Ensemble Perturbations and Forecast Errors

Mozheng Wei *

UCAR VSP, NCEP Environmental Modeling Center, Maryland

Zoltan Toth

SAIC at NCEP Environmental Modeling Center, Maryland

May 24, 2002

Submitted to Monthly Weather Review

*Corresponding author: Mozheng Wei, NCEP Environmental Modeling Center, 5200 Auth Road, Camp Springs, Maryland, 20746. E-mail: Mozheng.Wei@noaa.gov

ABSTRACT

Most ensemble forecast verification statistics are influenced by the quality of not only the ensemble generation scheme, but also the analysis scheme and forecast model. In this study, a new tool called Perturbation vs. Error Correlation Analysis (PECA) is introduced that lessens the influence of the latter two factors. PECA evaluates the ensemble perturbations, instead of the forecasts themselves, by measuring their ability to explain forecast error variance. As such, PECA offers a more appropriate tool for the comparison of ensembles generated by using different analysis schemes and models.

Ensemble perturbations from both NCEP and ECMWF were evaluated and found to perform similarly. The error variance explained by either ensemble increases with the number of members and the lead time. The dynamically conditioned NCEP and ECMWF perturbations outperform both randomly chosen perturbations and differences between lagged forecasts ("NMC" method). Therefore ensemble forecasts potentially could be used to construct flow dependent short-range forecast error covariance matrices for use in data assimilation schemes.

It is well understood that in a perfect ensemble the spread of ensemble members around the ensemble mean forecast equals the rms error of the mean. Adequate rms spread, however, does not guarantee sufficient variability among the ensemble forecast patterns. A comparison between PECA values and Pattern Anomaly Correlation (PAC) values among the ensemble members reveals that the perturbations in the NCEP ensemble exhibit too much similarity, especially on the smaller scales. Hence a regional orthogonalization of the perturbations may improve ensemble performance.

1 Introduction

There exist a large number of verification tools for the evaluation of ensemble forecasts (see, e. g., Stanski et al. 1989). One type of measures evaluates the performance of a summary indicator of a set of forecasts such as the mean or the median value of the ensemble distribution. Typically, the root mean square (RMS) error and/or the pattern anomaly correlation (PAC) is used for this purpose. A second set of measures evaluates probability distributions based on the ensemble forecasts. Such measures include, for example, the Brier Skill Score (BSS), and the Ranked Probability Skill Score (RPSS) that measure the reliability (statistical consistency with observations) and resolution (how different reliable forecast probability values are from the climatological distribution). A third set of related measures assesses the utility of the forecasts from a user's point of view. Related measures include the Relative Operating Characteristics (ROC) and the economic value of forecasts (both of which are related to resolution).

When a set of ensemble forecasts are evaluated through any of the above scores, the results will reflect the combined effect of the quality of (i) the analysis field around which the initial ensemble is centered; (ii) the forecast model(s) that is used for integrating the ensemble forecasts; and (iii) the way the initial ensemble perturbations are formed.

In the present paper we propose a new ensemble evaluation method that is less sensitive to the first two aspects of ensemble performance and measures more directly the effect of initial perturbations on ensemble performance. The proposed method, called Perturbation vs. Error Correlation Analysis (PECA), is based on the comparison of ensemble forecast perturbations (ensemble forecasts minus control) and forecast error patterns (control forecast minus verifying analysis).

The motivation for the development of such an ensemble evaluation measure is two-fold.

First, a tool that is less sensitive to differences in the quality of the analysis and forecast schemes used to generate the ensembles can be more readily applied in studies that compare ensemble forecasts generated by different NWP forecast centers, using various analysis schemes and models. Second, such a tool can be used to evaluate whether ensemble members are correlated with each other over various size areas at the proper level, i. e., at the level at which the error correlates with ensemble perturbations. Such an analysis in fact amounts to measuring the spread within the ensemble but in a manner different from earlier studies where spread is defined pointwise, in an rms sense. Here "pattern spread" is evaluated over predefined regions, revealing an aspect of ensemble perturbations that has not been thoroughly investigated before.

Here we refer to an earlier study by Molteni and Buizza (1999), based on an EOF analysis of ensemble perturbations. The method involves the comparison of ensemble perturbations and error patterns. This analysis, however, is carried out in a statistical sense only cases, in terms of a comparison of the perturbations and error EOF spectra. Therefore, unlike the method proposed here (PECA), it means only the statistical consistency, but not the forecast skill of ensemble system.

After a description of the proposed method and its properties in section 2, PECA results are presented for the National Centers for Environmental Prediction (NCEP) and the European Centre for Medium Range Weather Forecasts (ECMWF) ensemble forecast systems over different areas of the globe for a selected variable (500 hPa geopotential height) and for total energy in section 3. Also ensemble perturbations are compared with "NMC"-type perturbations (based on differences between short range forecasts valid at the same time, see Parrish and Derber, 1992). For a comparison, PECA results are also presented for randomly selected ensemble perturbations, and for "perfect" ensembles where the verifying analysis is taken as one of the ensemble members. These results are presented in section 4. Some conclusions are

offered in section 5.

