ENSEMBLE FORECASTING AT NCEP

Zoltan Toth[1] and Eugenia Kalnay

Environmental Modeling Center
National Centers for Environmental Prediction
NOAA/National Weather Service
Washington DC, USA

Summary: One of the most important aspects in ensemble forecasting is the creation of initial perturbations. Several different approaches have been proposed to solve this problem but there is not yet a consensus regarding an optimal strategy. In this paper we describe the path taken at NCEP, along with some comparison with other methods.

Creating initial ensemble perturbations envolves two steps: (1) the estimation of the uncertainty in the control analysis in a probabilistic fashion and (2) the sampling of that distribution. Regarding sampling one has to consider that in the atmosphere there are only a small number O(10) of fast growing independent perturbations in any region while there are $O(10^6)$ nominal degrees of freedom in our models. And since it is the fast growing errors that will impact the forecasts most, with a limited number of ensemble forecasts we should attempt to sample the subspace of the most unstable directions while ignoring the neutral or decaying part of the phase space of possible analysis errors, since they cannot be sampled well with a small number of ensemble forecasts anyway.

For estimating the fast growing component of the analysis errors, we use the breeding method (see Kalnay and Toth, 1996, same volume). By applying several independent breeding cycles we sample the subspace of the fastest growing possible analysis errors. We note here that the singular vector based method (when applied with commonly used norms) is not an estimation but rather a sampling method and that without good estimation of the initial error it may lead to suboptimal sampling.

Ensemble forecasting has been operational at NCEP (formerly NMC) since December 1992. In March 1994, more ensemble forecast members were added. In the new configuration, 17 forecasts with the NCEP global model are run every day, out to 16 days lead time. Beyond the 3 control forecasts (a T126 and a T62 resolution control at 00Z and a control at 12Z), 14 perturbed forecasts are made at the reduced T62 resolution. Global products from the ensemble forecasts are available from NCEP via anonymous ftp.

We found that the breeding based NCEP ensemble is able to extend the useful skill of numerical weather predictions by several days, making it possible to issue daily weather forecasts perhaps out to two weeks in advance during some winter periods. This may be especially useful since the low ensemble spread can identify these periods in advance. Analysis rank (or "Talagrand") diagrams show that the verification escapes the ensemble cloud only about 12 % of the time (in excess of what is expected due to the limited size of the ensemble).

1. INTRODUCTION

It is a common knowledge that weather forecasts fail with time. It has also been observed for a long time that the loss of skill in the forecasts does not occur at the same lead time every day. In the example of Fig. 1a, one could issue a confident 4-day forecast while in the case of Fig. 1b the ensemble correctly suggests a much greater uncertainty. Also, there are days when we have skill even beyond day 10 while on other days the skill may be lost as early as five days into the forecast. One reason for the ultimate failure of weather forecasts has been clear, namely our techniques, including our numerical models, are imperfect. Since the pioneering work of Lorenz (1963) we know, however, that this is not the only reason for forecast failure.

As Lorenz showed, the most fundamental cause of forecast failure is that the atmosphere is a chaotic system. This means that given an arbitrarily small error in our analysis of the initial state of the atmosphere, the forecasts are bound to fail after some finite time. This would happen even if we had a perfect model of the atmosphere. The estimated time of atmospheric predictability under "ideal" conditions (small initial error and perfect model) is of the order of 2-4 weeks (see, e. g., Toth, 1991).

1. General Sciences Corporation (Laurel, MD) at NCEP

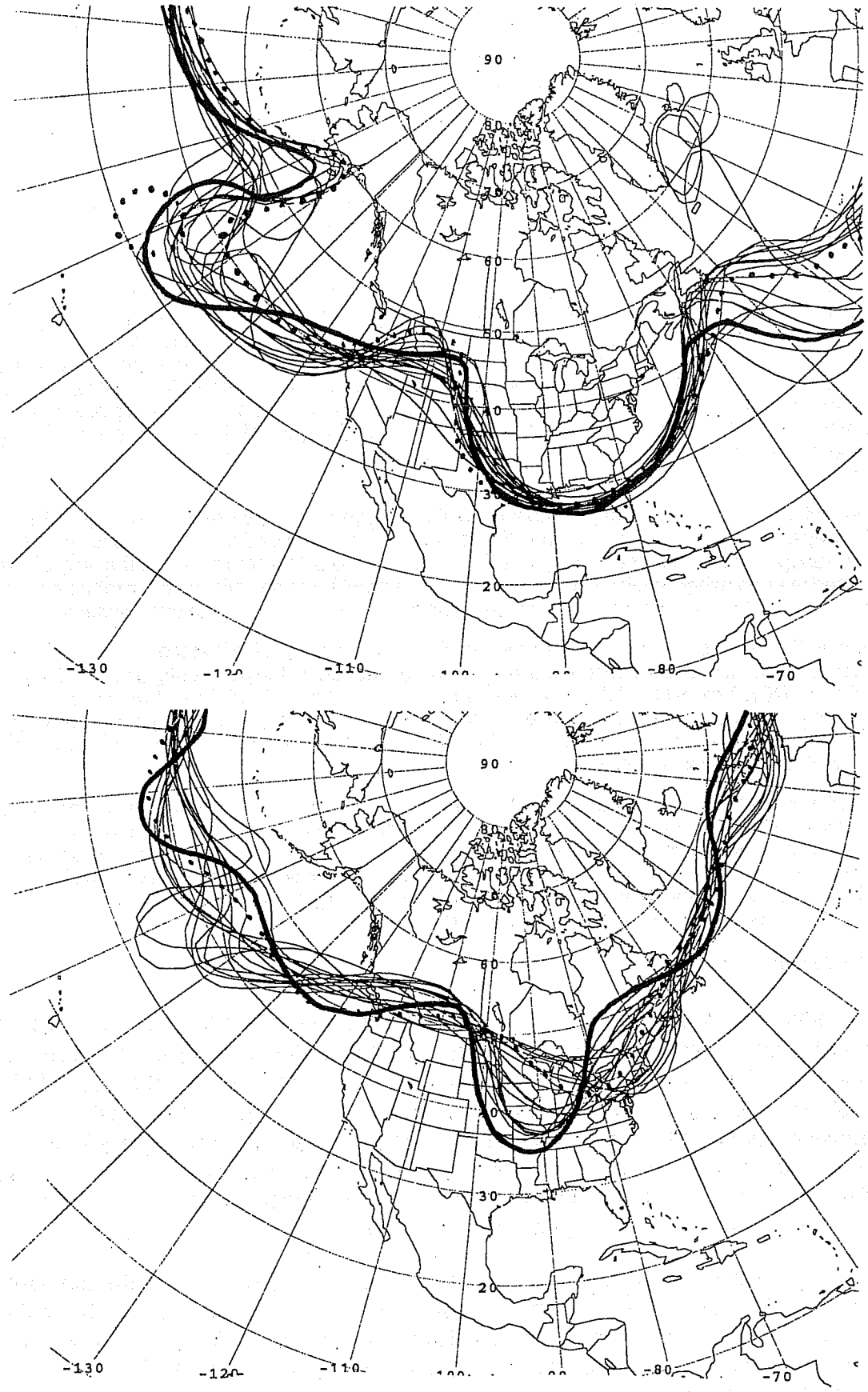

Fig. 1: 5640 m contour line of 500 hPa height field from all 17 108-hour lead time ensemble forecast members verifying 1996/03/20/12Z (top) and 1995/10/20/12Z (bottom). The dotted line marks the high resolution control forecast (MRF) and the heavy solid line is the verifying analysis.

There is nothing one can do about the ultimate loss of predictability. We will never be able to perfectly measure the initial state of the atmosphere and even so we would need to have a model that resolves *all scales* perfectly, otherwise errors from smaller scales would propagate into the resolved scales almost instantaneously. We can still aim at improving our models' physical parametrizations, spatial resolution, etc., but these improvements will *only extend*, and not change the fact of *finite predictability.* What we can do, however, is to predict the manner in which the loss of skill occurs in our forecasts. As described by Epstein (1969) and Leith (1974), one can artificially introduce small perturbations onto the analysis, representative of possible analysis errors, and integrate the same numerical model used for the control integration, starting from each of the perturbed analyses. With the availability of high speed computers and with the realization of how much can be gained from it, this technique, called ensemble forecasting, is gaining ground at major operational weather forecasting centers (Palmer et al., 1992; Tracton and Kalnay, 1993).

In this paper we will first take a look at what measures are available for tracking atmospheric instability that is responsible for the loss of skill (section 2). Then in section 3 we will discuss what kind of errors there are in the analysis. Based on the findings of sections 2 and 3, in section 4 we will present a practical procedure for creating ensemble forecasts. This procedure is based on the breeding technique that is discussed in more detail in a companion paper (Kalnay and Toth, 1996, this volume). Some results from the operational implementation of the ensemble forecast system at NCEP will also be presented in section 4. A discussion of open questions and other applications of the breeding based ensemble technique at NCEP is contained in section 5.

2. MEASURES OF INSTABILITY

2.1 Lyapunov characteristics

It is the presence of instabilities that give rise to chaos in certain dynamical systems. To measure how unstable a system is one can look at the global Lyapunov exponents:

$$\lambda_i = \lim_{t\to\infty} \frac{1}{t} \log_2 \frac{p_i(t)}{p_i(0)}$$

where p_i is an arbitrary linear perturbation introduced into the system at t=0. If one is interested in more than one Lyapunov exponent, a set of initial perturbations have to be evaluated and reorthogonalized periodically. If there is at least one positive global Lyapunov exponent in a system, the system will behave chaotically (Tsonis, 1992).

