A New Measure of Ensemble Performance: Perturbation versus Error Correlation Analysis (PECA)

MOZHENG WEI

UCAR and NCEP/Environmental Modeling Center, Camp Springs, Maryland

ZOLTAN TOTH

SAIC and NCEP/Environmental Modeling Center, Camp Springs, Maryland

(Manuscript received 6 June 2002, in final form 18 November 2002)

ABSTRACT

Most existing ensemble forecast verification statistics are influenced by the quality of not only the ensemble generation scheme, but also the forecast model and the analysis scheme. In this study, a new tool called perturbation versus error correlation analysis (PECA) is introduced that lessens the influence of the initial errors that affect the quality of the analysis. 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.

Ensemble perturbations from both the National Centers for Environmental Prediction (NCEP) and the European Centre for Medium-Range Weather Forecasts (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 [used in the ''NMC'' (for National Meteorological Center, the former name of NCEP) method for defining forecast error covariance matrices]. 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 perfectly reliable ensemble the spread of ensemble members around the ensemble mean forecast equals the root-mean-square (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; Atger 1999; Richardson 2000; Toth et al. 2002; Zhu et al. 2002). 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) are 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), which 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 aspect of ensemble performance and measures more directly the effect of ensemble perturbations on ensemble performance. The proposed method, called perturbation versus error correlation analysis (PECA), is based on the comparison of ensemble forecast perturbations (ensemble

Corresponding author address: Mozheng Wei, NCEP/Environmental Modeling Center, 5200 Auth Rd., Camp Springs, MD 20746. E-mail: Mozheng.Wei@noaa.gov

forecasts minus control) and forecast error patterns (control forecast minus verifying analysis).

The motivation for the development of such an ensemble evaluation measure is twofold. First, apart from the above-mentioned conventional measures, we need an additional tool that may be less sensitive to differences in the quality of the analysis scheme used to generate the ensembles, and that can be more readily applied in studies that compare ensemble forecasts generated by different NWP forecast centers. 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, that is, 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, in terms of a comparison of the perturbation and error EOF spectra. Therefore, the Molteni and Buizza (1999) approach, unlike the method proposed here (PECA), evaluates only the statistical consistency (reliability), but not the forecast skill (resolution) of ensemble systems. Another study in interpreting ensemble forecasts by Stephenson and Doblas-Reyes (2000) computed the multivariate skewness, kurtosis, and entropy from ensemble forecasts based on the Monte Carlo method. The tools proposed in this paper also were applied to the European Centre for Medium-Range Weather Forecasts (ECMWF) ensemble forecasts for the period of 20–30 December 1997.

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 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. Ensemble perturbations are also compared with ''NMC'' (for National Meteorological Center, the Former name of NCEP) type perturbations [based on differences between short-range forecasts valid at the same time; see Parrish and Derber (1992)]. For comparison, PECA results are also presented for randomly selected ensemble perturbations, and for ''perfect'' ensembles where one of the ensemble members is taken as the verifying analysis. 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) center, are defined as the differences between the perturbed forecasts and their respective control forecast (started from an unperturbed analysis):

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

where C, the originating center, is either NCEP or ECMWF, and $i = 1, 2, \ldots, N$, with $N = 10$ or 20 for NCEP and 50 for ECMWF. Here, $\mathbf{P}_i^C(t)$, $\mathbf{F}_i^C(t)$, and $\mathbf{F}_{\text{control}}^C(t)$ are ensemble perturbations, and perturbed and control forecasts, respectively.

The NCEP and ECMWF ensemble forecast systems are based on the breeding and singular-vector methods, respectively. In both systems, the initial perturbations are designed to span only a subspace in the phase space representing fast-growing errors. The NCEP ensemble perturbations, at 24-h 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). Note that stochastic perturbations are introduced in the ECMWF (but not the NCEP) perturbed forecasts to represent model-related errors. Neither ensemble system simulates the effect of imperfect boundary conditions. The ECMWF ensemble system had no initial perturbations in the Tropics until January 2002. The recent introduction of targeted tropical singular vectors (TSVs) (Barkmeijer et al. 2001) is expected to improve performance of the ECMWF ensemble in the vicinity of tropical areas.

We refer to Toth and Kalnay (1993, 1997), Buizza and Palmer (1995), and Molteni et al. (1996) for more details about the two ensemble forecast systems. At the time of writing, products consist of 10 ensemble forecasts both at 0000 and 1200 UTC at NCEP, and a 50-member ensemble at 1200 UTC at ECMWF each day.