2 Methodology

a. Description

Ensemble perturbations, at either numerical weather prediction (NWP) centers, are defined as the differences between the perturbed forecasts and their respective control forecast (started from an unperturbed analysis):

$$\mathbf{P}_i(t) = \mathbf{F}_i^C(t) - \mathbf{F}_{control}^C(t), \tag{1}$$

where C, the originating center, is either NCEP or ECMWF, and $i = 1, 2, \ldots, N$, with $N = 20$ for NCEP and 50 for ECMWF.

Note that the NCEP ensemble perturbations, at 24-hour lead time are, by definition, the bred vectors which, after rescaling are used as perturbations to initiate the next set of ensemble (Toth and Kalnay, 1997). The ECMWF initial perturbations are combinations of initial and evolved singular vectors (Buizza and Palmer, 1995; Molteni et al. 1996; Barkmeijer et al. 1999). At the time of writing, products consist of 10 ensemble forecasts both at 0000 UTC and 1200 UTC from NCEP, and a 50-member ensemble at 1200 UTC while ECMWF each day.

In both systems, forecast errors $\mathbf{E}(t)$ are defined as the difference between the control forecast and the verifying analysis from the same center,

$$\mathbf{E}(t) = \mathbf{F}_{control}^C(t) - \mathbf{F}_{analysis}^C(t). \tag{2}$$

A posteriori optimal, combination of n perturbations is obtained by solving the least-square problem

$$Min|\mathbf{E} - \sum_{i=1}^{n} \alpha_i \mathbf{P}_i|_{L2} \tag{3}$$

Having obtained α_i from the above equation, the optimally combined vector is defined as

$$\mathbf{P}_{optimal} = \sum_{i=1}^{n} \alpha_i \mathbf{P}_i. \tag{4}$$

Note that the optimal combinations as defined in (3) and (4) are based on actual error patterns. Therefore unlike the weighted ensemble mean of van den Dool and Rukhovets (1994) that is based on past statistics, the optimal perturbations can only be used a diagnostic value (and not as a prognostic tool).

The pattern anomaly correlation (PAC) between any two vectors $\mathbf{X}$ and $\mathbf{Y}$ is defined by

$$A_c(\mathbf{X}, \mathbf{Y}) = \frac{\{\mathbf{X}, \mathbf{Y}\}}{\{\mathbf{X}, \mathbf{X}\}^{\frac{1}{2}} \{\mathbf{Y}, \mathbf{Y}\}^{\frac{1}{2}}}. \tag{5}$$

PECA is defined as the PAC between $\mathbf{X} = \mathbf{P}_i$ (or $\mathbf{X} = \mathbf{P}_{optimal}$) and $\mathbf{Y} = \mathbf{E}$. Note that the square of the correlation, A_c^2, can be considered as explained error variance. In this study, PECA will be computed for different regions. In addition to the global domain, results will also be shown for the Northern (NH, 20N-77.5N) and Southern Hemisphere extratropics (SH, 20S-77.5S), the Tropics (20N-20S), and North America (NA, 140W-50W, 20N-60N) and Europe (EU, 20W-40E, 77.5N-30N). Correlation values computed between individual perturbations ($\mathbf{P_i}$) and the forecast errors ($\mathbf{E}$) will be averaged over the 10 individual perturbations in most cases studied. In addition, correlation values between $\mathbf{P}_{optimal}$ and $\mathbf{E}(t)$ will also be computed

for the same domains and forecast lead times.

b. Properties

All verification scores listed in the Introduction compare a forecast ensemble to the verifying analysis. The scores therefore reflect, beyond the quality of the ensemble generation technique, also that of the initial analysis around which the ensemble is centered, and the model that is used for integrating the ensemble forecasts. In contrast, PECA attempts to evaluate the quality of *ensemble perturbations*. This is achieved by measuring the amount of variance that individual and/or optimally combined ensemble perturbations can explain in forecast error fields. The higher the PECA values are, the more successful an ensemble is in achieving its goal of capturing forecast errors. PECA values are not directly influenced by how large the forecast errors are but rather by the ability of the ensemble perturbations to explain the forecast error.

The above point will be illustrated through a comparison of PECA with PAC as it is traditionally applied in Forecast Verification (PACFV). PACFV is defined as the PAC between $(\mathbf{F}_{control} - \mathbf{C})$ and $(\mathbf{X}_{analysis} - \mathbf{C})$, where $\mathbf{C}$ is the climate mean. Note that while PACFV compares a forecast anomaly from the climate mean with the verifying analysis anomaly from the climate mean, PECA compares the pattern of ensemble perturbations (perturbed minus control forecast) to that of the forecast error (control forecast minus analysis). While PACFV evaluates the overall quality of the forecast anomaly with respect to the climate mean, PECA focuses on the quality of the ensemble perturbations defined with respect to the control forecast. The fact that the anomalies defined by PECA are taken from the control forecast (and not from the climate mean) eliminates, to a large extent, the effect the quality of the initial analysis has on the verification measure.

The effect of differences in the quality of models used to generate two ensemble systems to be compared is also expected to be reduced when using PECA. Model differences, however, are expected to exert some influence on PECA. In particular, if a model, due to some imperfection, is not able to reproduce an instability that is present in nature, an ensemble generated by that model will not be able to capture forecast errors associated with that kind of instability. Alternatively, if a model exhibits a spurious instability that is not present either in nature or in another model, forecast errors associated with that model will not be captured by an ensemble generated by the other, more realistic model. Indications for both types of model imperfection related problems will discussed in the next section.

