaling is done in a smooth fashion, most of the balance naturally present in the bred perturbations is preserved. With regional rescaling we still retain the capability of changing the overall global or hemispheric amplitude but the smoothed relative geographical distribution is left intact. Based on the varying degree of hemispheric analysis uncertainty, operationally two geographical masks have been designed, one for the northern winter and another for the northern summer half year. More recently, monthly uncertainty estimates were derived by comparing the NCEP operational analysis and reanalysis fields (Iyengar et al. 1996).

Medium-range ensemble forecasts performed with the breeding method modified for regional rescaling (using self-breeding, see Fig. 5) showed an improvement in the skill of the ensemble mean over the Southern Hemisphere and also over the Tropics (especially for short lead times, when compared to the hemispherically rescaled perturbations) while there was no change over the northern extratropics. Over our 12-day experimental period between 23 May and 3 June 1992, we also tested applying the regional rescaling outside of the breeding cycle, to modify only the initial ensemble perturbations, but found that larger changes were necessary after each cycle and that the forecast results were not as good.

Since perturbation growth rates increase away from the equator, the above described regional rescaling procedure results in perturbations that are somewhat smaller/larger than desired over the Tropics/high latitudes (Iyengar et al. 1996). Let us assume, for example, that perturbation growth is largest over the poles. Since the rescaling factor applied at the pole is based on perturbation growth measured over a larger disc around the pole (over which average error growth is smaller than it is at the pole), the actual rescaling factor that is applied will not reduce the amplitudes enough at the higher latitudes. Work is under way to correct this problem.

b. Centering the ensemble around the control analysis

Since our best estimate of the true state of the atmosphere is the control analysis, we must center the ensemble perturbations around this field. This can be easily done by adding and subtracting the same perturbation to the control analysis (e.g., Ebisuzaki and Kalnay 1991). In this setup, 2n perturbations are derived from n independent breeding cycles (or from other orthogonal vectors). However, a case can be made for using each perturbation only once, thus possibly improving sampling (J. Purser 1992, personal communication). We tested this hypothesis by averaging 2n *independent* perturbations and then removing their average from each individual perturbation vector. The resulting medium-range ensemble integrations, however, had inferior forecast skill as compared to the identically sized paired ensemble setup: the improvement upon the control forecast obtained with the centered single perturbations was less than two-thirds of that obtained with the ensemble of positive and negative pairs of perturbations (see Table 2). The implication is that the nonlinear ensemble filtering mechanism discussed in section 2 is not as effective if the perturbations, though centered initially in a linear sense, are not paired. This must be

[4] We note here that while the relative magnitude of the perturbations (shown in Fig. 6) is realistic, their absolute magnitudes were several times below other estimates. When in one of the two analysis cycles the horizontal resolution was doubled to T126, the absolute values became in line with other error estimates (Iyengar et al. 1996).

TABLE 2. Comparison of ensembles generated by single bred perturbations (i.e., centering individual perturbations on control analysis, singles) and those generated by positive-negative pairs of perturbations (pairs) for May 23–28 1992 with 5%/10% initial perturbations for the Northern and Southern Hemispheres, respectively. (a) PAC skill scores at day 6. (b) Comparative verification as a function of lead time, Northern and Southern Hemispheres combined. Numbers indicate how many times either technique is better than the other in terms of PAC.

(a)	Control	Singles	Pairs
NH	0.680	0.687	0.692
SH	0.510	0.536	0.552
NH/SH Combined	0.595	0.611	0.622

(b) Lead time (days)	Pairs–Singles (wins)
1	8–4
3	7–5
4	7–5
5	9–3
6	10–2

due to higher than second-order nonlinear effects present in the ensemble perturbations (J. Purser 1995, personal communication). Our results are confirmed by independent experiments run at ECMWF (R. Buizza 1995, personal communication).

6. Ensemble forecasting results

In this section we will give an overview of ensemble forecasting experiments performed in order to test possible operational configurations. All experiments were done with a T62/18 level version of the NCEP Medium Range Forecast (MRF) model (Kanamitsu et al. 1991). The period used in these experiments is the 40 days between 6 May 1992 and 14 June 1992[5] (or a subperiod of it, where noted). Except in section 6g, 10-member ensemble forecasts are evaluated. The initial ensemble perturbations were derived from five independent breeding cycles with regional rescaling every 24 h, using the self-breeding algorithm of Fig. 5. To center the ensemble mean on the control analysis at initial time, each of the five perturbations was both added to and subtracted from the analysis. The quality of the ensemble forecasts is estimated using two measures: the skill of the ensemble mean forecast and the spread of the ensemble.

a. Measures of ensemble quality

At any lead time, members of the ensemble can be averaged. The mean ensemble forecast is then verified against the corresponding analysis much the same way as the control forecast. As a measure of skill, we use the forecast/analysis pattern anomaly correlation (PAC) measured over three separate belts over the globe: the Northern and Southern Hemisphere extratropics (20°–80° latitude belts) and the Tropics (±20° latitude). All scores are computed for the streamfunction field at a sigma layer close to the 500-hPa height level (resulting in scores very close to those for 500-hPa geopotential height). To compute the anomalies, monthly climatology is used based on analized data for the 1985–91 period. The rms errors were also computed but are not reported here because they led to identical conclusions as the PACs. Forecast PACs for different types of ensembles are compared to those for the control forecast to see if they represent an improvement due to nonlinear ensemble filtering.

The spread of the ensemble is determined as the average of the difference fields between the individual ensemble forecasts and the ensemble mean. Since kinetic energy is used as a norm in operations (see section 7), the difference at each grid point is defined as the square root of the kinetic energy in the difference (or error) field. (The use of streamfunction differences leads to very similar results.) The spatial distribution of the spread is considered as a prediction of the spatial distribution of the actual error in the control forecast, which is measured in the same way, in units of square root of kinetic energy. After setting the mean of both the forecast spread and observed error fields to zero, their correlation is computed (spread/error PAC). Spread/error PACs are computed only in the T3–T15 range of equivalent spatial resolution using the spectral filter mentioned above (Purser 1993, personal communication). Another (inverse) measure of spread is the average of PAC values computed between the control forecast and each ensemble member. Time correlations between spread and error statistics are also computed (in which case the spatial mean of the spread and error fields is not removed).

b. Size of the initial perturbation

In the section on regional rescaling (section 5a), we indicated that the overall size of the initial perturbations is an important parameter that has to be chosen to reflect the size of initial error in the analysis. An estimate of the analysis error can be derived from optimal interpolation analysis techniques (see, e.g., Gandin 1965; Buizza 1994). However, since these estimates are subject to the statistical approximations made within the analysis scheme, we attempted to optimize the overall perturbation size experimentally by verifying ensemble means for ensembles initiated with different initial amplitudes for the bred perturbations. The perturbation size is measured on the 500-hPa streamfunction field. We note that the wintertime Northern Hemisphere (NH) natural rms variability of the streamfunction field is around 8 500 000 $m^2\ s^{-1}$ (whereas it is around 80 m for geopotential height).

To estimate the optimal size of the initial perturba-

[5] Easy data access and computational convenience were the reasons for selecting this period. Clearly, forecasts separated by several days and covering all seasons would provide a more reliable dataset.

tions, over a 15-day period in 1992 we performed tests with different perturbation amplitudes between 3% and 20% of the NH winter variability for the NH and 6% and 40% for the Southern Hemisphere (SH), respectively, and recorded the skill score for the mean of the different ensembles. Since at T62 resolution much of the small perturbations develop linearly in the first 24 h time range, the ensemble mean of perturbations equal to or less than 10% of the rms variance (standard deviation) is not appreciably different from the control at one day. Though at this short lead time the skill of the ensemble mean cannot be directly used to determine the optimal perturbation size,[6] it is important to note that perturbed forecasts with 10% initial ''error'' for the NH and 20% for the SH diverged from the control as much as the control forecast diverged from the verifying analysis (not shown), suggesting that the optimal perturbation size is around this magnitude. This agrees well with other estimates for the error in global analysis fields. Kalnay et al. (1996) found that the difference between independent height analyses between 850 and 200 hPa from various centers is between 7 and 12 m for the NH and between 12 and 25 m for the SH. P. Caplan (1994, personal communication) estimated differences in the same range, with the SH uncertainty being about double of that for the NH. These estimates, along with other information such as improvement in forecast skill suggest that the quality of our atmospheric analysis has been considerably improved since the mid 1980s when Daley and Mayer (1986) estimated the global analysis error to be between 15 and 20 m at 500 hPa.