Knowing that a system is chaotic tells nothing about the possible changes in predictability over the attractor (i. e., changes depending on the initial condition). For that, one needs to look at the *local Lyapunov exponents* (LLVs, Trevisan and Legnani, 1995):

$$l_i(t) = \lim_{\tau\to 0} \frac{1}{\tau} \ln \frac{\|p(t+\tau)\|}{\|p(t)\|}$$

where t is some (ideally infinite) time after the initial perturbation has been introduced. Note that the global Lyapunov exponent is an integral of the local exponents over the whole attractor and the vectors corresponding to the exponents at any given point on the attractor are called the local Lyapunov vectors. The leading LLVs are, by definition, the vectors that are capable of producing the largest sustainable growth on the attractor

2.2 Finite time instability

It has been indicated by Lorenz (1965) and subsequently pointed out by Lacarra and Talagrand (1988) and Farrell (1988) that, over a prespecified period and for a chosen norm, growth faster than that of the LLVs can exist in certain systems. If L is a tangent linear model that evolves a perturbation ahead in time and L* is its adjoint then the eigenvectors of L*L (or the singular vectors, SVs of L) are the "optimal" vectors. These vectors are optimal in a sense that given the full phase space of a system, they represent the maximum possible growth for a prespecified time interval (for which L is defined) and with respect to a chosen norm that is used in the definition of the inner product.

2.3 Comparison of Lyapunov and singular vectors

Both the LLVs and the SVs offer a full orthogonal basis for the description of the phase space of dynamical systems and they both can be arranged into a spectrum of vectors with decreasing instantaneous growth rates (see Fig. 2). Beyond these similarities, there are important differences between the two sets of vectors:

i) While the LLVs are basic and general properties of a dynamical system, the SVs are specific in a sense that they pertain to a specific time interval and norm. Their structure may change drastically by changing these arbitrary parameters.

ii) By definition, all perturbations turn toward the leading LLVs. This includes the SVs as well which therefore turn away from their initial direction in the phase space into the direction of a leading LLV (Szunyogh et al., 1996). So while the LLVs act as attracting directions or magnets in the phase space, the SVs represent "repelling" directions.

iii) While the LLVs provide sustainable growth on the attractor, the SVs' growth cannot be maintained after the optimization period ends. The SVs' super-Lyapunov growth is actually due to a one-time abrupt phase space rotation of the initial vectors toward leading LLVs. In order to achieve similarly fast growth again, one would need to reintroduce an initial SV into the perturbation system again.

iv) LLVs represent natural perturbation development. The first D LLVs, where D is a system's Kaplan-Yorke dimension, span a subspace of the full phase space that the system can naturally visit, i. e., the subspace of the attractor. In contrast, nothing guarantees that the leading SVs would have to lie along the attractor. If this would be the case, for which we have several indications that we discuss further down, the initial SVs could oocur in a system only due to special forcing; the system, without specific forcing, could not naturally assume an initial SV perturbation.

v) Since analysis cycles can, as discussed by Kalnay and Toth (1995, this volume) be considered as perturbation models run on the true state of the atmosphere, the leading LLVs can naturally occur and dynamically amplify as analysis / first guess errors. In contrast, SVs, defined with norms used in everyday practice, cannot appear and amplify dynamically as perturbations (analysis errors) in a similar manner.

The above point constitutes a major difference between the LLVs and SVs. The result is that while the leading LLVs can dynamically develop from arbitrary errors in the analysis, SVs defined through commonly used norms can only appear as analysis errors due to specific random observational errors, projecting onto the SVs themselves. This question will be further discussed in section 3.

2.4 Do the singular vectors lie off the attractor?

The unique nature of the leading SVs is further highlighted by simple model experiments suggesting that the SVs may not lie along the attractor. First, we would like to quote Anderson (1995) who, in a three variable Lorenz model found that the leading SVs point toward phase space

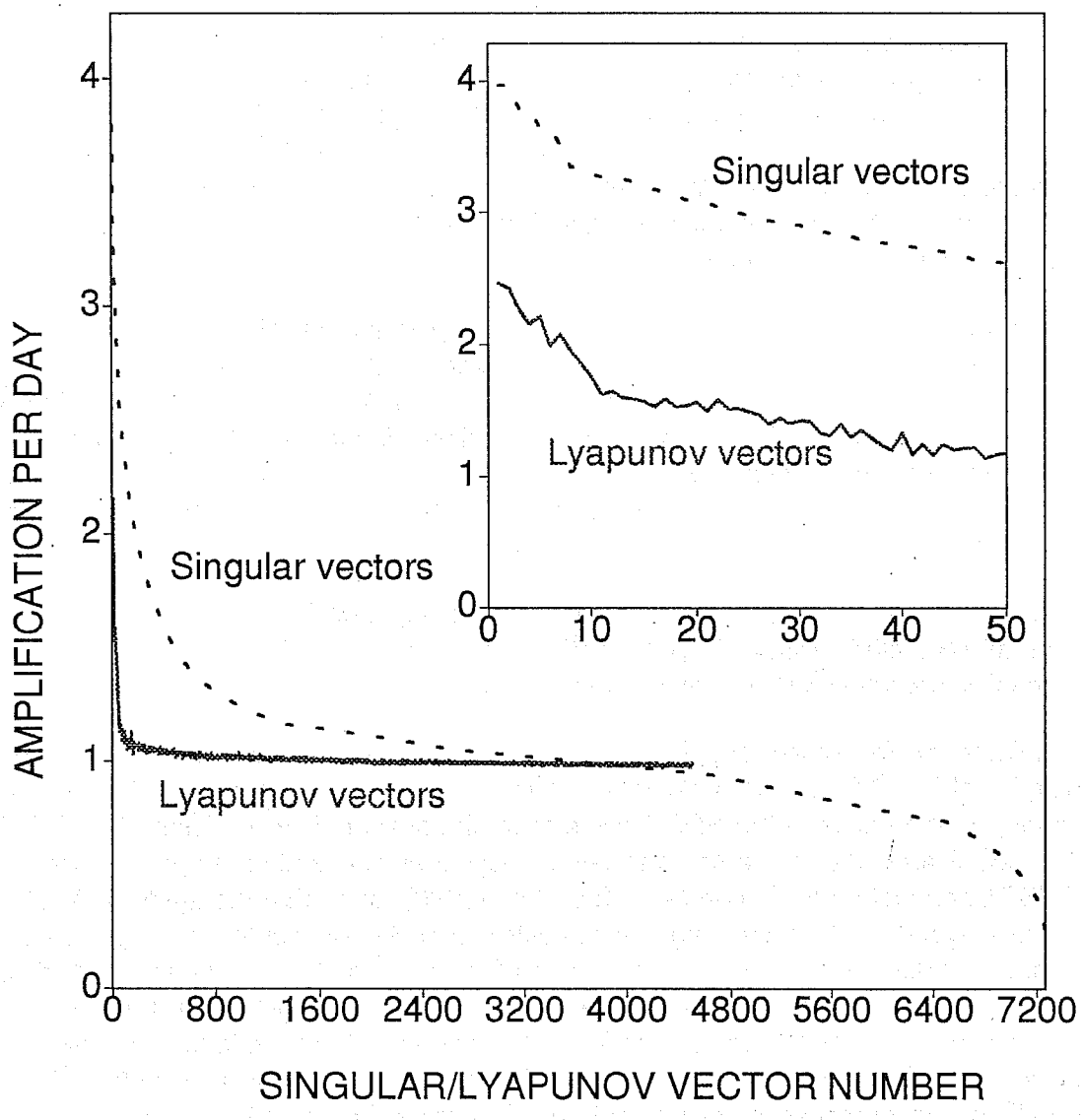

Fig. 2: Spectrum of 24-hour amplification rates for singular vectors (SVs) optimized for a 24-hour period in a T10, 18 level version of the NCEP MRF with totla energy norm (dotted line). The first1346 and the last 795 SVs have been computed and a simple interpolation is used to estimate the middle of the amplification spectrum. For the same flow pattern a large portion of the Lyapunov vectors (LLVs) have also been estimated with a 24-hour orthogonalization time (for more details see, for example, Legras and Vautard, 1996) and the associated 24-hour amplification values are also plotted (solid line). (From Toth et al., 1996.)

regions that are not visited by the system at all. In turn he found that when the SVs are used as initial ensemble perturbations, they do not perform as well as bred-type perturbations.

Anderson's results have been confirmed by Legras and Vautard (1996, this volume), who found that in the Lorenz 3-variable model the first two LLVs are tangential to the attractor while non of the SVs are on the attractor. Legras and Vautard (1996) also pointed out that, in the same way in which almost any perturbation in the far past, when integrated with the tangent linear model will become the leading LLV at current time, almost any perturbation in the far future will become the leading SV (with long optimization time) when integrated backwards in time to current time, using the adjoint model (a process which cannot occur naturally).

Vannitsem et al. (1995, personal communication) experimented with a 3-variable Lorenz model that has a dimension close to 2. It means that the first two LLVs can span most of the phase space visited by the model. Still, they found that three LLVs are needed to explain the SVs, indicating that in this model, the SVs lie off the attractor. Similarly, it was found that when explaining how random errors develop in that system, one needs all three SVs but only the two leading LLVs. The suggestion again is that the LLVs offer a more natural basis for describing natural perturbation development.

In a study of a 3-level quasi-geostrophic model, Oortwijn and Barkmeijer (1995) found that with very small magnitude initial SV perturbations they can force their system into phase space domains that are very rarely or never visited by the unperturbed system.

With a Cane-Zebiak simple coupled ocean atmosphere model Xue et al. (1996) found that the leading initial SV of the ocean part of this model, to a large extent, does not depend on the initial condition, on the phase of the ENSO cycle or on the seasonal cycle. Chen et al. (1995), with a similar model, arrived at the same conclusion though they note that the singular *values* do depend on ENSO/seasonal cycle. In contrast, Xue et al. (1996) found that the leading SV at final time, which must approximate the leading LLV well with a long, 6-month optimization period (see Legras and Vautard, 1996, in this volume), does depend on initial conditions. The fact that the SVs don't show sensitivity to drastic changes in the basic state may suggest that their extra-Lyapunov growth may be due to factors that are *independent of ENSO*, the major source of natural instability in the system. The question arises whether the instabilities associated with the SVs represent physically plausible processes or just patterns that are within the phase space of the model but would never realize naturally because they are off the attractor.