When comparing ensembles generated by different NWP centers, higher PECA values are thus indicative of an ensemble of higher quality, that can better explain its own center's forecast errors (however small or large they are). A superior ensemble in terms of PECA values may show inferior performance in terms of traditional measures like PACFV or probabilistic scores, when a NWP center's analysis and/or model performance is poorer than that of the others.

Beyond comparing the performance of ensembles generated by different NWP centers, PECA can also be used to evaluate an ensemble's performance in terms of the value of correlation among its members. In a perfect (reliable, statistically consistent) ensemble, the verifying analysis is indistinguishable from the ensemble forecasts. It follows that PECA values computed by substituting the actual error field by one of the ensemble perturbations (perfect model/ensemble assumption, PECA perfect) should be at the same level as those computed using the actual error field. Any discrepancy can be interpreted as a deficiency in the ensemble generation technique (lack of proper representation of initial and/or model related uncertainty). The PECA values computed in a perfect model/ensemble environment measure how similar perturbation patterns are over a selected geographical domain, hence the use of the term "pattern

spread".

If the PECA values in the perfect model/ensemble case are, for example, higher than those computed with real error fields, that is an indication for an under-dispersive ensemble. In such an ensemble the perturbation patterns show a lack of variability for properly explaining forecast error patterns. A careful comparison of perfect model/ensemble and regular PECA values computed over various size domains can provide quantitative guidance on whether the diversity of perturbation patterns is adequate or needs to be improved in an ensemble. Note that pattern spread, as will be shown in the next section, can be insufficient even if the rms spread, computed and averaged over individual grid points, is large enough. While the rms spread of an ensemble can be easily changed by multiplying the initial perturbations by a scalar number, PECA (pattern spread) is not affected by such a change. The pattern spread can only be changed through the introduction of more diversity in the initial ensemble perturbation *patterns*. The apparent difference between rms spread and pattern spread indicates that PECA evaluates an aspect of ensemble performance that has not been previously addressed.

3 NECP and ECMWF ensemble results

a. A case study

Before a statistical analysis of PECA results accumulated over a 30-day period is presented in the following subsections, two examples for the application of the proposed method over the North American region are shown below.

Fig. 1a shows NCEP 500 hPa geopotential height analysis field valid at 12 UTC May 8, 2001. The analysis field shows a wave-like structure across the higher latitudes of NA. The

corresponding error pattern of an 8-day lead time forecast initialized at 1200 UTC April 30, 2001, displayed in Fig. 1b, is dominated by a dipole pattern over eastern Canada. In this example, even the best combination of the NCEP ensemble perturbations (Fig. 1c), computed a posteriori based on the actual forecast error, fails to explain much of the actual error. The variance explained by the optimal ensemble perturbation is below 40%, compared with 69% explained variance for 8-day forecast errors averaged over the experimental 30-day period. A large part (60%) of the error remains unexplained by the ensemble. This is also evident from Fig. 1d that displays the residual error $\mathbf{R}(t)$ (Fig. 1d), defined as the part of the forecast error that cannot be explained by $\mathbf{P}_{optimal}(t)$, i.e. $\mathbf{R}(t) = \mathbf{E}(t) - \mathbf{P}_{optimal}(t)$. The magnitude of the optimal perturbation displayed was computed by projecting the forecast error $\mathbf{E}(t)$ onto the corresponding optimally combined ensemble perturbation defined by eq. (4).

The poor PECA performance of the ensemble initialized at 1200 April 30 2001 at 8-day lead time (Fig. 1) is in contrast with that at 10-day lead time, shown in Fig. 2. The corresponding verifying analysis field (1200 UTC May 10 2001, Fig. 2a) is dominated by a predominantly zonal flow. In this case the error field (Fig. 2b) is characterized by a wave-train type pattern north of 40N. The optimal perturbation (Fig. 2c) successfully explains the error pattern, including most of the details. 94% of the error variance is explained (compared with an average of 71%), leaving only a small fraction (6%) of the total error field unexplained (Fig. 2d).

b. Error variance explained by respective ensembles

Figs. 3 a-f show the PECA correlation values computed between the NCEP (solid lines) and ECMWF (dotted) forecast error and the corresponding ensemble perturbations for 500hPa geopotential height over the global, Northern and Southern Hemisphere, Tropical, North American and European domains respectively as a function of forecast lead time. Geopotential height

at 500hPa is one of the variables with the least amount of systematic error, thus the correlation values will reflect the ensemble's ability to explain initial value related forecast errors. The PECA values shown in Fig. 3 are averages over a 30-day period started at 1200 UTC April 1, 2001. The thin lines in Fig. 3 represent PECA values averaged over 10 individual ensemble perturbations, while the corresponding thick lines represent the PECA values for the optimally combined vectors (see eq. (4)).

As expected, the optimally combined perturbation vectors (thick lines) can explain a much larger portion of the forecast error than the individual perturbations (thin lines) at all lead times and over each domain for both forecast systems. Note that over smaller areas (NA and EU), both ensembles can explain a larger amount of forecast error variance. This is due to the fewer degrees of freedom over these smaller areas. This also explains the larger sampling fluctuations (noise) observed in the results valid for smaller geographical areas.