In Table 3 we show the results of using different perturbation sizes for day 3 to 9, comparing them with perturbations of size 10% for the NH and 20% for the SH. While some details in Table 3 must be due to sampling fluctuations, there is a clear signal apparent: for longer lead times, larger amplitudes give better results. For example, we find that for the NH, at day 3 an amplitude of 7.5% is slightly better than 10%, whereas at day 9, 12.5% is better. This increase in the optimal initial size with forecast length is also observed in the SH: at day 3 a size of about 25% is better, whereas at day 9 a size of 30% is more effective in increasing the skill of the ensemble average. The difference between the best and worst performing perturbation size at day 9 is around 0.02 and 0.05 for the NH and SH, respectively.

In a perfect model environment, the optimal perturbation size should not depend on lead time. However, our models are imperfect, which means that forecast errors are growing not only due to the initial difference but also due to model deficiencies (Reynolds et al. 1994). Part of the model-generated errors project on growing directions and act like amplifying errors due to the initial uncertainty, whereas others appear as a forecast bias. The model errors that project onto growing directions can be dealt with, to some extent, as an extra amplitude term in the initial error field, explaining why the optimal perturbation amplitude increases slightly with increasing lead time.

TABLE 3. The effect of the size of initial perturbations on the performance of 10-member ensembles for 23 May–6 June 1992. At the different lead times, PAC scores are computed for the mean forecast from different ensembles. Shown is the relative performance of each perturbation size with respect to (a) 10% (Northern Hemisphere) and (b) 20% (Southern Hemisphere) perturbation size, in terms of how many times the PAC for the experiment was better/worse (wins vs losses, W/L) and average improvement (AI). Results judged subjectively the best based on these two objective measures are highlighted in bold.

Perturb. size (% rms variance)	Day 3		Day 6		Day 9	
	W/L	AI	W/L	AI	W/L	AI
(a) Northern Hemisphere						
5	5–6	−	6–6	−	4–8	−
7.5	**7–5**	**+**	6–6	−	4–8	−
12.5	2–9	−	**5–7**	**+**	**6–6**	**+**
15	2–10	−	4–8	−	4–8	−
20	2–10	−	1–11	−	6–6	−
(b) Southern Hemisphere						
10	2–10	−	5–7	−	3–9	−
15	4–8	−	6–6	−	3–9	−
25	**7–5**	**+**	**8–4**	**+**	8–4	+
30	7–5	+	7–4	+	**7–5**	**+**
40	3–9	−	5–7	−	8–4	+

Based on the above results we have fixed the initial amplitude of perturbations in the remainder of this study at 12.5%/25% rms standard deviation for the NH/SH, respectively. Note that this amplitude is larger than optimal for short lead times but is about optimal for the medium and extended range.

c. Ensemble mean forecasts

Figure 7 shows the PAC scores for the control and ensemble mean forecasts for the experimental period. First we should note that ensemble averaging has a greater impact over the winter (in this case the SH) than over the summer hemisphere. This is in line with the fact that the natural variance, which was found to be linked with the impact of ensemble averaging in earlier studies, increases over the winter season. To understand the seasonal variations in the relative merit of the ensemble mean, one has to realize that baroclinic disturbances are probably the sole major source of instabilities in the winter. These instabilities have a relatively long life cycle (few days) and a large saturation amplitude. Consequently, baroclinic instabilities are directly responsible for a large portion of wintertime forecast errors. And since at T62 resolution these instabilities are well resolved, the ensemble based on these perturbations

[6] The signal is hard to detect because the errors in the verifying analysis are not much smaller than the short-range forecast errors. Had we used observational data for verifications instead of analysis fields, we may have been able to find a signal even at very short range.

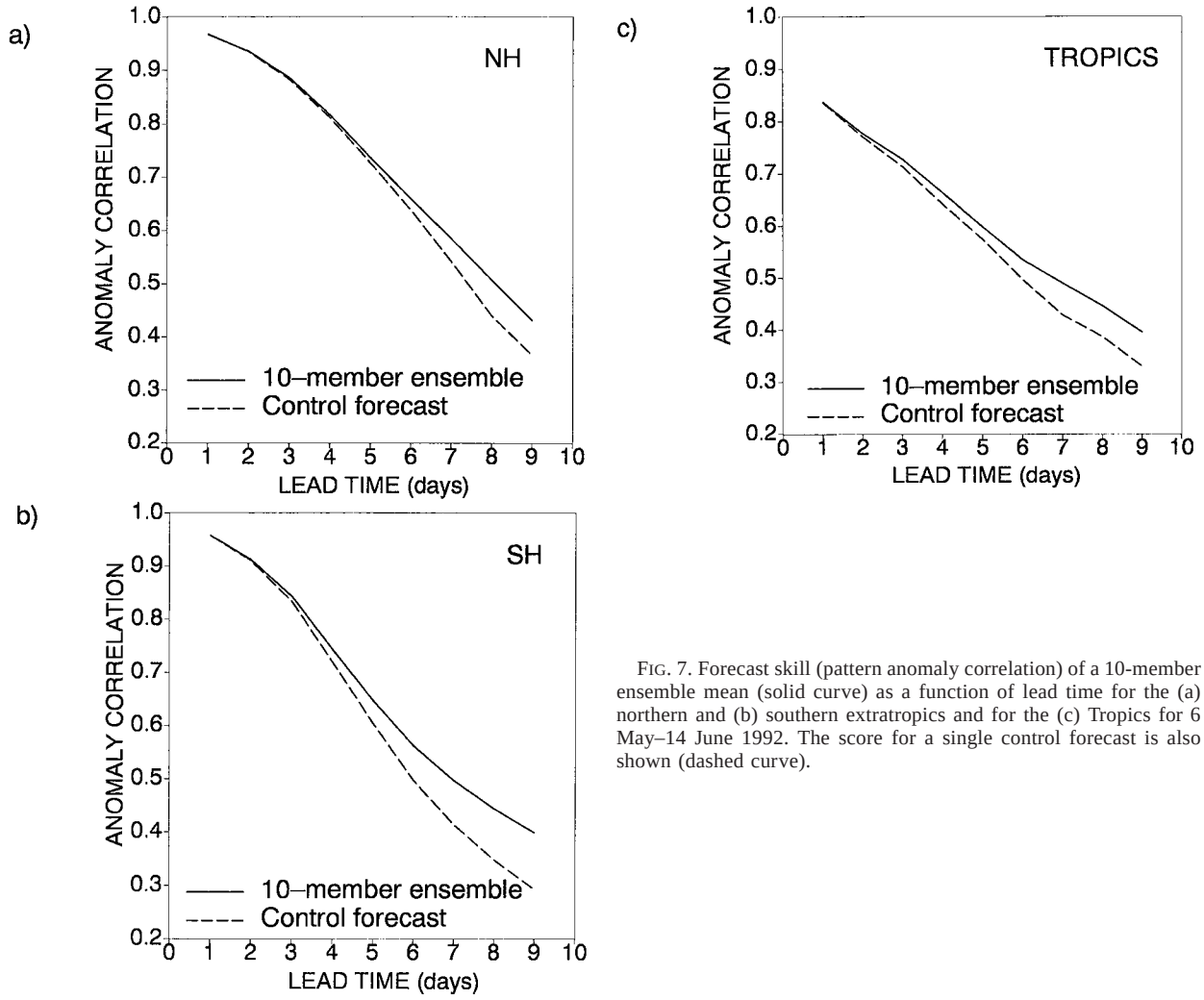

FIG. 7. Forecast skill (pattern anomaly correlation) of a 10-member ensemble mean (solid curve) as a function of lead time for the (a) northern and (b) southern extratropics and for the (c) Tropics for 6 May–14 June 1992. The score for a single control forecast is also shown (dashed curve).

is very effective in filtering out part of the forecast error that is due to initial error uncertainty. In contrast, the circulation in the summer is more ''local'' in nature, both in space and time. This is also reflected in the fact that the summer circulation has more spatial degrees of freedom (see, e.g., Fraedrich et al. 1995). Beyond large-scale dynamics, it is also strongly influenced by convection, which has a shorter lifetime and smaller saturation amplitude. It follows that a larger portion of the total error is left unexplained by baroclinic instabilities. As a consequence, our ensembles based primarily on baroclinic instabilities cannot provide as much improvement in skill in the summer as they can in the winter. Similar results were obtained by Molteni et al. (1996) using the ECMWF operational ensemble prediction system.