With a T10 truncated version of the NCEP MRF Toth et al. (1996) found that the initial SVs defined by the use of the total energy norm have an unusually strong vertical tilt and hence are geostrophically highly unbalanced. These perturbations, again, seem to be off the attractor. The initial fast perturbation growth of the SVs is associated with the fast geostrophic adjustment through which the SVs assume a vertical tilt that is characteristic of observed baroclinic instabilities. After this abrupt (24-hour or so) adjustment, the SVs lose their super Lyapunov growth.

We would like to emphasize that some of the characteristics of the SVs discussed above may be model specific and that further research is needed to better understand the nature of SVs. One has to note, however, that the calculation of the SVs through singular value decomposition is a mathematical procedure in which there is no built in mechanism to prevent the occurence of physically implausible results. Therefore, unless a norm that assumes a realistic initial error distribution is used in the computation of the SVs, any physical interpretation of the SVs has to be made with extra care.

2.5 Nonlinear perturbations

So far we have discussed linear measures of instability. In a linear world, the leading LLV(s) may be just perfect to describe possible error evolution since all perturbations with time turn toward the leading LLV. The atmosphere and our forecasts, however, evolve nonlinearly. In this situation the forecast error at any lead time will depend on the initial error field. So if our goal is to estimate forecast error, we need to consider each possible initial error pattern and follow its nonlinear evolution.

To do so, we need to have an estimate of the initial error distribution in a probabilistic fashion. Once this estimate is available - and this estimate is the subject of the extensive research on initial ensemble perturbations and analysis error covariance - the problem is relatively simple. Theoretically, one could take the Liouville equations (Ehrendorfer, 1994) and integrate them for any particular lead time to obtain the forecast probability distribution of the system, given its initial probabilities. This approach, however appealing and simple it sounds, is computationally not feasible for systems that contain more than a few variables. The only alternative is to take samples from the initial estimated probability distribution of the system with frequencies proportional to the probabilities and run the nonlinear forecast model for an ensemble of those initial conditions. Before we discuss in section 4 how this ensemble is generated at NCEP, we turn toward the more basic questions of estimating and sampling analysis errors in the next section.

3. ESTIMATING AND SAMPLING ANALYSIS ERRORS

3.1 Sampling analysis errors

As discussed in Kalnay and Toth (1995, this volume) the analysis contains both fast growing errors generated dynamically by the successive use of the first guess forecast fields in the analysis cycle, and random errors that originate from observational and analysis errors. There are indications that there may be only a few perturbations [O (10)] in any region that can grow dynamically fast at any given time. These directions can be well estimated by either the breeding method or the singular vector approach. In contrast, there are other directions, in which perturbations are neutral or decaying, and their number is on the order of the number of variables in our models [O (10^6)].

So we are in a situation where there are 10^6 possible independent error patterns. A large portion of the actual error (say, half) lies in a few directional subspace of the fast growing vectors that we can well estimate. Evidently, these directions can be well sampled as well with the number of perturbations used currently in ensemble prediction (O 10). The other half of the error lies in the rest of the phase space with a dimension of 10^6 or so. It is evident that we cannot sample the subspace of possible neutral and decaying errors well with 10 or so perturbations.

With an infinite number of ensemble perturbations the best approach would be that of Houtekamer et al. (1996), who realistically sample both the small dimensional growing and the large dimensional neutral/decaying part of the error. Houtekamer et al. run parallel analysis cycles with randomly generated "observational" error added in each cycle, to arrive at perturbations that have a realistic magnitude of both growing and neutral/decaying components. We would like to argue, however, that in case of a limited sample where the number of perturbations is largely below the number of directions to be sampled (like the case for the neutral/decaying part of the error) one needs to weight the magnitude of the error by the ratio between the sample size [O (10)] and the dimension of the phase space ([O (10^6)] for the neutral/decaying subspace). Since this ratio is so small, the ECMWF and NCEP perturbation strategies ignore the role of neutral/decaying perturbations and focus solely on the fast growing part of analysis errors.

Houtekamer et al.'s ensemble approach, which can be considered as an extension of the breeding cycle, may not result in a noticeable difference from that of an ensemble based on breeding. Both methods sample similarly the fast growing bred vectors which play the crucial role in error development. Hence the effect of the addition of the neutral/decaying perturbations by Houtekamer et al. can be expected to be relatively small.

3.2 Estimating growing analysis errors: Bred vs. singular vectors in ensembles

So far breeding (see Kalnay and Toth, 1995, this volume, and Toth and Kalnay, 1993, 1996) is the only method that has been advanced as a way of estimating the fast growing part of the analysis error. The singular vector approach, though it also determines fast growing directions in the phase space, is not an error estimating technique. It is actually based on the simple assumption that the initial error distribution is *white noise*, i. e. that all directions in the phase space are equally likely as analysis errors (see, for example, Molteni et al., 1995). When we choose a norm for the computation of the fastest SVs we assume that the analysis error, with respect to that norm, has a random white noise distribution. We would like to argue that with any obvious choice of norm (such as rms, kinetic or total energy, etc.) this assumption is not valid. This is because of the dynamical error evolution within the analysis cycle that results in excessive magnitude perturbations in the directions of the leading LLVs or nonlinear bred vectors.

Once the analysis error initial distribution is estimated one could use the SV approach. However, an appropriate norm needs to be found first, in which the analysis error looks like white noise. Solving this inverse problem is not a trivial task and so far, apart from Houtekamer's (1995) approach that incorporated some statistical (but not dynamical) properties of the estimated analysis error, it has not been addressed.

If one would assume that the bred vectors sample well the fast growing analysis errors and would be able to find a norm in which the bred vectors would look like white noise, with the singular vector approach one would be able to determine orthogonal directions within the subspace span by a set of bred vectors, in order of descending growth rates. At NCEP, instead of solving this difficult inverse problem, we use a set of bred vectors from separate breeding cycles started with independent initial seeds, as initial ensemble perturbations. These bred vectors, unlike the perturbations generated via the SV approach (Buizza et al., 1993) have some correlation (typically 0.2 - 0.3 globally). We speculate that strict orthogonality may not be a crucial requirement in a nonlinear situtation where orthogonality is quickly lost anyway. The bred vectors thus offer a sample of quasi-orthogonal vectors that are plausible fast growing analysis errors with roughly equal growth rates; the fastest LLVs are naturally combined in a random manner, with weights statistically proportional to their growth rates.

The resulting bred vectors are similar to the initial perturbations used at ECMWF in a sense that there, too, the SVs are combined randomly and hence the growth rates of initial perturbations are similar to each other. However, since no special norm is used in the generation of the singular vectors, certain directions in the phase space may be overemphasized. Following earlier discussions in sections 2 and 3, one needs to consider that the bred vectors appear and amplify in the analysis via dynamical means. Thus their amplitude must be much larger in the analysis error field than arbitrarily chosen error patterns. The leading initial SVs are just such arbitrary perturbations since they don't seem to appear naturally. Considering the large number of dimensions [O (10^6)] in which random error is spread, much of the actual forecast error may come from initial LLVs or bred vectors as analysis errors and not from initial SVs that are present only due to random errors. This would be true even if the bred vectors would have only a few times larger

amplitude in the analysis than arbitrary perturbations. This question obviously needs to be addressed in a more quantitative manner; the next subsection offers only a first look at the problem.

Ensemble forecasts based on SVs at ECMWF and on bred vectors at NCEP may still behave very similarly, though. And some verification numbers we subjectively compared looked indeed similar. This may happen because, as Szunyogh et al. (1996) found with a T10 version of the NCEP MRF model, the initial SVs, if optimized for a 3-day or longer period, turn practically into the direction of the leading LLVs by the end of the optimization period. At ECMWF, the optimization period is only slightly shorter, 2 days long. So it is possible that after 2-3 days, the two sets of perturbations would project relatively well on each other. And since nonlinearities on the larger scales become important only after 2-3 days, the two systems, despite the difference in initial perturbations, may yield similar results.

3.3 Bred vectors as analysis errors

We will present two attempts to compare the properties of bred vectors to independent estimates of analysis uncertainty. In the first comparison we tried to establish if the bred vectors are able to reproduce the statistical characteristics of analysis uncertainty. Here we measured analysis uncertainty as the difference between two independently run analysis cycles. The two analysis cycles were identical except that the initial first guess fields were different (and in one of the two cycles the first guess was produced as an average of two slightly perturbed 6-hour forecasts). The two analysis cycles were run for more than 30 days in May-June, 1992 and then the rms difference between the two series of analyses for the last 24 days of the experiment were averaged. Similarly, we computed the rms average of the bred perturbations from a breeding cycle (with hemispherically fixed rescaling) for the same period. As we can see from the vertical cross sections of the vorticity fields (Fig. 3) the bred vectors reproduce well the varying degree of uncertainty present in the analysis. Regarding the horizontal distribution of analysis errors (not shown) breeding is able to capture the variability over the Southern Hemisphere (where satellites provide a quasi-homogeneous data coverage). Through the use of a geographical mask for rescaling within the breeding cycle (see next section) we are also able to reproduce spatially dependent uncertainty over the Northern Hemisphere where data coverage is far from uniform.

In the second comparison we tried to establish if the bred vectors are able to capture the dynamical characteristics of analysis error on a day by day basis. In this approach we estimated analysis error as a difference between the analysis (A) or first guess (FG) and observations (O), at observation locations. In particular, we measured how much of the difference field (O-FG) or (O-A) projects onto the bred growing vector (G, estimated from one breeding cycle). The procedure used is identical to that reported in Kalnay and Toth (1996, see their Fig. 7 in this volume) except that here we use the whole Northern and Southern Hemisphere extratropics (instead of smaller overlapping regions). The results indicate that in the first guess (analysis) error fields defined above there is a roughly 2.0 (0.9) % projection, in terms of rms total climate variance, onto the bred growing vector of the day when the two hemispheres are averaged (equivalent to 1-2 m explained error at 500 hPa height in the first guess). Random perturbations would have a much smaller projection. We also computed the same projection using bred vectors valid on other days within the same months and obtained 2.5-3 times smaller values.