For individual perturbations, the NCEP ensemble performs better out to 7 days lead times (after which the correlation values for the two systems become similar) over all domains except the tropics. Over the tropics, the NCEP/ECMWF ensemble shows superior performance before/after 3 days lead time. When the systems are compared using the optimal perturbations, the NCEP ensemble exhibits higher correlations up to 2-3 days lead time, after which the two ensembles perform rather similarly. We note that the initial ensemble perturbations likely play a more important role at short lead times (0-5 days), while model related errors, in a relative sense, may influence on the results more at longer (6 to 10 days) lead times. In particular, the presence of model bias that would not be well explained by the dynamically evolving ensemble perturbations, is expected to lead to lower PECA values.

Note that until January 2002, ECMWF ensemble system had no initial perturbations in the tropics. The recent introduction of targeted tropical singular vectors (Barkmeijer et al 2001) is

expected to improve performance of the ECMWF ensemble in the vicinity of tropical areas.

Another important observation is that both the individual and the optimal vectors can explain an increasing amount of error variance as the lead time increases. This can be explained by a collapse of the phase space containing possible forecast errors into a smaller dimensional subspace due to 2 factors. Perturbations, including the error fields that evolve quasi linearly, are attracted to the fastest growing nonlinear perturbations (Toth and Kalnay, 1993, 1997), that are related to the leading Lyapunov vectors (LVs) (e.g. Buizza and Palmer 1995; Szunyogh et al. 1997; Reynolds and Errico 1999; Wei 2000; Wei and Frederiksen 2002). This may affect the rapid increase of correlation values in the first 5 days. Second, when nonlinearities become dominant, the error (and perturbation) fields become dominated by larger scales, leading to a further collapse of the error subspace. This process may explain the slower increase in PECA values beyond 5-day lead time.

c. Explained error variance with swapped ensembles.

To gain a better understanding of the relative role of ensemble perturbations and model errors, here we discuss PECA results where the first 10 NCEP perturbations (NPs) are used to explain the ECMWF forecast errors, and the first 10 ECMWF perturbations (EPs) to explain the NCEP forecast errors. The results are shown as dashed and dash-dotted lines respectively in Fig. 3. For short lead times of up to a few days, the optimally combined NPs can explain the ECMWF forecast errors slightly better than the optimally combined EPs can explain the NCEP forecast errors over the global and NH domains. After that the optimally combined EPs have a slight advantage. Over the Southern Hemisphere and the tropics however, the optimally combined EPs can explain the NCEP forecast errors better than the optimally combined NPs can explain the ECMWF forecast errors.

While at very short lead times, the results are similar to the "unswapped" cases discussed above, the individual and optimal swapped perturbations display a much reduced ability to explain the other center's forecast errors at longer lead times over the large domains. At 10-day lead time for the global domain, for example, the NCEP ensemble can explain about 40% variance in the NCEP control forecast error whereas only 10% in the ECMWF forecast error. Over the smaller North American and European areas, however, no such large discrepancy is present (see Fig. 3e-f).

This suggests that at longer lead times, the error fields have a strong large scale model specific component that only an ensemble generated via the same model can capture. This error may be due to some unrealistic or *spurious* instabilities that are specific to each model, but are not present in nature. The unstable structures that appear in the error fields will appear only in perturbations generated by the same model.

The fact that the NCEP ensemble performance shows more degradation than the ECMWF ensemble when used to explain errors in forecasts from the other center suggests that it may be more affected by the presence of spurious large scale instabilities. This result is consistent with that of Saha (2001) who, using a technique developed by van den Dool et al. (2000), found that ECMWF forecasts contain less large scale systematic error than NCEP forecasts do.

Interestingly over several domains, the inclusion of a few ECMWF members with the NCEP ensemble leads to higher explained variances for the NCEP forecast errors (dash dot dotted lines) at intermediate lead times. At 7-day or longer lead times, however, such a mixed ensemble performs worse than a pure NCEP ensemble. The inclusion of NCEP perturbations improves (degrades) the ECMWF ensemble performance before (after) 3 days lead times.

d. The effect of ensemble size.

To the extent ensemble perturbations are independent, optimally combining more ensemble members is expected to lead to higher explained error variances (compare thick and thin lines in Fig. 3). In this subsection, we explore the effect of increased ensemble membership in more detail.

Fig. 4 shows the PECA values between various number of optimally combined NPs and EPs and the respective NCEP and ECMWF forecast error. The results are displayed for various lead times (1, 2, 3, 5 and 7 days). The results from NCEP and ECMWF are indicated by thick and thin lines respectively.

While the ECMWF ensemble has 50 members initiated at 1200 UTC, NCEP has only 10 members at both 0 and 1200 UTC. To study the effect of a larger ensemble for the NCEP system, we combined the 1200 UTC and the subsequent 0000 UTC NCEP ensembles. The choice for the use of the subsequent (and not preceding) ensemble was motivated by the fact that the preceding, longer lead time ensembles would have led to higher correlations (see Fig. 3).

As expected, increasing the number of ensemble perturbations increases the correlations between the forecast errors and the optimally combined perturbations for both centers. For the global domain (Fig. 4a), for lead times up to 5 days (thick and thin dotted lines respectively), any available number of optimally combined NPs can explain a slightly larger percentage of forecast error than the same number of optimally combined EPs can. At 7-day lead time, the situation is reversed in that the ECMWF perturbations become more effective in explaining forecast errors (thick and thin long dashed lines).