As can be seen from Fig. 7, the skill for the control and ensemble mean at day 1 are practically identical when verified against the control analysis (see also footnote 6). However, as expected from Leith (1974) and from section 2, the ensemble mean develops an advantage over the control forecast that becomes appreciable by day 3 and reaches a substantial 0.07–0.11 by day 9. If we consider 0.5 PAC as the minimum level of useful skill, ensemble forecasting extends predictability by a day or so (17–25 h), out to 8 days over the NH and 7 days over the SH and the Tropics. Note that the improvements from ensemble averaging are as large in the Tropics as they are over the summer hemisphere extratropics.

The gain from ensemble forecasting in the medium and extended range compares favorably with the increase obtained by doubling the horizontal resolution: at day 5, the difference between the scores obtained using the NCEP operational T126 model and a nearly identical, ''parallel'' T62 system is slightly below 0.02, averaged over 32 months of operations. The gain obtained by ensemble averaging with 10 members over the 40-day experimental period is substantially larger, although both procedures take about the same computer time. We should point out that increasing the resolution of global NWP models has a clear benefit during the

first few days of a forecast (Tracton and Kalnay 1993). Running ensembles at a lower resolution, however, has a substantial advantage for the range beyond 5 days, where nonlinearities become important. We mention here that ensemble forecasting can also be beneficial for the shorter range, as long as the nonlinear aspects of the flow are relatively well modeled and analyzed (see Brooks et al. 1995).

d. Ensemble averaging versus spatial smoothing

It might be argued that the gain in skill from ensemble averaging may be dominated by smoothing resulting from averaging the different perturbed forecasts. In the framework of quasigeostrophic integrations, for example, Houtekamer and Derome (1995) found that optimally smoothing the control forecast can improve the forecast verification as much as some ensemble configurations.

Figure 8a shows the verifying analysis for a 9-day forecast started from 30 May 1992. A comparison of the control forecast (Fig. 8b) with the 10-member ensemble average forecast (Fig. 8c), and their corresponding errors (Figs. 8d and 8e, respectively) suggests that ensemble averaging does indeed have a smoothing effect. It is more appropriate to call this effect ''filtering,'' since it depends on the flow, particularly upon the varying degree of similarity amongst the ensemble members. Ensemble averaging results in a selective smoothing of those features that cannot be forecast with certainty. Consider, for example, the forecasts in Fig. 8 over North America. The trough over the southeastern United States is well predicted by the control and is hardly changed by the 10-member ensemble mean. The southern portion of the trough predicted by the control over the west coast, however, did not verify. The ensemble mean filtered out part of this system, resulting in smaller overall errors in this region. Undoubtedly, there are several other areas/cases where the changes in the ensemble mean do not verify well but overall it still provides an improvement over the single control forecast.

To quantify how much of the improvement due to ensemble averaging is connected to simple spatial smoothing (as compared to nonlinear filtering), we performed experiments where both the control and the ensemble mean forecasts were spatially smoothed with a filter that has a Gaussian response in the grid domain (J. Purser 1992, personal communication) till they reached their maximum PAC verification scores. The results, presented in Table 4, show that not much smoothing is needed to maximize the scores in the extratropics. Even at 9 days lead time, a truncation of T20 has to be retained in the control while, as expected, the ensemble average requires somewhat less smoothing. In the Tropics (not shown) no amount of smoothing improves the scores. The main result here is that the ensemble average retains a considerable advantage (more than 60%) over the control even after both fields had been optimally smoothed.

e. Forecast of the spatial distribution of the errors

Ensemble forecasting should offer more than an improved best estimate of the evolution of the atmosphere (ensemble mean forecast). It should also provide the means to estimate higher moments, and ultimately the full probability distribution of the forecasts. A first step in achieving this goal is the derivation of an estimate of forecast reliability in the spatial domain. Ideally, we would like to know in which areas errors are more likely. We have used the spatially smoothed ensemble spread of the kinetic energy introduced in section 6a for estimating the magnitude of the expected forecast errors. Figures 8f and 8g show, for the same 9-day forecast example of the previous subsection, the spatial distribution of the kinetic energy of the error and of the ensemble spread, respectively. Several important aspects of the error field are indicated quite realistically in the ensemble spread field. Note, for example, that the absolute maxima in the error field over the two extratropics is well predicted by the ensemble over southern Australia and over eastern Asia. Several error features turn out to be well predicted in the subtropics and Tropics as well—see, for example, the correspondence between the actual and predicted large errors over Western Sahara and east of the Hawaiian Islands.

The spread/error PAC scores based on the ensemble forecasts are displayed in Fig. 9. The fact that the spread/error PAC is low at short lead times is due to the presence of random errors in the initial conditions and verifying analyses (see also Barker 1991; Wobus and Kalnay 1995). Since there is a strong zonally symmetric component in the error fields, we computed the PAC of the spread/error both with (not shown) and without the zonal mean included. The spread/skill spatial correlation is about 0.4 without the zonal mean and is above 0.7 with the zonal mean included. This result is encouraging, suggesting that ensemble forecasting can result in skillful predictions of the spatial distribution of the errors.

f. Forecast of the temporal variations in skill

The ensemble forecasts can also be used to predict the variations of forecast skill (or the reliability of forecasts) in the time domain. This has been a subject of considerable research because of its importance for medium- and extended-range forecasts (e.g., Branstator 1986; Kalnay and Dalcher 1987; Palmer and Tibaldi 1988). If we can determine a priori which forecasts are going to be most skillful, the utility of extended-range forecasts can be considerably enhanced (e.g., Tracton et al. 1989). Here we will demonstrate the relationship in time between ensemble spread and error through a typical example.

In Fig. 10 the PAC scores between 9-day lead ensemble mean forecasts and the verifying analysis are