The above results provide strong evidence that the analysis uncertainty is far from being white noise, and that the bred vectors can reproduce the major characteristics - both statistical and dynamical - of this uncertainty.

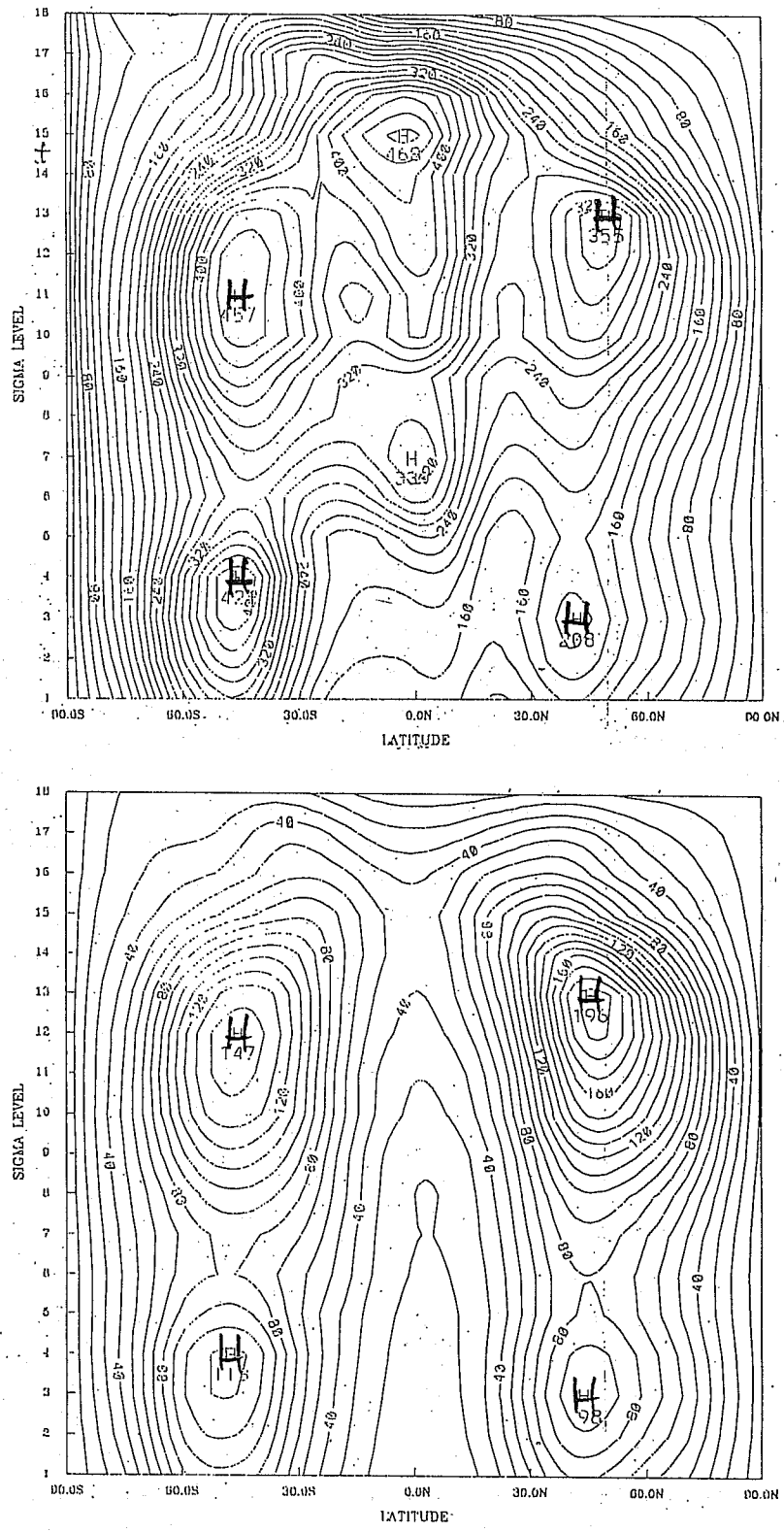

Fig. 3: Vertical distribution of uncertainty in the vorticity fields present in the control analysis as determined from the rms difference between two analyses from independently run NCEP analysis cycles between May 23 and June 15, 1992 (top, labels multiplied by 10^8). The analysis cycles were practically identical except that the initial first guesses differed slightly. The values shown are smoothed and the overall global mean is scaled to one. The same rms difference between a pair of positively and negatively perturbed short range ensemble forecasts from a breeding cycle for the same period is also shown (bottom, labels are multiplied by 10^7).

4. ENSEMBLE FORECASTING WITH BREEDING

4.1 Operational implementation

Ensemble forecasting has been operational at NCEP since December 1992. As indicated by Kalnay and Toth (1996, see their Fig. 3 in this volume) the current operational ensemble configuration at NCEP consists of 17 model runs every day out to 16 days lead time. At 00Z, a T126 control is run out to 7 days lead time, after which it is truncated and run at a T62 reduced resolution. Another control is also run started from identical but truncated T62 initial conditions. In addition, there are ten perturbed forecasts, started with five bred perturbations both added to and subtracted from the control analysis. The breeding cycle is incorporated within the extended-range forecast ensemble so the generation of the initial perturbations is basically cost free (see Fig. 4). At 1-day lead time we take the difference between a pair of positively and negatively perturbed forecasts. After rescaling its amplitude, this difference field provides as the initial perturbation for next day's ensemble forecasts.

As mentioned in section 3.3, the breeding method, when used with a hemispherically fixed rescaling factor, cannot reproduce the horizontally varying uncertainty in the analysis that is present due to the inhomogeneous data coverage. To account for this factor, we follow a simple procedure where we compute a rescaling factor for each gridpoint, based on observed growth in a surrounding area. The perturbations are actually scaled back to a fixed geographical mask illustrated in Fig. 5. After some tuning experiments we set the overall perturbation amplitude at around 12 % of rms climate variability. This value is perhaps slighly above the estimated analysis error. However, the non-systematic part of the model error would also project onto the fast growing patterns at later lead times, which may explain why this perturbation size gives close to optimal performance in the medium and extended range.

Considerable efforts have been directed toward displaying the information contained in the ensemble in a user friendly manner (Tracton, 1994). Beyond the "spaghetti" plots shown in Fig. 1, we have the ensemble and cluster means, variances, different probability and other charts available to the forecasters on line, shortly after the forecast integrations finish. These products are also accessible to the different field forecast offices throughout the country and are available to the wider user community as well via ftp (at nic@fb4.noaa.gov). Experience with the use of the ensemble is building up at all levels and at NCEP we would like to help this process by the development of new products that are based on the ensemble

4.2 Verification results

The ensemble results should be subjected to both objective and subjective evaluation. As for the latter, Toth et al. (1996) found that the operational NCEP ensemble offers valuable tools for the synoptician for the medium and extended range and occasionally even for the short range. At NCEP and also at more and more regional offices of the National Weather Service, the global ensemble forecasts constitute an integral part of the numerical forecast tools used every day. At this point, the ensemble is used primarily to assess the reliability of the forecasts and to establish scenarios alternative to that offered by the high resolution MRF control forecast.

An exhaustive objective verification of the NCEP operational ensemble forecast system would involve the computation of a number of statistics over a long period of time and is beyond the scope of this study. Instead, we will focus on a two-week period in December1995 in which the NCEP global forecasts were particularly skillful. Fig. 6 shows the pattern anomaly correlation

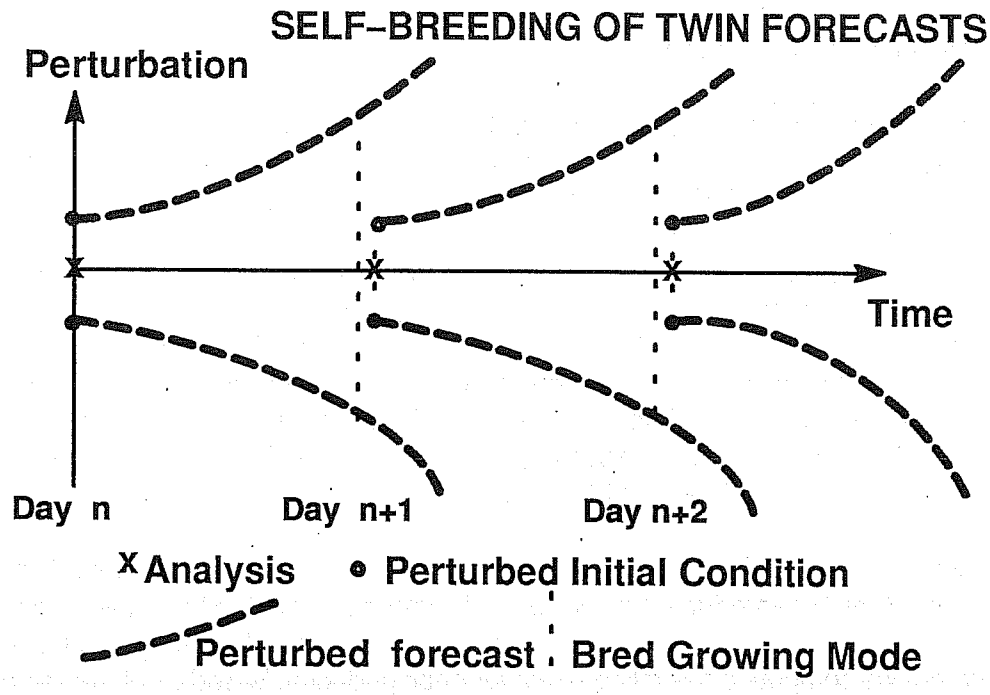

Fig. 4: Schematic of a self contained breeding pair of ensemble forecasts. Note that breeding is part of the extended ensemble forecasts at NCEP and that the creation of efficient initial ensemble perturbations requires no additional computing resources beyond that needed to run the forecasts themselves. (From Toth and Kalnay, 1996.)