Results over the Northern and Southern Hemispheres are similar to those over the global domain, while the NCEP optimally combined ensemble performs better at all lead times in the tropics. Over the smaller North American and European domains the advantage of the

NCEP ensemble is evident only at 1 and 2-day lead times. For example, the NCEP ensemble can explain a similar amount of variance in the 1-day error field as the ECMWF ensemble can in the 2-day error fields. ECMWF PECA values are very close to and sometimes higher than those for NCEP at 3 days or longer lead times.

The PECA results presented in Fig. 4 are shown as a function of both number of members and lead time in more detail in Figs. 5 and 6. When comparing the results from the NCEP and ECMWF ensembles, one should keep in mind that the NCEP ensemble has 20 members, while the ECMWF has 50. An additional difference is that the maximum lead time for NCEP and ECMWF ensembles are 16 and 10 days respectively.

As discussed earlier, the PECA values are the lowest over the largest, global domain, while highest over the smallest, EU domain. This indicates that ensemble perturbations can explain much more forecast error variance over smaller domains than over larger ones. A regionally optimized ensemble would improve the forecast capability greatly.

It is interesting to note in Figs. 6 (e) and (f) that at short lead times and over smaller areas, an increase in ensemble membership brings significant improvement over beyond 25-30 members. This is not so at larger lead times when the error fields, on average are, are rather well explained even by smaller ensembles.

e. Comparison with lagged forecast difference fields.

Parrish and Derber (1992) proposed to use a set of difference fields taken between 1 and 2-day forecasts verifying on the same day, in the construction of forecast error covariance matrices. Forecast difference fields between 24 and 48 hour lead time were generated for both the NCEP and ECMWF control forecasts over a preceding period (1200 UTC March 5, 2001 to 1200 UTC April 3, 2001) for explaining 1-day forecast error fields. In our experiment, we computed 30

consecutive vector fields, i.e.

$$\mathbf{F}_{NMC}(t_i) = \mathbf{F}^{2-day}_{control}(t_i) - \mathbf{F}^{1-day}_{control}(t_i), \tag{6}$$

$i = 1, 2, \ldots, 30$ for both NCEP and ECMWF.

The procedure, called "NMC method" (NCEP was formerly called NMC - National Meteorological Center), has been widely used in data assimilation schemes worldwide. In order to determine if the use of flow dependent ensemble perturbations will provide better error covariance information than a fixed set of short range forecast difference fields, the "NMC" perturbation vectors were evaluated in a fashion similar to the ensemble perturbations.

The "NMC" method assumes that difference fields between different lead time forecasts valid at the same time can be used to describe forecast error statistics. It is clear that the correlation between the NCEP NMC vectors and NCEP forecast error (thick solid lines) is generally higher than that between the ECMWF NMC vectors and ECMWF forecast error (thin solid lines).

For most domains, both NPs and EPs can better explain their respective 1-day forecast error than the "NMC" perturbations (compare dotted and solid lines in Fig. 4). The tropics is the only domain where the optimally combined ECMWF NMC vectors perform better than the optimally combined EPs which is probably due to the lack of initial perturbation in the tropics in the ECMWF ensemble during this period of time (thin solid and dotted lines in Fig. 4 (d)). As we mentioned above, the introduction of TSVs on Jan 22, 2002 in the ECMWF EPS is expected to improve its performance in the tropics. Note that the NCEP NMC vectors exhibit clearly higher correlation with NCEP forecast error fields than ECMWF NMC vectors with their forecast error fields (except for the Southern Hemisphere domain). The difference is

more pronounced for smaller domains.

We also note that forecast error covariance matrices currently used in the ECMWF data assimilation scheme are not computed by using the "NMC method". Instead, they are based on an ensemble of analyses generated by running data assimilations cycles with perturbed observations (Houtekamer et al. 1996; Buizza and Palmer 1999). Our results suggest that the construction of ensemble-based forecast error covariance matrices in place of the "NMC method" may in general improve the performance of data assimilation schemes.

f. 3-dimensional error structure.

So far, results have been presented for one variable at one level only (500 hPa geopotential height). In this section, we analyze the ability of the ensemble perturbations to capture forecast error fields defined by 3 variables, temperature (T) and velocity (U, V) at 3 levels. Based on data at 850hPa, 500hPa and 250hPa (200hPa for ECMWF), we define a new variable p: $p = [U, V, \alpha T]$, where $\alpha = \sqrt{C_p/T_r}$, $C_p = 1004.0 Jkg^{-1}K^{-1}$ is the specific heat at constant pressure for dry air (Holton 1992) and T_r is a reference temperature. For each pressure level T_r is obtained by linear interpolation from the US standard atmospheric data in NOAA/NASA/USAF (1976). Thus the inner product $< p, p >$ has the form of total energy as defined by Rabier et al. (1996) and Barkmeijer et al. (1999). The dimension of p is $9m$, where m is the number of grid points over a given domain.

The results corresponding to the use of variable p are presented in Fig. 7 in a manner similar to Fig. 3. First note that due to an increase in the degrees of freedom, correlation values in Fig. 7 are considerably lower than in Fig. 3. While the NCEP ensemble performed better than the ECMWF in case of one variable at one level, the ECMWF ensemble becomes more efficient when the multi-level/multi-variable error fields are considered. This is true especially at short

lead times, and over the smaller Northern American and European domains.