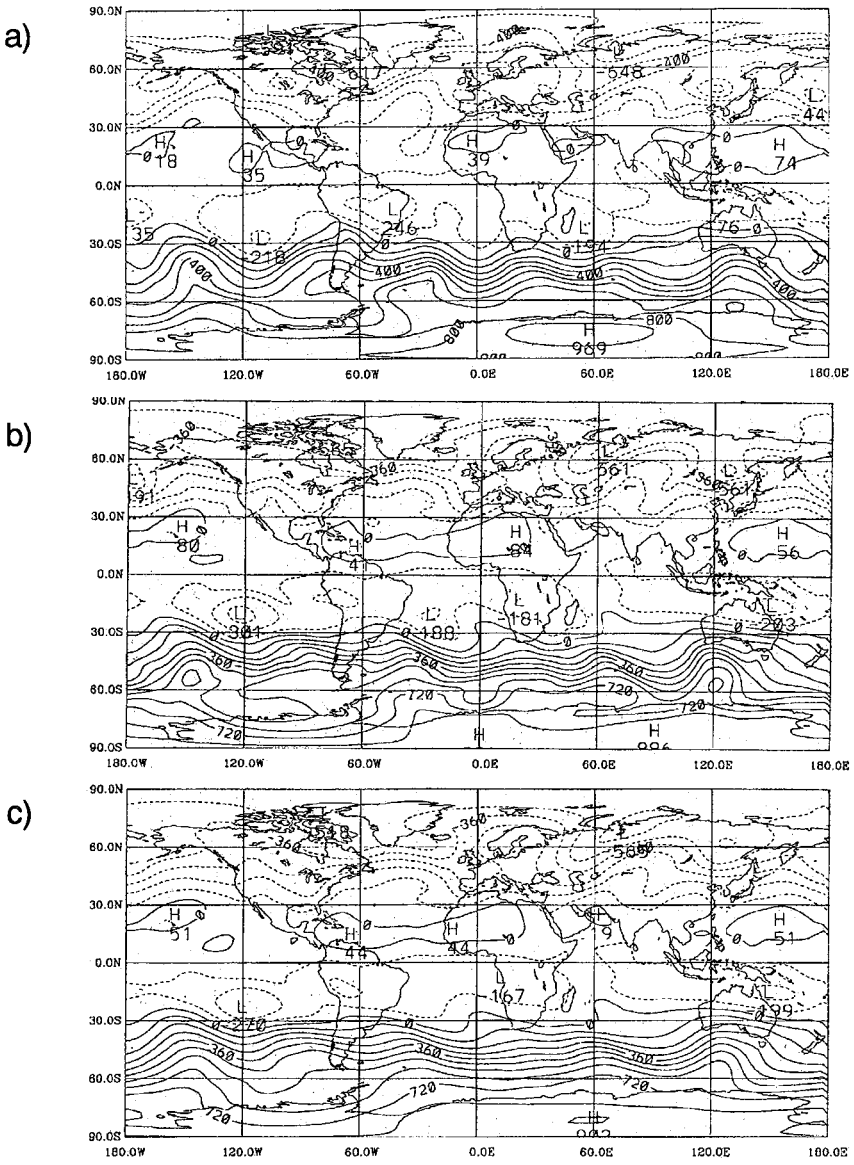

FIG. 8. Ten-member, 9-day lead time ensemble forecast started on 30 May 1992. Shown are the 500-hPa streamfunction fields for (a) verifying analysis (contour interval: CI = 10^7); (b) control forecast (CI = 9×10^6); (c) ensemble mean forecast (CI = 9×10^6); (d) control error (CI = 3×10^6); (e) ensemble mean error (CI = 3×10^6); (f) as in (d) but in rms and smoothed; (g) forecast of the error (f) as in (e) but in rms and smoothed. Labels for (a)–(e) are scaled by 10^{-5}.

plotted along with, as an indicator of skill, the average of the PAC values between the control forecast and each ensemble member for a 40-day experimental period started 6 May 1992. During the first 30 days of the period the two curves have a correlation of 0.62—a value similar to those reported by Barker (1991) with a perfect model approach, by Wobus and Kalnay (1995) using a statistical approach based partly on an ensemble of control forecasts from different NWP centers, and by Molteni et al. (1996), using the operational ECMWF ensemble. Spread-skill temporal correlations from the operational NCEP ensemble, above 0.6 beyond day 5 lead time, confirm the above experimental results (J. Whitaker 1996, personal communication). We should also note that the spread PAC values are somewhat higher than the skill PACs, an indication that the spread in the ensemble is somewhat deficient.

A drastic change occurs in the behavior of the two curves in Fig. 10, however, around the 30th day. Though there is still a relationship between ensemble spread and skill during the last 10 days of the period (with an anomaly correlation of 0.47), the spread PACs now are much higher than the skill PAC values. This may be associated with the establishment of a summer circulation regime

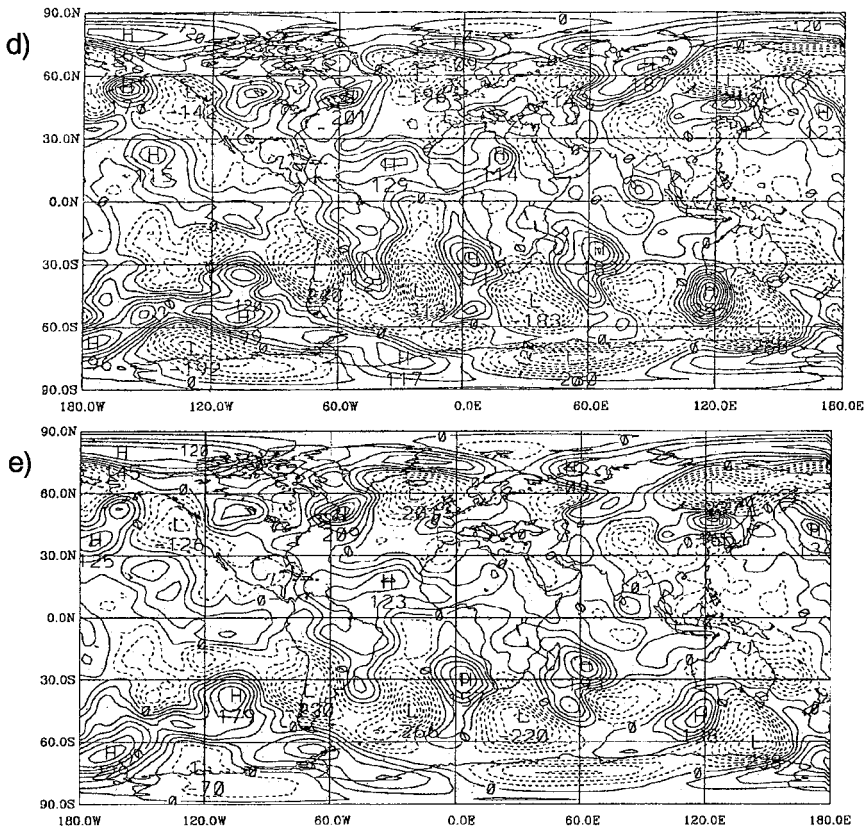

FIG. 8. (*Continued*)

in early June 1992. As it is well known, the MRF (and global GCMs in general) have more systematic errors in the summer that can introduce additional errors not accounted for by the ensemble. Notwithstanding the problem posed by systematic model errors that can ruin the spread/skill relationship, an objective evaluation of the NCEP ensemble during the 1995/96 winter season strongly indicates that the system can effectively distinguish between cases with high and low uncertainty (Zhu et al. 1996).

g. Size of the ensemble

It was Leith (1974) who first considered the question of how many ensemble members are needed to improve the skill of the control forecast by ensemble averaging. Using a simple model he found that eight members are enough to realize most of the gain attainable through ensemble averaging. Houtekamer and Derome (1995), also using a perfect model environment but with a three-layer, T21 resolution quasigeostrophic model, basically confirmed Leith's results. Barker (1991), using a setup similar to that of Houtekamer and Derome, examined the effect of ensemble size on the temporal correlation between ensemble spread and control skill. We now consider the same question using a setup equivalent to the operational NCEP ensemble system.

In Fig. 11, the skill of the ensemble mean, the skill in forecasting the spatial error pattern, and the temporal correlation between ensemble spread and control error are displayed as a function of ensemble size between 1 and 40 members. The results are shown for 9 days lead time, based on a 40-day experimental period during which a total of 40 forecasts from 20 independently run breeding cycles were generated. The gain from enlarging the ensemble is most obvious when going from 2 to 4 and then to 10-member ensembles, a result in agreement with earlier studies. Regarding forecast skill, only minimal improvement is obtained beyond 20 members. However, as the steeper curves in Fig. 11 indicate, the temporal and spatial relationship between spread and error continues to improve even up to 40 members. From the shape of these curves it seems there is still a lot to be gained from increasing the size of the ensemble beyond 40 members. Certainly it is clear from the figures that for higher forecast moments it is necessary to have many more members in order to reduce the sampling problem.[7]

[7] Tracton (1993, personal communication) also pointed out that ensemble-based reliability estimates can be improved upon by including one or even two days old lagged ensemble members, even if the lagged ensemble members do not improve the ensemble mean appreciably.