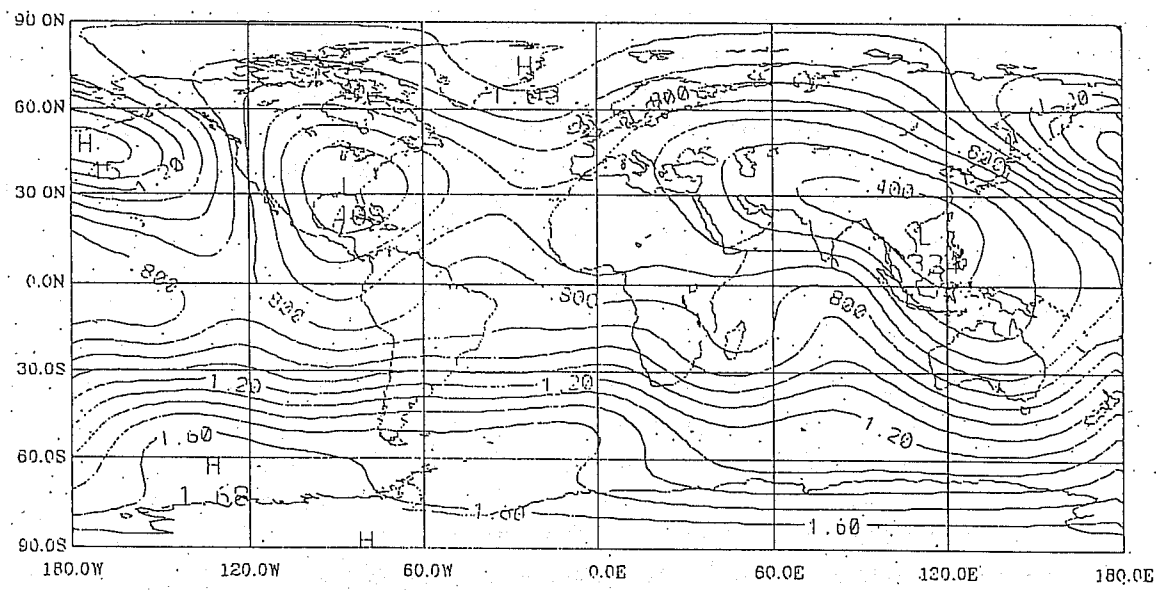

Fig. 5: Relative regional uncertainty (for 500 hPa streamfunction) present in the control analysis as determined from the rms difference between two analyses from independently run NCEP analysis cycles for a period in April-May 1992. The analysis cycles were practically identical except that the initial first guesses differed slightly. The values shown are smoothed and the overall global mean is scaled to one.

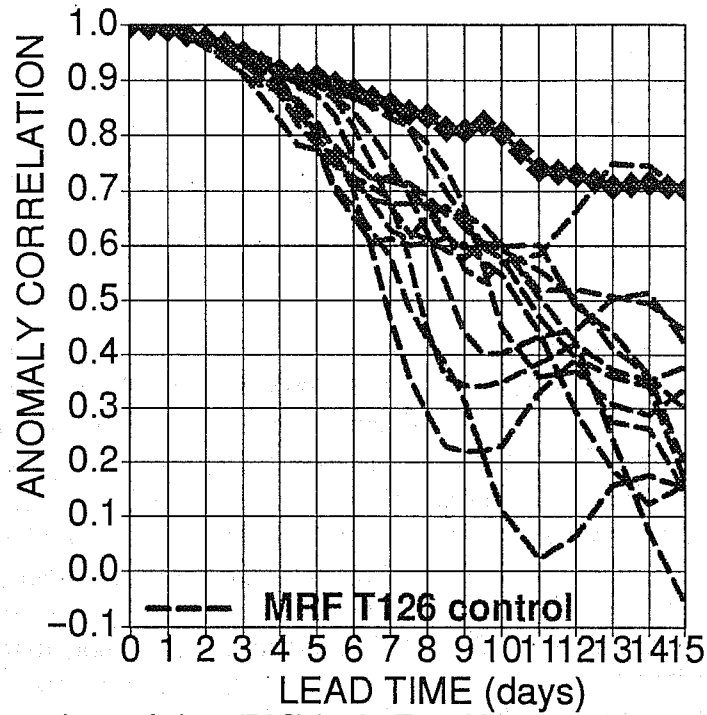

Fig. 6: Pattern anomaly correlations (PAC) for the T126 MRF control forecast (blue, dotted) for the 500 hPa height field over the Northern Hemisphere extratropics, for 13 forecasts initiated between 9 and 22 of December 1995 (verification for the forecast started 18 December is missing). The scores for the forecast started on 10 December is highlighted by diamonds.

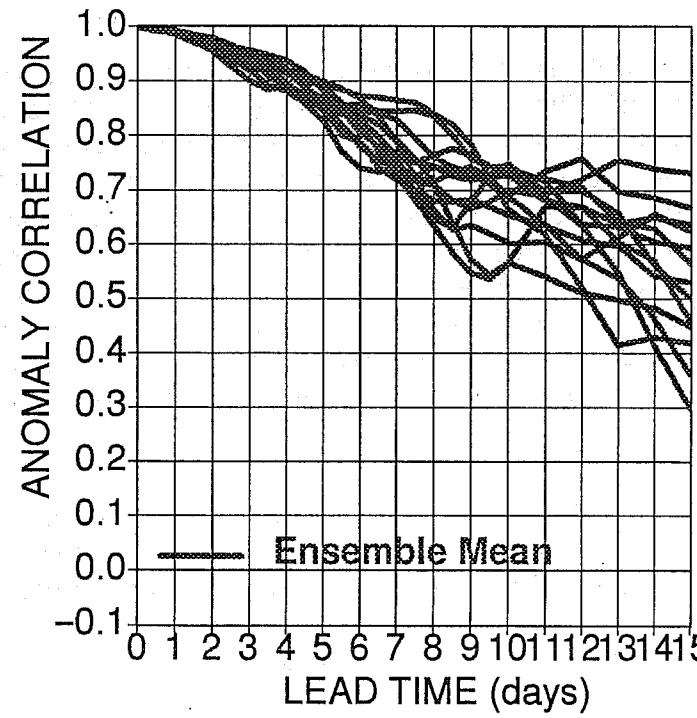

Fig. 7: Pattern anomaly correlations (PAC) for th 17-member NCEP ensemble mean forecast (solid, red) for the 500 hPa height field over the Northern Hemisphere extratropics, for 13 forecasts initiated between 9 and 22 of December 1995 (verification for the forecast started 18 December is missing).

(PAC) score for the 15 days long MRF forecast started on 10 December 1995. The score stays above .7 all the way out to 15 days lead time. (On average, PAC drops below 0.7 around 6 days lead time during winter.) This is probably the first time that any weather forecast showed such a high skill beyond two weeks that is close to what has been considered as the ultimate limit of atmospheric predictability (Lorenz, 1969). Obviously, the model itself did not introduce substantial error or bias into the forecast. We can also note in Fig. 6, however, that during the period 9–22 December 1995 this forecast was exceptional: there was only one other forecast (from 12 December 1995) that could get close to its performance, and only for the last few days of the forecast period. If one forecast is successful, why the others from the same circulation regime are not?

In Fig. 7 we present similar verification scores for the operational 17-member NCEP ensemble mean forecast. One notices two major differences between Figs. 5 and 6: the ensemble scores are, overall, considerably higher than the MRF control scores, and they have much less variability from one forecast to the next. This means that the ensemble mean consistently provides a more reliable forecast than a single control forecast. This is one of the main advantages of the ensemble forecast technique. If one considers a PAC of 0.5 or 0.6 to be the useful level of skill for daily weather prediction, then the ensemble mean, as seen from the averaged scores in Fig. 8, is extending the practical limit of predictability by 5–6 days during this period. For example, while the MRF scores at day 9 range between 0.2 and 0.8, those for the ensemble mean are between 0.55 and 0.8, with the majority above 0.6. The separation between the scores for the MRF and the ensemble mean become even larger by day 15: most MRF forecasts score below 0.45, while the majority of ensemble forecasts score above that value. Since there are always errors (however small they may be) in the initial analysis, they will amplify, due to the chaotic nature of the atmosphere, when only a single forecast is used. One can get high scores for the control only by chance, due to some fortunate arrangement in the initial error field. The ensemble mean forecast, on the other hand, can filter out some of those nonlinear errors amplifying due to the atmospheric instabilities.

A good example of the impact of ensemble averaging can be seen in Fig. 9, where the PAC for both the control MRF forecast and for the ensemble mean are shown for 12 December 1995. While the skill of the control forecast drops into the range of 0.5–0.6 for a couple of days – not a bad performance on its own – the mean of the ensemble remains practically above 0.7 all the way out to 15 days. It is important to note that the spread of the ensemble members around the ensemble mean was unusually low in this forecast – only 66 % of the spread for the ensemble averaged for the winter of 1994/95 (S. Tracton, personal communication). So not only was the model accurate, but the atmosphere itself was in a circulation regime with higher than average inherent predictability during a long period in December 1995. And, very importantly, it was possible to know in advance that the atmosphere was very predictable from the lower than average spread of the forecasts.