The explanation, again, is not clear but one may speculate that the vertical and /or cross-variable structure of the ECMWF model may be more realistic than that of the NCEP model. Note that the ECMWF ensemble is run at a higher vertical and horizontal resolution (T255L40) than the NCEP ensemble (T126L28 for first 3.5 days, T62L28 thereafter). This explanation is supported by the results of Richardson (2001, personal communication) who found that when started from the same analysis the ECMWF forecast model produces higher quality forecasts than the NCEP model.

Beyond 2-3 days lead time, the NCEP individual multi-layer/multi-variable perturbations perform better than the ECMWF perturbations over the larger domains (global, NH, SH and tropics). For short lead time, ECMWF ensemble forecasts gain more from optimally combined ensembles, presumably due to the fact that they are orthogonalized at initial time while the NCEP perturbations are not. With these gains the ECMWF optimal perturbations perform better than the NCEP perturbations over the Northern Hemisphere, and the North American and European regions while NCEP performs better over the Tropics. The two systems perform similarly over the global, SH and European domains. The largest differences are observed over the tropics where the dynamically conditioned NCEP perturbations apparently have a large advantage over ECMWF perturbations that are purely stochastic in this area.

4 Perfect and random ensembles results

In the above section, the ability of ensemble perturbations in explaining forecast errors was investigated. To place the results in a broader context, here we will study how well random (lower bound of skill) or perfect (upper bound of skill) perturbations compare with the above

results. Only NCEP ensembles will be used in the experiments below.

The dotted lines shown in Fig. 8 are identical to the solid lines in Fig. 3, except here only 8 perturbations are combined optimally to estimate the actual skill of the NCEP ensemble. To estimate a lower bound for skill, ensemble perturbations that are valid 8 days earlier than the forecast error, are used. These "random" perturbations have the same statistical characteristics as the appropriate ensemble perturbations used above, but have no (or little) dynamically relevant information. The results for this random case using 8 perturbations are presented as the dashed lines in Fig. 8. An important result is that while at very short lead time, the random and actual perturbations perform rather similarly, once the errors become dynamically more organized only the actual perturbations can explain them well. This is true both for individual and optimally combined perturbations. The results indicate that the randomly chosen perturbations are dynamically not relevant and cannot explain flow dependent forecast errors.

If both the model and ensemble generation were perfect, the truth could be simulated by one of the ensemble members. Under a perfect model, perfect ensemble scenario, one of the ensemble members will be considered truth and the remaining 4 pairs of members will be used to explain the "error". In this case, $\mathbf{NP}_i(t) = \mathbf{F}_i^{NCEP}(t) - \mathbf{F}_{control}^{NCEP}(t)$, where $i = 1, 2, \ldots, 8$. The forecast error is defined as

$$\mathbf{E}_j(t) = \mathbf{F}_{control}^{NCEP}(t) - \mathbf{F}_j^{NCEP}(t), \tag{7}$$

where $j \neq i$. It is conceivable that the correlation between $\mathbf{E}_j(t)$ and $\mathbf{NP}_i(t)$ will be relatively high if $\mathbf{F}_j^{NCEP}(t)$ is correlated with $\mathbf{F}_i^{NCEP}(t)$ to certain extent. The results for this perfect model/ensemble case are shown as the solid lines in Fig. 8.

Note that the curve for the perfect case over the global domain runs more or less parallel to the actual PECA curve. This confirms that, as discussed earlier, the phase space of the forecast error undergoes a similar contraction as that of the ensemble perturbations. The fact that the perfect curve starts well above the actual curve, on the other hand, clearly indicates that the ensemble perturbations are too correlated with each other at initial time.

For the global domain, for example, the initial correlation value for the perfect case (thin solid line in Fig. 8a) is as high as the correlation values between the error and individual ensemble perturbations at around 3-day lead time (thin dotted lines in Fig. 8a). The problem is excerberated over the smaller areas. The results suggest that by imposing more diversity among the ensemble members on the smaller scales, the introduction of regional orthogonalization in the rescaling of the bred vectors (Toth and Kalnay 1997) may alleviate the problem and potentially lead to improved ensemble performance.

5 Discussion and Conclusions

One of the goals of ensemble forecasting is to generate a set of forecast scenarios that encompass truth. The success of ensemble forecasts can be measured by a number of ways. Most existing verification tools measure the overall skill of an ensemble forecast system. The verification results are strongly influenced by the quality of the analysis around which the ensemble is initialized, and the forecast model used. In this study a new metric is introduced that measures how well individual or optimally combined ensemble perturbations can explain forecast error variance (perturbation vs. error correlation analysis – PECA). This measure evaluates the performance of ensemble perturbations, and not the full forecast fields, and hence deemphasizes the effect of analysis and model differences on ensemble performance. The more closely ensemble

perturbations, on average, are correlated with forecast error, the better the ensemble represents truth.

Explained forecast error variance statistics were evaluated and compared for the bred vector based NCEP and singular vector based ECMWF ensembles. The main findings of this study are as follows:

1) The phase space of ensemble perturbations and that of forecast errors collapse into a smaller subspace with increasing lead time. This explains the higher correlation between ensemble perturbations and forecast errors at longer lead times.