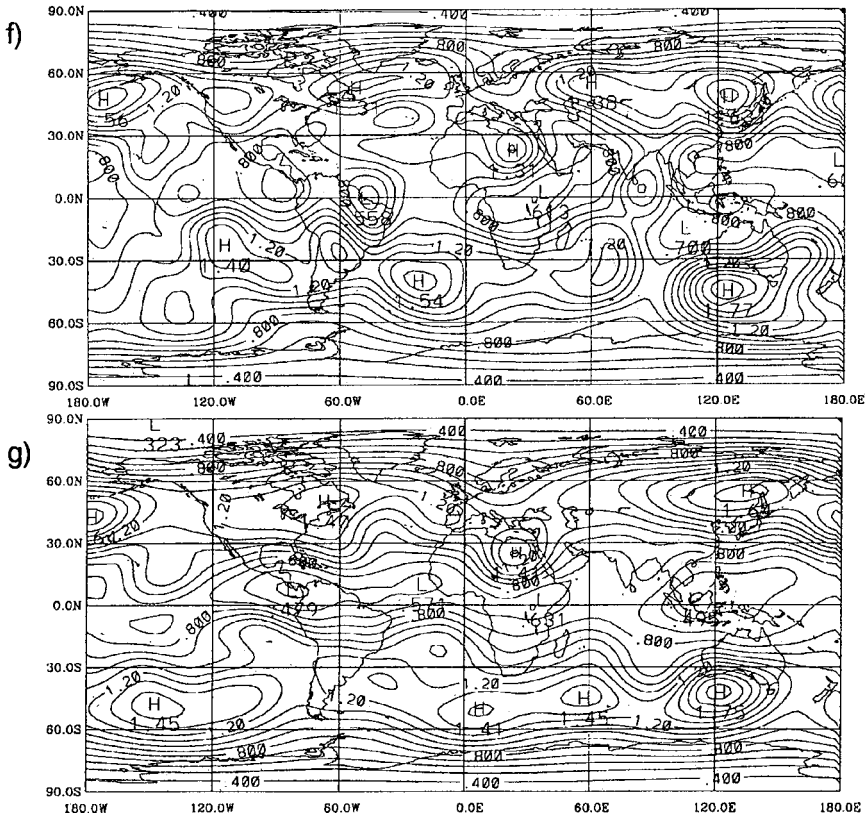

FIG. 8. (*Continued*)

h. Comparison of bred versus random initial perturbations

Finally, we compare the effectiveness of random and bred initial perturbations. The random perturbations are created by linearly combining difference fields between randomly selected analyses with random weights. These random perturbations are, by construction, not dependent on the flow of the day and hence expected to perform poorer than the bred vectors. Toth and Kalnay (1993) showed that two-member bred ensembles outperformed similarly sized ensembles with random initial perturbations in terms of ensemble mean scores. As indicated in Table 5, this is also true for 10-member ensembles. The advantage of bred perturbations is more pronounced over the winter hemisphere, where baroclinic instabilities have possibly a greater contribution to initial errors. The absolute differences in the performance of the two ensembles are not large but when expressed in terms of the difference in skill between the control forecast and the random ensemble, or between high and low resolution controls, they can reach values up to 40%. Since random perturbations initially grow

TABLE 4. The effect of optimal spatial smoothing on the control and 10-member ensemble mean forecasts for the period 23 May–3 June 1992 with 10%/20% initial perturbations for the Northern and Southern Hemispheres, respectively. For further details, see text.

Lead time (days)	Optimal smoothing (~triangular truncation)		Ensemble advantage over control retained	
	Control	Ensemble	PAC	Percent total
5	T30	T40	0.02	62.5%
7	T25	T35	0.033	63.8%
9	T20	T30	0.042	60.5%

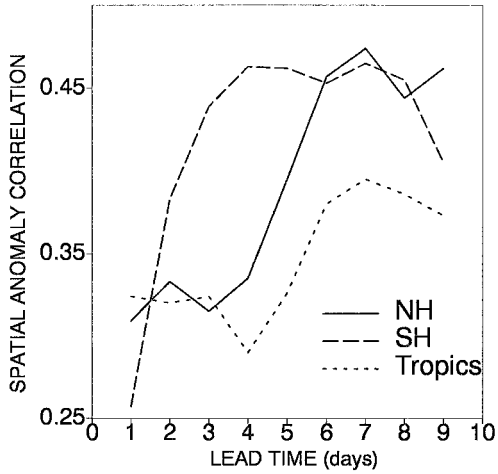

FIG. 9. Pattern correlation between predicted and actual error in the forecasts, averaged for the period 6 May–14 June 1992.

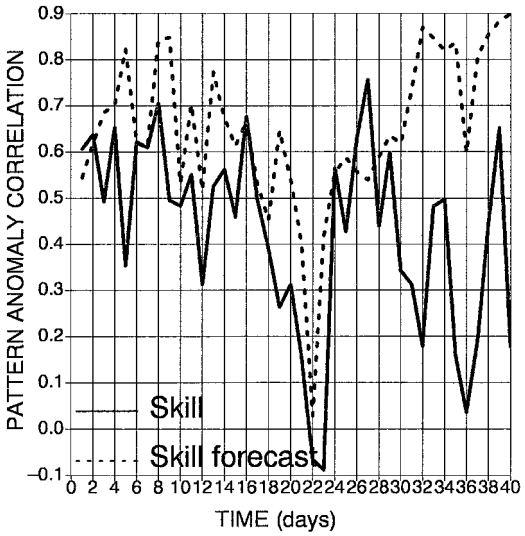

FIG. 10. PAC skill values for 9-day lead time forecasts for the North American region (solid line) along with average PAC between the control forecast and each of the 10 ensemble members (dashed line) for 6 May–14 June 1996.

slower than the bred ones, for a fair comparison the random ensemble was started with a larger, 18.5%/37% initial perturbation amplitude to ensure that over the medium- and extended-ranges the two ensembles have comparable perturbation amplitudes.

7. Operational implementation

The initial operational ensemble configuration implemented at NCEP in December 1992 consisted of one pair of bred perturbed forecasts, one T126 and a T62 control forecast, plus a 12-h delayed control forecast (Tracton and Kalnay 1993). All forecasts initiated in the most recent 48 h were included, making an ensemble of 14 valid for 10 days. Based on the experimental results presented in the previous sections, and following the installation at NCEP of a new Cray C90 supercomputer, the ensemble forecasting system was upgraded on 30 March 1994. In addition to the T62 and T126 control forecasts, five bred pairs of forecasts are run at 0000 UTC and two pairs at 1200 UTC, and all the forecasts are extended to 16 days. The new configuration amounts