As an example, in Fig. 10 we present the forecast started on 12 December 1995, along with the verifying analysis. One can see that the ensemble mean forecasts well captured many of the major changes in the hemispgeric flow configuration even with long lead times. For example, the position of the low pressure area on the east coast of the United States is well predicted at 216 hours (9 days) lead time. It is interesting to note that this low pressure area is less pronounced in the MRF control forecast (Fig. 11), indicating that though the ensemble mean is generally a smoother field, it is far from being a washed-out product. Another area of difference between the ensemble mean (Fig. 10) and the control forecast (Fig. 11) at 216 hours lead time is around longitude 80 E. Here the control forecast built up a major high pressure system that did not verify. In contrast, the ensemble

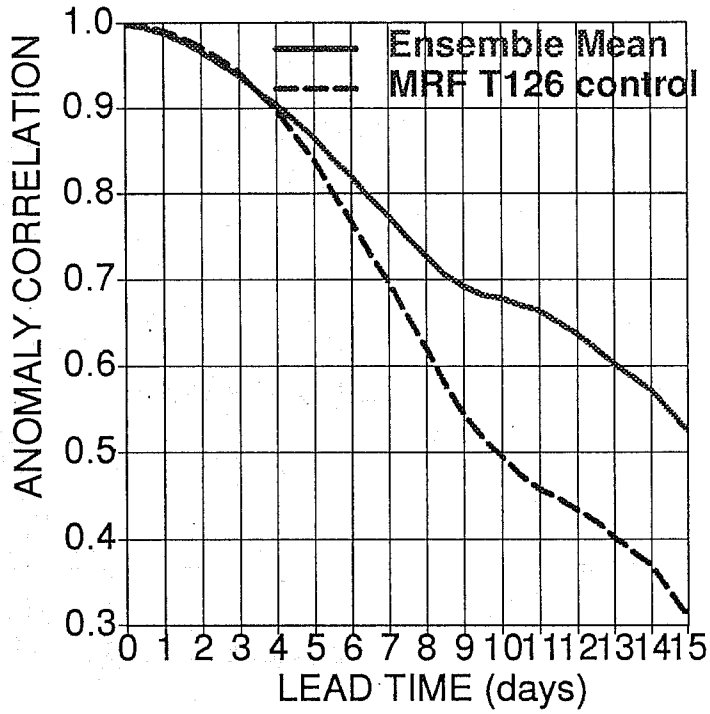

Fig. 8: Average of pattern anomaly correlations (PAC) for the 17-member NCEP ensemble mean forecast (solid, red) for the 500 hPa height field over the Northern Hemisphere extratropics, for 13 forecasts initiated between 9 and 22 of December 1995 (verification for the forecast started 18 December is missing). PAC for the MRF T126 control forecast (dashed, blue) for the same period is also shown.

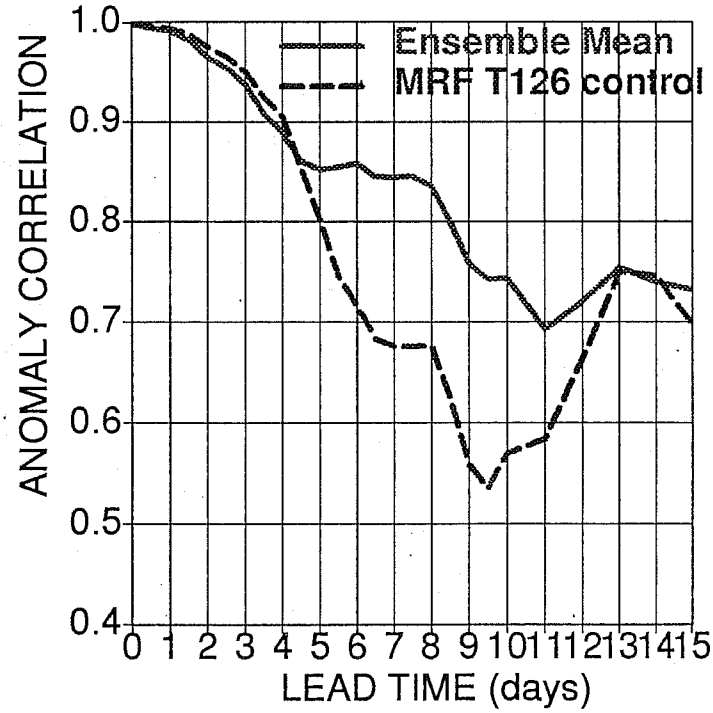

Fig. 9: Pattern anomaly correlation (PAC) for the 17-member NCEP ensemble mean forecast (solid, red) for the 500 hPa height field over the Northern Hemisphere extratropics, for the forecast initiated on 2 December 1995. PAC for the MRF T126 control forecast (dashed, blue) for the same day is also shown.

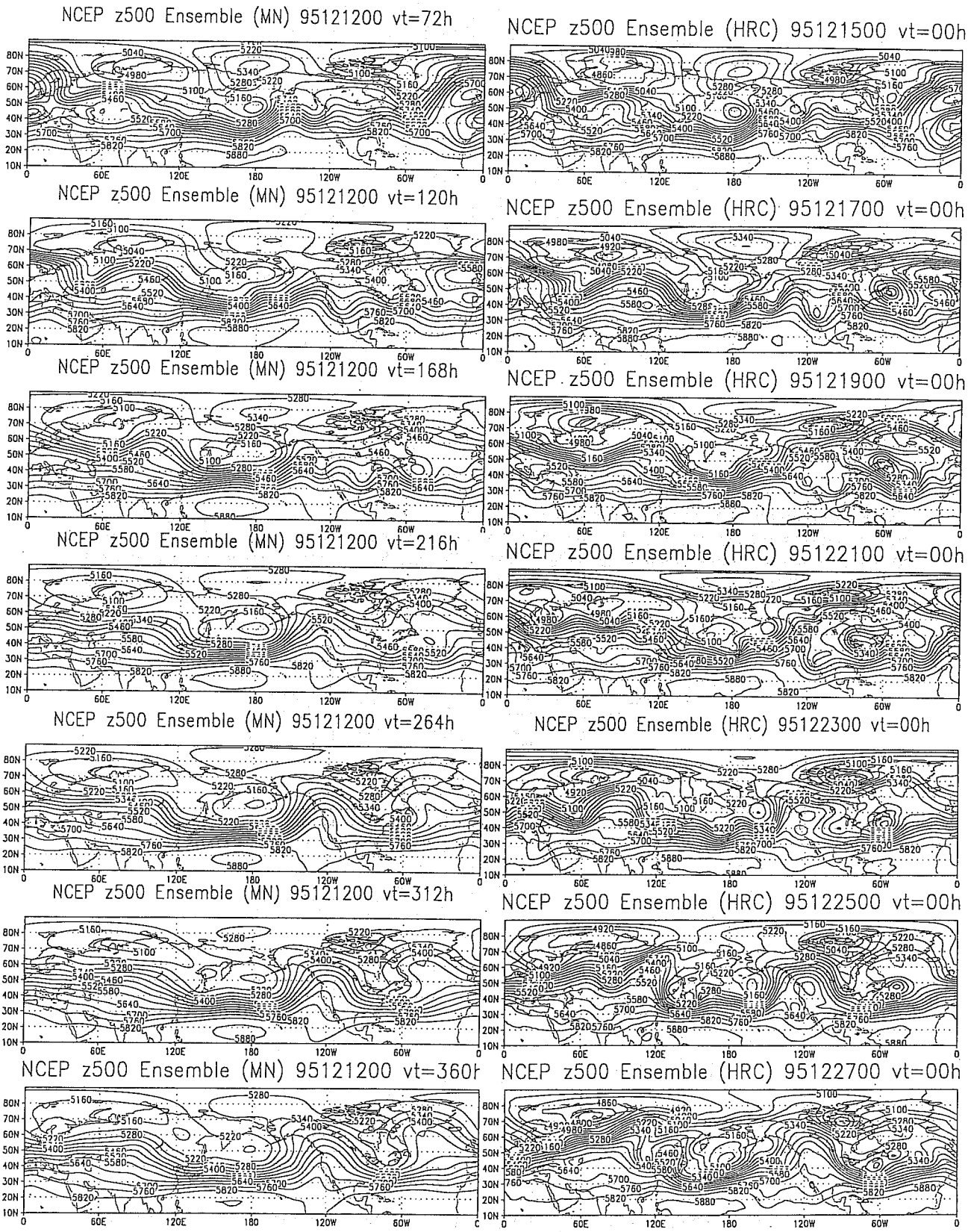

Fig. 10: 500 hPa geopotential height of the 17-member NCEP ensemble mean forecast initiated on 12 December 1995 (left) at 3, 5, ..., and15 days lead time, and corresponding analyses (right).

mean flow is more zonal, closer to verification, testifying again that the ensemble mean can provide a useful forecast guidance. Note also how well the emergence and maintainance of the ridge over the west coast of the US is predicted during the second week of the forecast in the ensemble mean.

Though the skill of the ensemble mean forecast from the 12 December was perhaps the best during the two weeks period studied here, the other forecasts, as can be inferred from the skill scores, did also well. The forecasts and the corresponding verifying analyses in Fig. 12 again demonstrate that the ensemble mean forecast was able to predict many of the important synoptic scale features and their change during December 1995, even 15 days in advance. For example, the ensemble mean forecasts well indicated the retrogression of the ridge over the west coast of the US during the first half of the verification period. As a result of this change, the Pacific trough became shorter from the east by the end of the period, which was also indicated by the ensemble. Over the east coast of the US, the continual reinforcement of the trough with arctic air from the north throughout the verification period was also well predicted by the ensemble. In the Atlantic, the ensemble mean forecast well captured the more and more westward tilt of the ridge/blocking pattern that developed during the period. The split flow over much of Europe near the end of the period was also predicted by the ensemble. Finally, a high amplitude ridge around longitude 80 E was also well indicated by the ensemble around 28 December 1995.

We would like to emphasize again that the NCEP MRF model performed exceptionally well during December 1995. This is because, first, model systematic errors were apparently small. And second, the atmosphere's intrinsic predictability was high – note that strong zonal flow and blocked flow configurations prevailed over the Pacific and Atlantic, respectively, during much of the period studied here (see Figs. 10 and 12). This is a flow configuration that Toth (1993) found, in an observational study, most stable. So such excellent ensemble mean scores may not be typical for other periods, and we should generally expect a more modest performance, especially during the summer months that is a difficult season for the global ensemble. However, our study shows that there are periods during which useful daily weather forecasts can at least be made through 15 days lead time, with relatively high confidence.