The rotation of all linear perturbations toward the leading LLVs on one hand, and an upscale propagation of perturbation energy in the nonlinear phase on the other were called upon as possible explanations for this phenomenon. The typically enhanced performance of ensemble forecasts with increasing lead time (e. g., higher skill of ensemble mean forecast compared to a control forecast) is probably also related to this behavior. As Toth and Kalnay (1997) pointed out, ensemble averaging is effective in reducing errors only if the perturbations project on actual errors in the forecasts; otherwise it can even increase forecast errors.

2) The dynamically conditioned ensembles exhibit substantially more skill than randomly chosen perturbations with the same statistical characteristics. Moreover, the ensembles perform better than a set of lagged forecast differences (that are used at several NWP centers to construct forecast error covariance matrices in data assimilation schemes, the "NMC method") in explaining short range forecast errors. This indicates that ensembles could provide the basis for the construction of flow dependent error covariance matrices.

3) The NCEP and ECMWF ensembles generally exhibit a similar level of skill. The following, relatively minor differences were noted:

a) Individual NCEP perturbations were found more skillful in explaining errors in a single

variable (500 hPa geopotential height) over the first 5-7 days lead time. This may be an indication of more efficient initial ensemble perturbations in the NCEP ensemble.

b) The ECMWF ensemble was found better in explaining multiple level/variable error fields in the short range (up to 3 days), especially on the smaller scales. This result, as the finding in subsection (3c), suggests that the higher resolution ECMWF model (T255L40) may be more realistic than the NCEP model (T126 or T62, L28). This suggestion is supported by the results of D. Richardson (2001, personal communication) who found that the ECMWF model generates more skillful forecasts than the NCEP model when started from the same (NCEP) initial analysis field.

c) Optimal combinations of perturbations added more value to the ECMWF than to the NCEP ensemble. This may be due to an orthogonalization of initial perturbations performed for the ECMWF but not for the NCEP ensemble.

4) Interestingly, when ensembles were used to explain errors in a control forecast made with the other center's model their skill was dramatically reduced on the larger spatial scales. This suggests that some large scale errors may arise due to unrealistic instabilities that are model specific. These model specific errors can be captured only through an ensemble generated by the same model.

5) NCEP ensemble perturbations exhibit too high correlation among themselves, especially on smaller scales.

This suggests that an introduction of more diversity in the ensemble initial perturbations through a regional orthogonalization procedure applied on the smaller scales may make the ensemble more realistic and lead to improved forecast performance. The perturbation vs error correlation analysis (PECA) scheme introduced in this study provides a useful diagnostic and verification tool to achieve this goal.

Acknowledgments. The research described in this paper is an outgrowth of early experiments carried out by Jun Du (EMC), in collaboration with the second author. The authors had stimulating discussions with Roberto Buizza (ECMWF), Peter Houtekamer (CMC Environment Canada), and Jeff Anderson (NCAR) prior to, and with Istvan Szunyogh (Uni. of Maryland) and John Derber (EMC) during the research. It is a pleasure to thank Yuejian Zhu and Richard Wobus for their technical help. The authors are grateful to David Burridge, director, and the staff of ECMWF, in particular Horst Boettger and John Henessy, for participating in an exchange of ensemble forecast data between ECMWF and NCEP. We would also like to thank Kenneth Campana and Glenn White from NCEP for reading the manuscript carefully and helpful suggestions.

REFERENCES

Atger, F., 1999: The skill of ensemble prediction systems. *Mon. Wea. Rev.*, **127**, 1941-1953.

Barkmeijer, J., R. Buizza and T.N. Palmer, 1999: 3D-Var Hessian singular vectors and their potential use in the ECMWF Ensemble Prediction System. *Quart. J. Roy. Meteor. Soc.*, **125**, 2333-2351.

Barkmeijer, J., R. Buizza, T. N. Palmer, K. Puri, and Mahfouf, J.-F., 2001: Tropical singular vectors computed with linearized diabatic physics. *Quart. J. Roy. Meteor. Soc.*, **127**, 685-708.

Bishop, C.H., B.J. Etherton and S.J. Majumdar, 2000: Adaptive Sampling with the Ensemble Transform Kalman Filter. Part I: Theoretical Aspects. *Mon. Wea. Rev.*, **129**, 420-436.

Buizza, R. 2001: Accuracy and potential economical value of categorical and probabilistic forecasts of discrete events. *Mon. Wea. Rev.*, **129**, 2329-2345.

Buizza, R., and Palmer, T. N., 1999: Ensemble data assimilation. Proceedings of the 17th

Conference on Weather Analysis and Forecasting, 13-17 September 1999, Denver, Colorado, US, pp 241.

Buizza, R. and T. N. Palmer, 1995: The singular-vector structure of the atmospheric global circulation. *J. Atmos. Sci.*, **52**, 1434-1456.

Evans, R.E., M.S.J.Harrison, R.J. Graham and K.R.Mylne, 2000: Joint medium-range ensembles from the Met. Office and ECMWF systems. *Mon. Wea. Rev.*, **128**, 3104-3127.

Hamill, T. M. and C. Snyder, 2000: A comparison of probabilistic forecast from bred, singular-vector, and perturbed observation ensembles. *Mon. Wea. Rev.*, **128**, 1835-1851.

Holton, J. 1992: An Introduction to Dynamic Meteorology. Academic Press, 511pp.