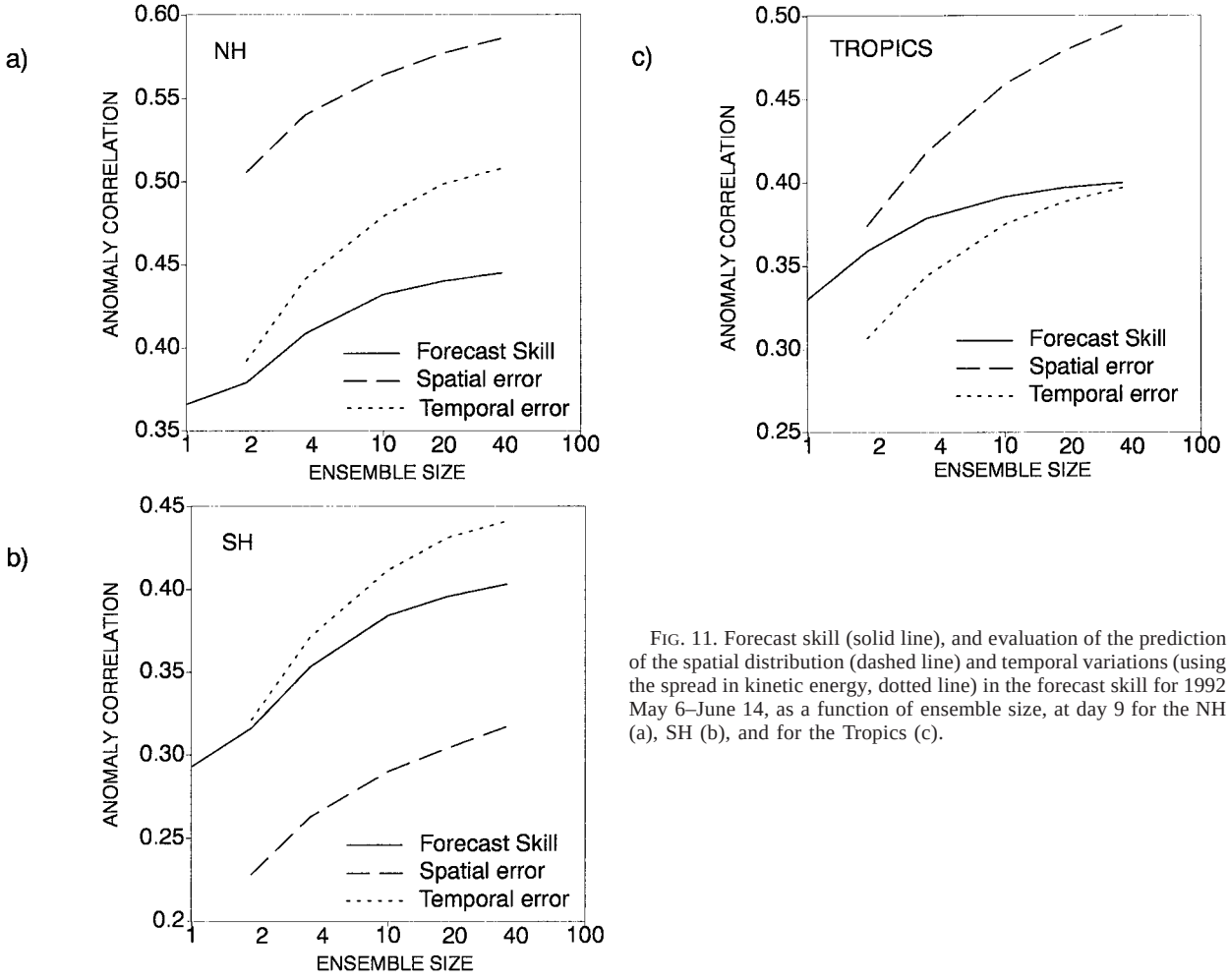

FIG. 11. Forecast skill (solid line), and evaluation of the prediction of the spatial distribution (dashed line) and temporal variations (using the spread in kinetic energy, dotted line) in the forecast skill for 1992 May 6–June 14, as a function of ensemble size, at day 9 for the NH (a), SH (b), and for the Tropics (c).

TABLE 5. Comparison of the control forecasts (CNT) and 10-member, randomly generated (18.5%/37% rms variance initial amplitude) and bred ensembles (12.5%/25% rms variance initial amplitude) for 23 May–6 June 1992. PAC skill score for T126 control is estimated based on average difference between high- and low-resolution controls for 3 years.

	Forecast skill (PAC)					
	Day 5			Day 9		
	NH	Tropics	SH	NH	Tropics	SH
Forecast						
CNT T62	0.735	0.503	0.583	0.395	0.264	0.313
CNT T126	0.753	—	0.601	—	—	—
Random	0.747	0.516	0.624	0.474	0.282	0.374
Bred	0.752	0.546	0.635	0.472	0.302	0.403

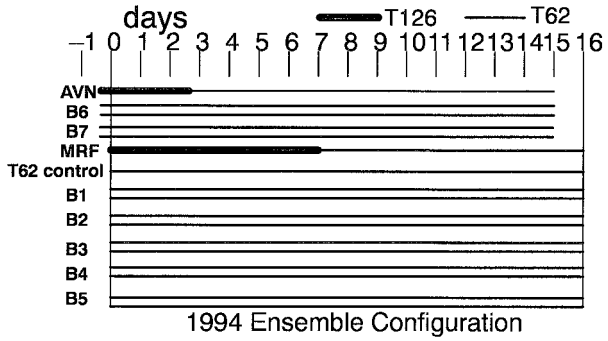

FIG. 12. Schematic of the configuration of the operational ensemble forecasting system at NMC. Each horizontal line represents a numerical forecast. High-resolution T126 forecasts are marked with heavy lines while the other forecasts are run at T62 resolution. Note that at 0000 UTC there are two control forecasts: one started at T126 resolution and then truncated to T62 at day 7 and one started at a T62 truncated resolution. At 1200 UTC, the high-resolution control is truncated after 3 days of integration. Pairs of perturbed forecasts based on the breeding method are marked as B1–B7.

to 17 individual ensemble members every day. When the forecasts from the last two days are also considered for the extended range, the total number of ensemble members valid for two weeks is 46 (see Fig. 12 and also Tracton 1994).

Based on the results of section 6b, the size of the initial perturbations is set at 12.5% and 25% of the total rms variance in the NH and SH, respectively. (During SH summer, the perturbation size there is reduced to 12.5% rms variance.) In the regional rescaling procedure, the kinetic energy of the flow (rather than the previously used rms streamfunction norm) is applied. We use 24-h breeding cycles, and the bred perturbations are determined as the difference between two perturbed ensemble forecasts at 24-h lead time. It should be noted that with this procedure the generation of bred perturbations is performed at no cost beyond that of running the ensemble forecasts themselves (see Fig. 5). In this configuration, breeding is part of the extended ensemble forecasts and the creation of efficient initial ensemble perturbations requires no additional computing resources beyond that needed to run the forecasts themselves. Both subjective (Toth et al. 1997) and objective (Zhu et al. 1996) evaluation of the results from the new operational ensemble system support the experimental results reported in this paper.

8. Conclusions

In this paper different aspects of ensemble forecasting were examined. First, it was emphasized that the perturbations applied to the control initial state of the atmosphere (analysis) must be chosen in the directions of possible growing errors in the analysis. Using Lorenz's error growth equation we showed that if the perturbations project on the analysis errors, averaging pairs of perturbed forecasts results in a nonlinear filtering of forecast errors. On the other hand, if the initial perturbations do not project on the errors in the analysis, the same nonlinear processes can lessen the positive impact of ensemble averaging. We argued that the analysis errors are composed not only by random errors as assumed in the operational analysis methods, but also by fast growing ''errors of the day'' introduced by the successive use of dynamical short-range forecasts as first-guess fields within the analysis cycle.

The growing errors in the analysis cycle develop as perturbations upon the evolving true state of the atmosphere. The perturbations (i.e., the analysis errors), carried forward in the first guess forecasts, are reduced (or ''scaled down'') at regular intervals by the use of observations. However, because of the inhomogeneous distribution of observations some errors in the analysis will grow without suppression by observational data. Examples of this can be found over the oceans and the SH, where radiosonde data is scarce, and on small scales anywhere. Due to this process, growing errors associated with the changing state of the atmosphere develop within the analysis cycle and dominate subsequent forecast error growth.

We argued that these errors or perturbations can be well estimated by the method of ''breeding growing vectors,'' which simulates the development of growing errors in the analysis cycle. In a breeding cycle, the difference field between two nonlinear forecasts is carried forward (and scaled down at regular intervals) upon the evolving atmospheric analysis fields. The bred vectors, in fact, offer a generalization of the linear Lyapunov vectors into the nonlinear perturbation regime. In complex systems like the atmosphere where perturbation growth may critically depend on perturbation amplitude, the bred vectors interpreted at a given amplitude offer the nonlinear equivalent of the Lyapunov vectors. Thus, we surmise that due to their construction, the bred vectors are superpositions of the leading local (time-dependent) finite amplitude perturbation vectors that can grow fastest in a given perturbation range (i.e., nonlinear equivalent of the local Lyapunov vectors, LLVs). Breeding cycles with different initial perturbations converge,

in any geographical region after a few days, to a subspace of perturbations that comprises the leading local (in phase space) fastest growing finite-amplitude perturbations of the atmosphere. The unique role played by the leading LLVs in analysis/forecasting was emphasized by pointing out that all linear perturbations, after a transient period of 3–4 days, assume the shape of leading LLVs. In any bred perturbation the weight on the individual leading LLVs is random, determined by how the preceeding perturbation projects on the different LLVs at that time and also on the impact of smaller-scale stochastic forcing (convection) on the dominant baroclinic processes. It was shown that perturbations from independent breeding cycles are quasi-orthogonal in a global sense without the introduction of any constraint. For these reasons the bred perturbations lend themselves as good candidates to be used as initial ensemble perturbations.