Another important characteristic of the NCEP ensemble is revealed in Fig. 13. To arrive at the analysis rank (or "Talagrand") distribution for the 17-member NCEP ensemble, at each grid point and for each forecast the 17 members of the ensemble are first arranged in an order of increasing height values, thus defining 18 bins (including 2 open ended bins). The verifying analysis falls into one of these bins, and the frequency observed for each bin is accumulated over all grid points and initial times. A perfect model and ensemble would have an expected value of 5.55% (100%/18 categories). The underlying assumption is that each ensemble member is equally likely and so is each forecast bin, including the two extremes, both with an open end. If the ensemble spread is insufficient, or the model has some kind of systematic error (bias), the distribution would be U-shaped, with excess cases where the verifying analysis falls outside the range of the ensemble.

The distributions shown in Fig. 13 have a U shape (except for the 12-hr lead time forecasts) but the values in the extreme categories are not too high: if we average all lead times, only in 12 % of the cases does the verification fall outside the cloud of the ensemble, in excess of the value expected from the limited size of the ensemble. This means that the forecaster can be reasonably confident that the verification will be within (or close to) the cloud of the ensemble. Note that though the ensemble starts out with a good estimation of errors at 12-hour lead time, the most excess cases occur around days 2 and 4. This is the period during which the forecasts, which start from observed initial conditions, swiftly adjust to the model climatology (model drift) and the forecasts may show some small but systematic errors due to this adjustment (see, for example, Anderson, 1995). This

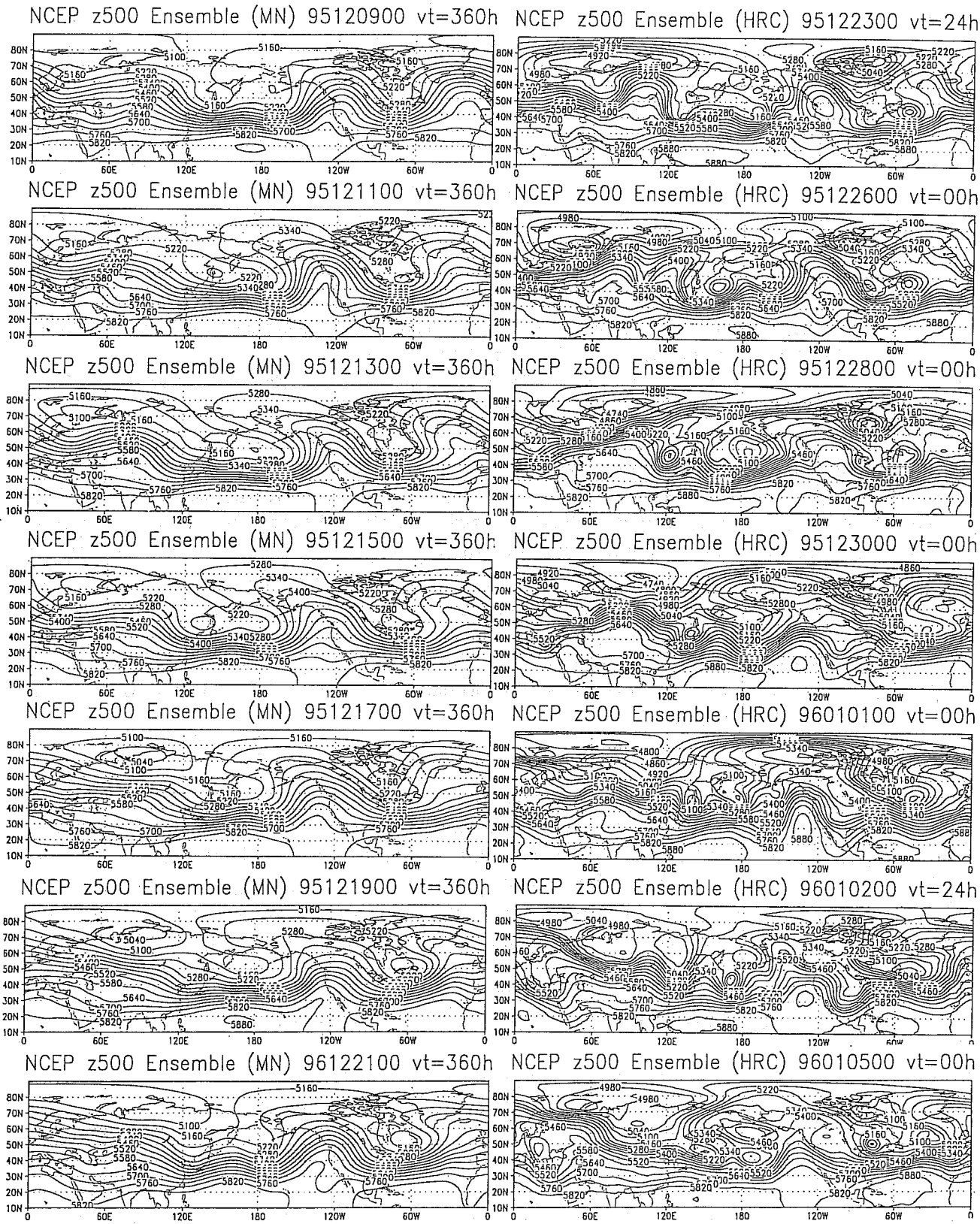

Fig. 12: 500 hPa geopotential height of the 17-member NCEP ensemble mean forecasts at 15 days lead time, initiated on every second day between 9 and 21 December 1995 (left), and corresponding analyses (right; for 1995/12/24 and 1996/01/03, a 24-hour forecast is shown). (Fig. 11 is placed on next page.)

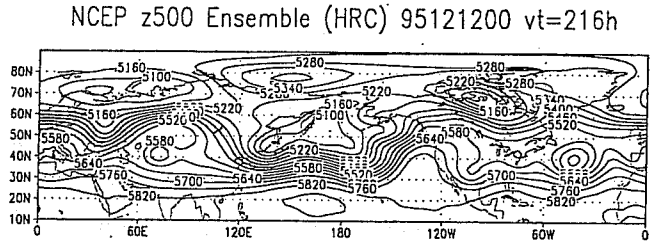

Fig. 11: 500 hPa geopotential height of the NCEP MRF T126 resolution control forecast at 9 days lead time, initiated on 12 December 1995.

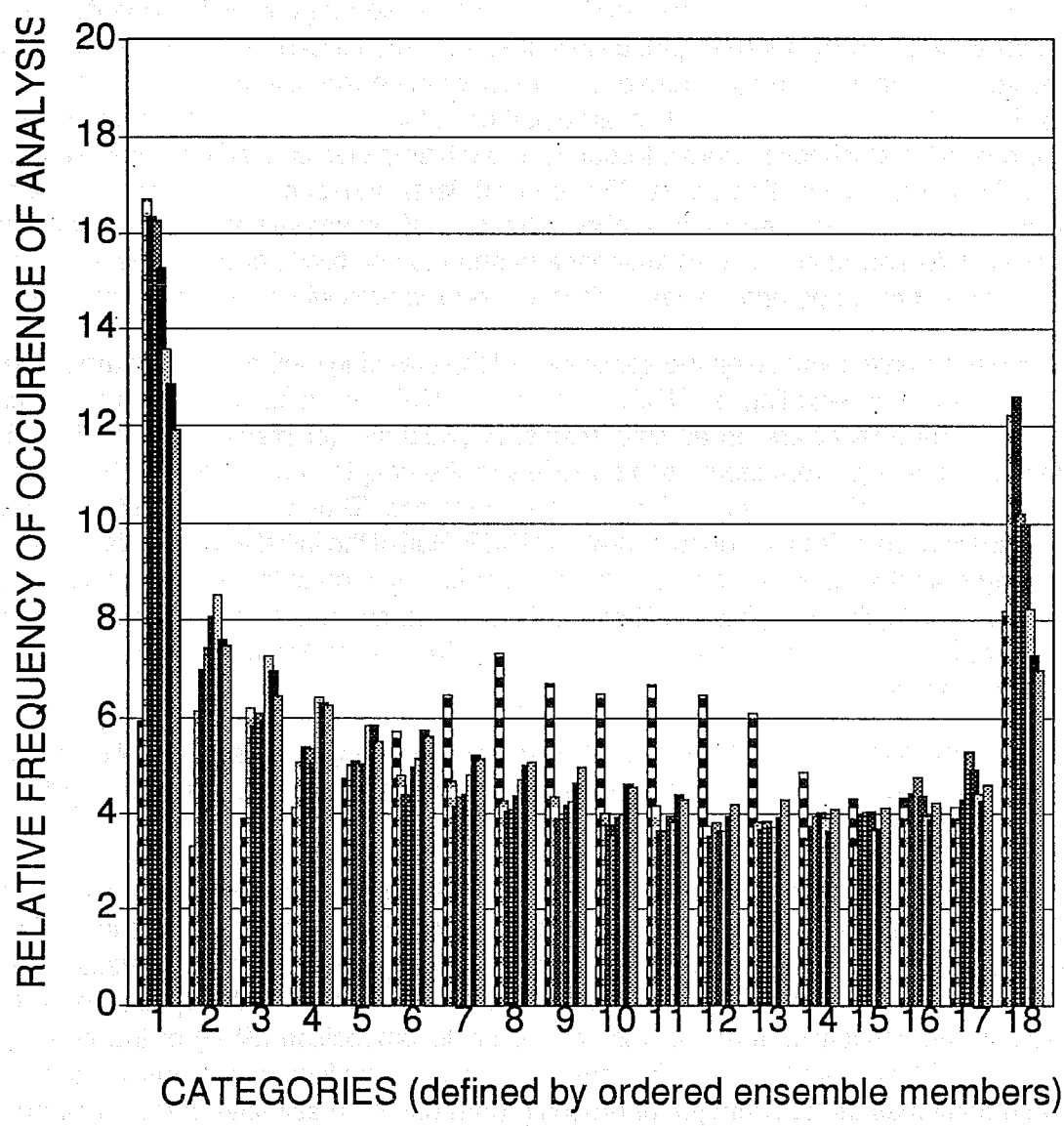

Fig. 13: Analysis rank ("Talagrand") distribution for the17-member NCEP ensemble for the 500 hPa height forecasts over the Northern Hemisphere extratropics, accumulated for 13 days between 9 and 22 December 1995. (Data from 18 December are missing.) Results are shown (from left to right within each group of bars) for 12-hour, 2-, 4-, 6-, 8-, 10-, 12-, and 14-day lead times. For further details see text.

may explain why the distribution is more U-shaped at intermediate lead times. By 14 days (at which lead time the forecasts are still skillful) the distribution becomes much flatter and only in 7 % of the cases does the ensemble miss the verification.