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

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

Majumdar, S.J. C.H. Bishop, B.J. Etherton, and Z. Toth, 2001: Adaptive Sampling with the Ensemble Transform Kalman Filter. Part II: Field Program Implementation. *Mon. Wea. Rev.*, **127** (in press).

Molteni, F., R. Buizza, T. Palmer, and T. Petroliagis, 1996: The ECMWF ensemble prediction system: Methodology and validation. *Quart. J. Roy. Meteor. Soc.*, **122**, 73-119.

Molteni, F., and Buizza, R., 1999: Validation of the ECMWF Ensemble Prediction System using empirical orthogonal functions. *Mon. Wea. Rev.*, **127**, 2346-2358.

NOAA/NASA/USAF, 1976: US Standard Atmosphere 1976, Washington DC, 227pp.

Parrish, D. F., and J. Derber, 1992: The National Meteorological Center's spectral statistical-

interpolation analysis system. *Mon. Wea. Rev.*, **120**, 1747-1763.

Rabier, F., E. Klinker, P. Courtier and A. Hollingsworth, 1996: Sensitivity of forecast errors to initial conditions. *Quart. J. Roy. Meteor. Soc.*, **122**, 121-150.

Richardson, D. S., 2000: Skill and economic value of the ECMWF ensemble prediction system. *Quart. J. Roy. Meteor. Soc.*, **126**, 649-668.

Reynolds, C. A., and R. M. Errico, 1999: Convergence of singular vectors toward Lyapunov vectors. *Mon. Wea. Rev.*, **127**, 2309-2323.

Saha, S., 2001: Empirical Orthogonal Teleconnections (EOT). An analysis of NCEP and ECMWF forecast error patterns, available online at: http://lnx40.ncep.noaa.gov/

Stanski, H. R., L. J. Wilson, and W. R. Burrows, 1989: Survey of common verification methods in meteorology. WMO World Weather Watch Technical Report No. 8, WMO/TD. No. 358.

Szunyogh, I., E. Kalnay, and Z. Toth, 1997: A comparison of Lyapunov and optimal vectors in a low-resolution GCM. *Tellus*, **48A**, 200-227.

Toth, Z., and E. Kalnay, 1993: Ensemble forecasting at NMC: the generation of perturbations. *Bull. Amer. Meteor. Soc.*, **174**, 2317-2330.

Toth, Z., and E. Kalnay, 1997: Ensemble forecasting at NCEP and the breeding method. *Mon. Wea. Rev.*, **125**, 3297-3319.

Toth, Z., Y. Zhu, and T. Marchok, 2001: On the ability of ensembles to distinguish between forecasts with small and large uncertainty. *Wea. Forecasting*, **16**, 436-477.

Van den Dool, H. M., and L. Rukhovets, 1994: On the weights for an ensemble averaged 6-10 day forecast at NMC. *Weather and Forecasting*, **9**, No. 3, 457-465.

Van den Dool, H. M., S. Saha and A. Johansson, 2000: Empirical Orthogonal Teleconnections.

J. Climate, 13, 1421-1435.

Wei, M., 2000: Quantifying local instability and predictability of chaotic dynamical system by means of local metric entropy. *Int. J. of Bifurcation and Chaos*, **10**, 135-154.

Wei, M., and J. S. Frederiksen 2002: Error growth and dynamical vectors during Southern Hemisphere blocking (submitted to Nonlinear Pro. Geophy.)

Zhu, Y., G. Iyengar, Z. Toth, S. M. Tracton and T. Marchok, 1996: Objective evaluation of the NCEP global ensemble forecasting system . Preprints , 15th AMS Conference on Weather Analysis and Forecasting, Norfolk, Virginia.

Zhu, Y., Z. Toth, R. Wobus, D. Richardson, and K. Mylne, 2002: On the economic value of ensemble based weather forecasts. *Bull. Amer. Meteor. Soc.*, in print.

Figure Captions

Fig. 1. NCEP analyzed 500 hPa geopotential height field over North America (a), corresponding 8-day lead time forecast error field (b), optimally combined perturbation (c), residual error (d), all valid at 1200 UTC May 8, 2001.

Fig. 2. As in Fig.1, but for 10-day lead time valid at 1200 UTC May 10, 2001.

Fig. 3. Correlation between 500hPa geopotential height in NCEP and ECMWF control forecast error and the corresponding ensemble perturbations (EPs and NPs), averaged over a 30-day period starting at 1200 UTC April 01, 2001, over (a) the global, (b) Northern Hemisphere, (c) Southern Hemisphere, (d) Tropical, (e) North American and (f) European domains.

Fig. 4. Correlation between various number of optimally combined NPs, EPs and NMC perturbations, and the respective forecast error for lead times of 1, 2, 3, 5 and 7 days over the same domains as in Fig. 3.

Fig. 5. Contour display of the correlation between all different number of optimally combined NPs and NCEP forecast errors over the same domains as in Fig. 3.

Fig. 6. As in Fig. 5, but for ECMWF ensembles.

Fig. 7. Correlation between NCEP (U, V, T at 250hPa, 500hPa and 850hPa) and ECMWF (U, V, T at 200hPa, 500hPa and 850hPa) forecast errors and the corresponding ensemble perturbations (EPs and NPs) over the same domains as in Fig. 3.

Fig. 8. Correlation between forecast error and 8 randomly chosen (dashed), "perfect" and actual (dotted) ensemble perturbations over the same domain as in Fig. 3.