The growing component of the regionally varying uncertainty in the analysis was measured as the difference between parallel analysis cycles. The average difference field is then used as a mask in the regular rescaling process of the bred vectors to ensure that the initial ensemble perturbations have a spatial distribution of amplitudes similar to that of the analysis errors. Each bred perturbation is both added to and subtracted from the control analysis.

Results from 10-member experimental ensembles indicate that for short-range forecasts the optimal size of the initial perturbations is about the same as the estimated size of analysis errors. For longer forecasts, the optimal size is somewhat larger, presumably because of the presence of model deficiencies that generate additional forecast errors that can be, to some extent, treated just as the initial uncertainty.

We showed that for medium-range forecasting the mean of the bred ensemble has skill superior to that of 1) a double horizontal resolution control, 2) a control smoothed optimally, and 3) an ensemble initiated with random (flow independent) perturbations. We also pointed out that ensemble averaging removes the unpredictable components of the flow while leaving the predictable part virtually intact. These results attest that the bred ensemble mean offers an economic way for improving the control forecast and thus can replace the control as our best estimate of the future state of the atmosphere.

Higher moments of the probability distribution of future states should also be estimated through ensemble forecasting. For limited samples, we showed that bred ensemble spread correlates with forecast error both in space and in time. This information about the reliability of the forecasts is especially critical at longer lead times where model performance is known to be case dependent.

The improvement in forecast skill at and beyond 7 days lead times (0.04–0.11 in AC), together with a robust estimate of forecast reliability ($\sim$0.6 temporal correlation between ensemble spread and forecast errors) indicates that the bred ensemble system has the potential of extending ''weather outlooks'' into the second forecast week. To capitalize on this potential, NCEP started on 30 March 1994 an ensemble forecasting system with 14 perturbed (bred) and 3 control forecasts, each extending out to 16 days in lead time. When all forecasts initiated within the past 48 h are considered, there is a 46-member ensemble valid for two weeks available every day.

There has been much discussion recently about the properties and relative merits of bred vectors and optimal (or singular) vectors. Despite the many differences, we would like to emphasize here that both types of perturbations represent a subspace of possible growing perturbations. The adjoint technique used to compute the optimal vectors is capable of finding the fastest growing linear perturbations. When using this linear methodology, care should be taken so that the likelihood of different analysis errors is adequately described. This problem is currently under inverstigation (Barkmeijer et al. 1997, manuscript submitted to *Quart. J. Roy. Meteor. Soc.*). Until this problem is solved, a simple approach, proposed by Houtekamer (1995) could be used to obtain singular vectors that are statistically representative of analysis errors in terms of baroclinic shear and spatial error magnitude. Work is also in progress at different centers to include more physical processes in the tangent linear and adjoint formulations of NWP models (e.g., Zupanski and Mesinger 1995), which will increase the utility of the adjoint technique.

If practical aspects such as simplicity and computational costs are considered, the breeding method has a clear advantage over competing methodologies. The bred vectors are computed using the full physics package of NWP models at the highest required horizontal resolution. Without the introduction of any special constraint, they are well balanced and correspond well with the estimated structure of analysis errors (Houtekamer and Derome 1995; Iyengar et al. 1996). However, the bred vectors are not optimized for future growth. This may be a disadvantage of the breeding method. However, it is possible that on the time and spatial scales considered here the structure of the optimal vectors, given the likelihood of analysis errors, are similar to those of the bred vectors. This would be true if nonmodal behavior would not be dominant, given again the likelihood of different analysis errors. Recent results by Fisher and Coutier (1995) and Barkmeijer et al. (1997, manuscript submitted to *Quart. J. Roy. Meteor. Soc.*) point into this direction; nevertheless, more research is needed to further clarify these issues.

At ECMWF the adjoint perturbation technique is used. A similar technique was adopted for the 30-day range at the Japan Meteorological Agency more recently, while research is under way regarding the configuration of a medium-range ensemble (Takano 1996, personal communication). At FNMOC and at the South

African Weather Bureau a breeding-based ensemble is operational (Rennick 1995; W. Tennant 1995, personal communication), while at the National Center for Medium Range Weather Forecasting in India a breeding-based ensemble system is being prepared to be implemented (G. Iyengar 1996, personal communication). At the Atmospheric Environment Service of Canada experiments have been carried out with an ensemble system in which, beyond the initial atmospheric conditions, the initial surface parameters, as well as some model parameters are also perturbed (Houtekamer et al. 1996). The perturbed atmospheric initial conditions are derived from running independent analysis cycles, in each of which randomly generated ''measurement errors'' are added to the real observational data. The independent analysis cycles can be considered as breeding cycles, where, beyond the growing perturbations, random analysis errors are also well represented in a statistical sense. In addition to the above sites, ensemble forecast experiments are under way at several other places, including NCAR (Baumhefner 1996).

The different perturbation techniques have various potential advantages. Their impact on the quality of ensemble forecasts can be evaluated only after a careful comparison of experimental results. We conclude by noting that a combination of ensemble forecasts from different numerical prediction centers may give further improvement to the quality of an ensemble (Harrison et al. 1995). The benefits from having a larger number of forecasts, and using different analysis schemes, forecast models, and perturbation techniques may all contribute to the success of numerical weather prediction.

Acknowledgments. We greatly appreciate the many discussions and help offered by M. Kanamitsu, H. Juang, M. Iredell, S. Saha, J. Purser, J. Derber, S. Tracton, J. Irwin, and P. Caplan at different stages of this work. Jim Purser kindly provided us with the spherical filter used in sections 5 and 6 while Steve Tracton suggested the use of the schematic in Fig. 12. The fact that the breeding method has a strong connection to the Lyapunov vectors was first pointed out to us by Jon Ahlquist and Anna Trevisan. Joe Gerrity, Peter Houtekamer, Tim Palmer, Istvan Szunyogh, and Steve Tracton offered valuable comments on earlier versions of this manuscript.

REFERENCES

Augustine, S. J., S. L. Mullen, and D. P. Baumhefner, 1992: Examination of actual analysis differences for use in Monte Carlo forecasting. *Proc. 16th Climate Diagnostics Workshop*, Los Angeles, CA, NOAA/NWS/NMC/CPC, 375–378.

Barker, T. W., 1991: The relationship between spread and forecast error in extended-range forecasts. *J. Climate,* **4,** 733–742.

Baumhefner, D. P., 1996: Numerical extended-range prediction: Forecast skill using a low-resolution climate model. *Mon. Wea. Rev.*, **124,** 1965–1979.

Benettin, G., L. Galgani, and J. M. Strelcyn, 1976: Kolmogorov entropy and numerical experiments. *Phys. Rev. A,* **14,** 2338–2345.

——, ——, A. Giorgilli, and J. M. Strelcyn, 1980: Lyapunov characteristic exponents for smooth dynamical systems and for Hamiltonian systems: A method for computing all of them. *Meccanica,* **15,** 9.

Bishop, C., and Z. Toth, 1996: Using ensembles to identify observations likely to improve forecasts. Preprints, *11th Conf. on Numerical Weather Prediction,* Norfolk, VA, Amer. Meteor. Soc., 72–74.

Bouttier, F., 1994: A dynamical estimation of forecast error covariances in an assimilation system. *Mon. Wea. Rev.*, **122,** 2376–2390.