5. DISCUSSION

No consensus has been reached within the ensemble forecasting community regarding the optimal estimation and sampling of analysis uncertainty. Nor has there been any thorough study comparing the existing techniques in a realistic forecast environment. In this situation, the best approach perhaps is to combine ensemble forecasts generated at different numerical weather prediction centers. Harrison et al, 1995, for example, found that using models and/or control analyses from different centers yields in an ensemble spread that is closer to optimal.

Currently ensemble forecasts from all centers have an ensemble spread that is smaller than the size of the error in the ensemble mean forecast. The insufficient spread must be related to the use of models that differ from reality. Different solutions have been suggested to alleviate this problem. Houtekamer et al. (1996), for example, believe that the model used in ensemble forecasting has to be (perhaps drastically) changed from one member of the ensemble to the next. According to the suggestion of Kalnay and Toth (1994), a large portion of the errors that are due to imperfect models may be related to stochastic errors introduced at each time step caused by limited resolution and numerical and parametrization errors. These errors later project onto fast growing phase space directions, enlarging the forecast error above the level of ensemble spread. To account for these processes, Toth and Kalnay (1996) suggest a method to add bred perturbations to the ensemble forecasts *during the integration.* Again, this is a research area with no definite answers yet.

In this paper we have focused on the application of the breeding method for medium and extended range, ensemble forecasting at NCEP. However, NCEP is committed to provide ensemble guidance to the forecasters on all time scales. Applications (or plans thereof) of the breeding method are underway from storm-scale models to the coupled ocean-atmosphere model. For example, following the recommendations of a workshop (Brooks et al., 1995) a 15-member regional scale ensemble is run once a week at NCEP. Half of the initial and boundary perturbations are coming from the global ensemble while the rest is based on different in-house analyses. For these experiments, the ETA (Black, 1994) and the regional spectral models (Juang and Kanamitsu (1994) are used, thus providing an estimate of the variability caused by different model formulations as well.

As mentioned in Kalnay and Toth (1996, in this volume) breeding can also be applied in the context of climate forecasts, in the framework of the NCEP coupled ocean-atmosphere model (Ji et al., 1994).

Finally we would like to mention the potential applicability of the bred global ensemble at NCEP in targeting observations. The goal here is to find the "source"area from which the initial (or 12 hour lead) errors will most impact the forecast errors at 3-5 days later, over a preselected verification region. Additional observations in this source area may improve the quality of subsequent forecasts in the verification area. A singular vector decomposition (SVD) of the bred ensemble perturbations (Bishop and Toth, 1996) offers an easy to implement solution to targeting. This approach combines the advantages of breeding (that offers an estimate of fast growing analysis errors) and SVD that allows for regionalization of the perturbations. Preliminary results are encouraging and we hope that the method can be first applied in the winter of 1996/97, during the FASTEX experiments in which the goal is to study and better predict smaller scale cyclonic waves in the Eastern Atlantic.

Acknowledgements. The verification results are based on the computations of Mr. Y. Zhu, and were supplemented by Mr. G. Iyengar. Figures 10–12 were kindly prepared by T. Marchok.

REFERENCES:

Anderson, J. L., 1993: The Climatology of Blocking in a Numerical Forecast Model *J. Climate*, **6**, 1041–1056.

Anderson, J. L., 1995: Selection of initial conditions for ensemble forecasts in a simple perfect model framework. *JAS*, in press.

Bishop, C., and Z. Toth, 1996: Using ensembles to identify observations likely to improve forecasts. Preprints of the 11th AMS Conference on Numerical Weather Prediction, 19–23 August 1996, Norfolk, Virginia.

Black, T. L., 1994: The new NMC msoscale eta model: description and forecast examples. *Wea. Forecasting*, **9**, 265–278.

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. Meteorol. Soc.*, **76**, 1617–1624.

Buizza, R., J. Tribbia, F. Molteni, and T. Palmer, 1993: Computation of optimal unstable structures for a numerical weather prediction model. *Tellus*, **45A**, 388–407.

Chen, Y.-Q., D. S. Battisti, T. N. Palmer, J. Barsugli, and E. S. Sarachik, 1995: A study of the predictability of tropical Pacific SST in a coupled atmosphere/ocean modfel using singular vector analysis: The role of the annual cycle and the ENSO cycle. *Mon. Wea. Rev.*, under review.

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.

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

Farrel, B., 1988: Optimal excitation of neutral Rossby waves. *J. Atmos. Sci.*, **45**, 163–172.

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. Technical Report No. 153, UK Met. Office. [Available from: Forecasting Research Division, Meteorological Office, London Road, Bracknell, Berkshire RG12 2SZ, UK.]

Houtekamer, P. L., 1995: The construction of optimal perturbations. *Mon. Wea. Rev.*, in press.

Houtekamer, P. L., L. Lefaivre, J. Derome, H. Ritchie, and H. L. Mitchell, 1996: A system simulation approach to ensemble prediction. *Mon. Wea. Rev.*, in press.

Ji, M., A. Kumar, and A. Leetmaa, 1994: An experimental coupled forecast system at the National Meteorological Center. Some early results. *Tellus*, **46A**, 398–418.

Juang, H.-M. H., and M. Kanamitsu, 1994: The NMC nested regional spectral model. *Mon. Wea. Rev.*, **122**, 3–26.

Kalnay, E., and Z. Toth, 1994: Ensemble forecasting with imperfect models. WGNE Research Report, 1994.

Kalnay, E., and Z. Toth, 1996: The breeding method. Proceedings of the ECMWF Seminar on Predictability. September 4–8, 1995, Reading, England, in press.

Lacarra, J. F, and O. Talagrand, 1988: Short range evolution of small perturbations in a barotropic model. *Tellus*, **40A**, 81–95.

Legras, B., and R. Vautard, 1996: A guide to Lyapunov vectors. Proceedings of the ECMWF Seminar on Predictability. September 4–8, 1995, Reading, England, in press.

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

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

Lorenz, E. N., 1969: Three approaches to atmospheric predictability. *Bull. Amer. Meteorol. Soc.*, **50**, 345–349.

Molteni, F., R. Buizza, T. N. Palmer, and T. Petroliagis, 1995: The ECMWF ensemble system: Methodology and validation. *Q. J. R. Meteorol. Soc.*, in press.

Oortwijn,, J, and J. Barkmeijer, 1995: Perturbations which optimally trigger weather regimes. *J. Atmos. Sci.*, in press.

Palmer, T., F. Molteni, R. Mureau, R. Buizza, P. Chapelet, and J. Tribbia, 1992: Ensemble prediction. ECMWF Research Department Tech. Memo. No. 188. (Available from ECMWF, Reading, England.)

Szunyogh, I., E. Kalnay, and Z. Toth, 1995: A comparison of Lyapunov vectors and optimal vectors in a low resolution GCM. Tellus, under review.

Toth, Z., 1991: Estimation of atmospheric predictability by circulation analogs. *Mon. Wea. Rev.*, **119**, 65–72.

Toth, Z., 1993: Preferred and unpreferred circulation types in the Northern Hemisphere wintertime phase space. *J. Atmos. Sci.*, **50**, 2868–2888.

Toth, Z., and E. Kalnay, 1993: Ensemble Forecasting at the NMC: The generation of perturbations. *Bull. Amer. Meteorol. Soc.*, **74**, 2317–2330.

Toth, Z., and E. Kalnay, 1996: Ensemble forecasting at NMC and the breeding method. *Mon. Wea. Rev.*, under review.

Toth, Z., E. Kalnay, S. Tracton, R. Wobus, and J. Irwin, 1996: A synoptic evaluation of the NCEP ensemble. Proceedings of the Fifth ECMWF Workshop on Meteorological Operational Systems. November 13–17, 1995, Reading, England, in press.

Toth, Z., I. Szunyogh, and E. Kalnay, 1996: Singular, Lyapunov and bred vectors in ensemble forecasting. Preprints of the 11th AMS Conference on Numerical Weather Prediction, 19–23 August 1996, Norfolk, Virginia.

Tracton, M. S., 1994: Operational ensemble prediction – The NMC experience. Preprints of the Tenth Conference on Numerical Weather Prediction, July 18–22, 1994, Portland, Oregon, p. 206–208. Available from: AMS, 45 Beacon Str., Boston, MA 02108-3693.

Tracton, M. S. and E. Kalnay, 1993: Ensemble forecasting at NMC: Operational implementation. *Wea. Forecasting*, **8**, 379–398.

Trevisan, A., and R. Legnani, 1995: Transient error growth and local predictability: a study in the Lorenz system. *Tellus*, **A47**, 103–117.

Tsonis, A. A., 1992: Chaos: From theory to applications. Plenum Press, New York, pp. 83–96.

Xue, Y., M. A. Cane, and S. E. Zebiak, 1996: Predictability of ENSO using singular vector analysis. Part I: Optimal growth in seasonal background and ENSO cycles. Mon. Wea. Rev., under review.