Branstator, G., 1986: The variability in skill of 72-hour global-scale NMC forecasts. *Mon. Wea. Rev.*, **114,** 2628–2639.

Brooks, H. E., M. S. Tracton, D. J. Stensrud, G. DiMego, and Z. Toth, 1995: Short-range ensemble forecasting (SREF): Report from a workshop. *Bull. Amer. Meteor. Soc.*, **76,** 1617–1624.

Buizza, R., 1994: Localization of optimal perturbations using a projection operator. *Quart. J. Roy. Meteor. Soc.*, **120,** 1647–1681.

——, 1995: Optimal perturbation time evolution and sensitivity of ensemble prediction to perturbation amplitude. *Quart. J. Roy. Meteor. Soc.*, **121,** 1705–1738.

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

Daley, R., 1991: *Atmospheric Data Analysis.* Cambridge University Press, 471 pp.

——, and T. Mayer, 1986: Estimate of global analysis error from the Global Weather Experiment Observational Network. *Mon. Wea. Rev.*, **114,** 1642–1653.

Ebisuzaki, W., and E. Kalnay, 1991: Ensemble experiments with a new lagged analysis forecasting scheme. WMO, Research Activities in Atmospheric and Oceanic Modelling Rep. 15, 6.31–6.32. [Available from WMO, C. P. No. 2300, CH1211, Geneva, Switzerland.]

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

——, and J. J. Tribbia, 1997: Optimal prediction of forecast error covariances through singular vectors. *J. Atmos. Sci.*, **54,** 286–313.

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

Fraedrich, K., Ch. Ziehmann, and F. Sielmann, 1995: Estimates of spatial degrees of freedom. *J. Climate,* **8,** 361–369.

Gandin, L. S., 1965: *Objective Analysis of Meteorological Fields.* Israel Program for Scientific Translations, 242 pp.

Harrison, M. S. J., D. S. Richardson, K. Robertson, and A. Woodcock, 1995: Medium–range ensembles using both the ECMWF T63 and unified models—An initial report. U.K. Meteorological Office Tech. Rep. 153, 25 pp. [Available from Forecasting Research Division, Meteorological Office, London Road, Bracknell, Berkshire RG12 2SZ, United Kingdom.]

Houtekamer, P. L., 1995: The construction of optimal perturbations. *Mon. Wea. Rev.*, **123,** 2888–2898.

——, and J. Derome, 1994: Prediction experiments with two-member ensembles. *Mon. Wea. Rev.*, **122,** 2179–2191.

——, and ——, 1995: Methods for ensemble prediction. *Mon. Wea. Rev.*, **123,** 2181–2196.

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

Iyengar, G., Z. Toth, E. Kalnay, and J. Woollen, 1996: Are the bred vectors representative of analysis errors? Preprints, *11th Conf. on Numerical Weather Prediction,* Norfolk, VA, Amer. Meteor. Soc., J64–J66.

Kalnay, E., and A. Dalcher, 1987: Forecasting forecast skill. *Mon. Wea. Rev.*, **115,** 349–356.

——, and Z. Toth, 1994: Removing growing errors in the analysis.

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.
Kanamitsu, M., and Coauthors, 1991: Recent changes in the global forecast system at NMC. *Wea. Forecasting,* **6,** 425–435.
Langland, R. H., and G. D. Rohaly, 1996: Analysis error and adjoint sensitivity in prediction of a North Atlantic frontal cyclone. Preprints, *11th Conf. on Numerical Weather Prediction,* Norfolk, VA, Amer. Meteor. Soc., 150–152.
Leith, C. E., 1974: Theoretical skill of Monte Carlo forecasts. *Mon. Wea. Rev.,* **102,** 409–418.
Lorenc, A., 1981: A global three-dimensional multivariate statistical interpolation scheme. *Mon. Wea. Rev.,* **109,** 701–721.
Lorenz, E. N., 1965: A study of the predictability of a 28-variable atmospheric model. *Tellus,* **17,** 321–333.
——, 1982: Atmospheric predictability experiments with a large numerical model. *Tellus,* **34,** 505–513.
Molteni, F., and T. N. Palmer, 1993: Predictability and finite-time instability of the northern winter circulation. *Quart. J. Roy. Meteor. Soc.,* **119,** 269–298.
——, 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.
Palmer, T. N., and S. Tibaldi, 1988: On the prediction of forecast skill. *Mon. Wea. Rev.,* **116,** 2453–2480.
——, F. Molteni, R. Mureau, R. Buizza, P. Chapelet, and J. Tribbia, 1992: Ensemble prediction. ECMWF Research Department Tech. Memo. 188, 45 pp.
Parrish, D. F., and J. C. Derber, 1992: The National Meteorological Center's spectral statistical-interpolation analysis system. *Mon. Wea. Rev.,* **120,** 1747–1763.
Rennick, M. A., 1995: The ensemble forecast system (EFS). Models Department Tech. Note 2-95, Fleet Numerical Meteorology and Oceanography Center, 19 pp. [Available from Models Department, FLENUMMETOCCEN, 7 Grace Hopper Ave., Monterey, CA 93943.]
Reynolds, C. A., P. J. Webster, and E. Kalnay, 1994: Random error growth in NMC's global forecasts. *Mon. Wea. Rev.,* **122,** 1281–1305.
Shimada, I., and T. Nagashima, 1979: A numerical approach to ergodic problem of dissipative dynamical systems. *Prog. Theor. Phys.,* **61,** 1605–1615.
Smagorinsky, J., 1969: Problems and promises of deterministic extended range forecasting. *Bull. Amer. Meteor. Soc.,* **50,** 286–311.
Szunyogh, I., E. Kalnay, and Z. Toth, 1997: A comparison of Lyapunov and optimal vectors in a low resolution GCM. *Tellus,* **49A,** 200–227.
Toth, Z., and Kalnay, E., 1993: Ensemble forecasting at NMC: The generation of perturbations. *Bull. Amer. Meteor. Soc.,* **74,** 2317–2330.
——, I. Szunyogh, and E. Kalnay, 1996: Singular, Lyapunov and bred vectors in ensemble forecasting. Preprints, *11th Conf. on Numerical Weather Prediction,* Norfolk, VA, Amer. Meteor. Soc., 53–55.
——, E. Kalnay, S. Tracton, R. Wobus, and J. Irwin, 1997: A synoptic evaluation of the NCEP ensemble. *Wea. Forecasting,* **12,** 140–153.
Tracton, M. S., 1994: Operational ensemble prediction—The NMC experience. Preprints, *Tenth Conf. on Numerical Weather Prediction,* Portland, OR, Amer. Meteor. Soc., 206–208.
——, and E. Kalnay, 1993: Operational ensemble prediction at the National Meteorological Center: Practical aspects. *Wea. Forecasting,* **8,** 379–398.
——, K. Mo, W. Chen, E. Kalnay, R. Kistler, and G. White, 1989: Dynamical extended range forecasting (DERF) at the National Meteorological Center. *Mon. Wea. Rev.,* **117,** 1604–1635.
Trevisan, A., and R. Legnani, 1995: Transient error growth and local predictability: A study in the Lorenz system. *Tellus,* **47A,** 103–117.
Wobus, R. L, and E. Kalnay, 1995: Three years of operational prediction of forecast skill at NMC. *Mon. Wea. Rev.,* **123,** 2132–2148.
Wolf, A., J. B. Swift, H. L. Swinney, and J. A. Vastano, 1985: Determining Lyapunov exponents from a time series. *Physica,* **16D,** 285–317.
Zhu, Y., G. Iyengar, Z. Toth, M. S. Tracton, and T. Marchok, 1996: Objective evaluation of the NCEP global ensemble forecasting system. Preprints, *15th Conf. on Weather Analysis and Forecasting,* Norfolk, VA, Amer. Meteor. Soc., J79–J82.
Zupanski, D., and F. Mesinger, 1995: Four-dimensional variational assimilation of precipitation data. *Mon. Wea. Rev.,* **123,** 1112–1127.