In both systems, forecast errors, $\mathbf{E}^C(t)$, are defined as the difference between the control forecast, $[\mathbf{F}_{\text{control}}^C(t)]$, and the verifying analysis, $[\mathbf{F}_{\text{analysis}}^C(t)]$, from the same center:

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

Since the analysis field itself has errors in it, the forecast error defined by Eq. (2) is the difference between the true forecast error (forecast minus the true state of the atmosphere) and the true analysis error (analysis minus truth). This fact will have to be considered when interpreting the results of this study, especially at short lead times, when the magnitudes of analysis and forecast errors are comparable. An a posteriori optimal combination of n perturbations is obtained by solving the least squares problem:

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

Having obtained α_i from the above equation, the optimally combined vector ($\mathbf{P}_{\text{optimal}}^C$) is defined as

$$\mathbf{P}^C_{\text{optimal}} = \sum_{i=1}^{n} \alpha_i \mathbf{P}^C_i. \quad (4)$$

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 as a diagnostic (not as a prognostic) tool.

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

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

PECA is defined as the PAC between $\mathbf{X} = \mathbf{P}^C_i$ (or $\mathbf{X} = \mathbf{P}^C_{\text{optimal}}$) and $\mathbf{Y} = \mathbf{E}^C$. Note that the square of the correlation, A_c^2, can be considered to be the 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, 20°–77.5°N) and Southern Hemisphere extratropics (SH, 20°–77.5°S), the Tropics (20°S–20°N), North America (NA, 20°–60°N, 140°–50°W), and Europe (EU, 30°–77.5°N, 20°W–40°E). Correlation values computed between individual perturbations ($\mathbf{P}^C_i$) and the forecast errors ($\mathbf{E}^C$) will be averaged over the first 10 individual perturbations at 1200 UTC in most cases studied. In addition, correlation values between $\mathbf{P}^C_{\text{optimal}}$ and $\mathbf{E}^C(t)$ 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 also reflect, beyond the quality of the ensemble generation technique, 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 $(\mathbf{F}_{\text{control}} - \mathbf{C})$ and $(\mathbf{X}_{\text{analysis}} - \mathbf{C})$, 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 may also be reduced when using PECA. This may be true for random types of model errors (see Toth and Vannitsem 2002). Systematic 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 of both types of model imperfection related problems will be discussed in the next section.

When comparing ensembles generated by different NWP centers, higher PECA values, which mean that each ensemble perturbation can better explain its own center's forecast errors, are thus indication of higher quality. 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, perfect PECA experiment) 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 the similarity between perturbation patterns over a selected geographical domain (pattern spread).

If the PECA values in the perfect model/ensemble case are, for example, higher than those computed with real error fields, it is an indication of an underdispersive ensemble. In such an ensemble the perturbation patterns show a lack of variability for properly explaining fore-

cast 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, perfect 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. Different kinds of daily results based on the conventional measures, including the period we have analyzed in this paper, can be found online at the NCEP Web site (http://sgi62.wwb.noaa.gov:8080/ens/verif.html).

3. NECP and ECMWF ensemble results

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.

Figure 1a shows the NCEP 500-hPa geopotential height analysis field valid at 1200 UTC 8 May 2001. The analysis field shows a wavelike structure across the higher latitudes of NA. The corresponding error pattern of an 8-day lead time forecast initialized at 1200 UTC 30 April 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 an average of 69% explained variance for 8-day forecast errors over the experimental 30-day period. A large part (60%) of the error in Fig. 1b remains unexplained by the ensemble. This is also evident from Fig. 1d, which displays the residual error $\mathbf{R}(t)$, defined as the part of the forecast error that cannot be explained by $\mathbf{P}_{\text{optimal}}(t)$; that is, $\mathbf{R}(t) = \mathbf{E}(t) - \mathbf{P}_{\text{optimal}}(t)$. 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 UTC 30 April 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 10 May 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 40°N. The optimal perturbation (Fig. 2c) successfully explains the error pattern, including most of the details. Ninety-four percent 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

Figures 3a–f show the PECA correlation values computed between the NCEP (solid lines) and ECMWF (dotted) forecast error and the corresponding ensemble perturbations for 500-hPa 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 500 hPa is one of the variables with the least amount of systematic error; thus, the correlation values, at least in extratropics, will mainly 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 1 April 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 an optimal combination of the individual vectors [see Eq. (4)]. Since the forecast error is not known exactly [see Eq. (2) and associated discussion], the relationship between the true forecast error and the dynamically evolving ensemble perturbations, especially at short lead times, is somewhat underestimated by the PECA values.

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), optimal perturbations can explain a larger amount of forecast error variance. This is due to the fewer degrees of freedom in the error (and perturbation) fields over the smaller areas. This may also explain the larger sampling fluctuations (noise) observed in the results valid for smaller geographical areas. In contrast, the PECA values computed from and averaged over individual perturbations are not influenced by the size of the regions.

For individual perturbations, the NCEP ensemble performs better out to 7-day lead time (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-day lead time. When the systems are compared using the optimal perturbations, the NCEP ensemble exhibits higher correlation up to 2–3-day 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–

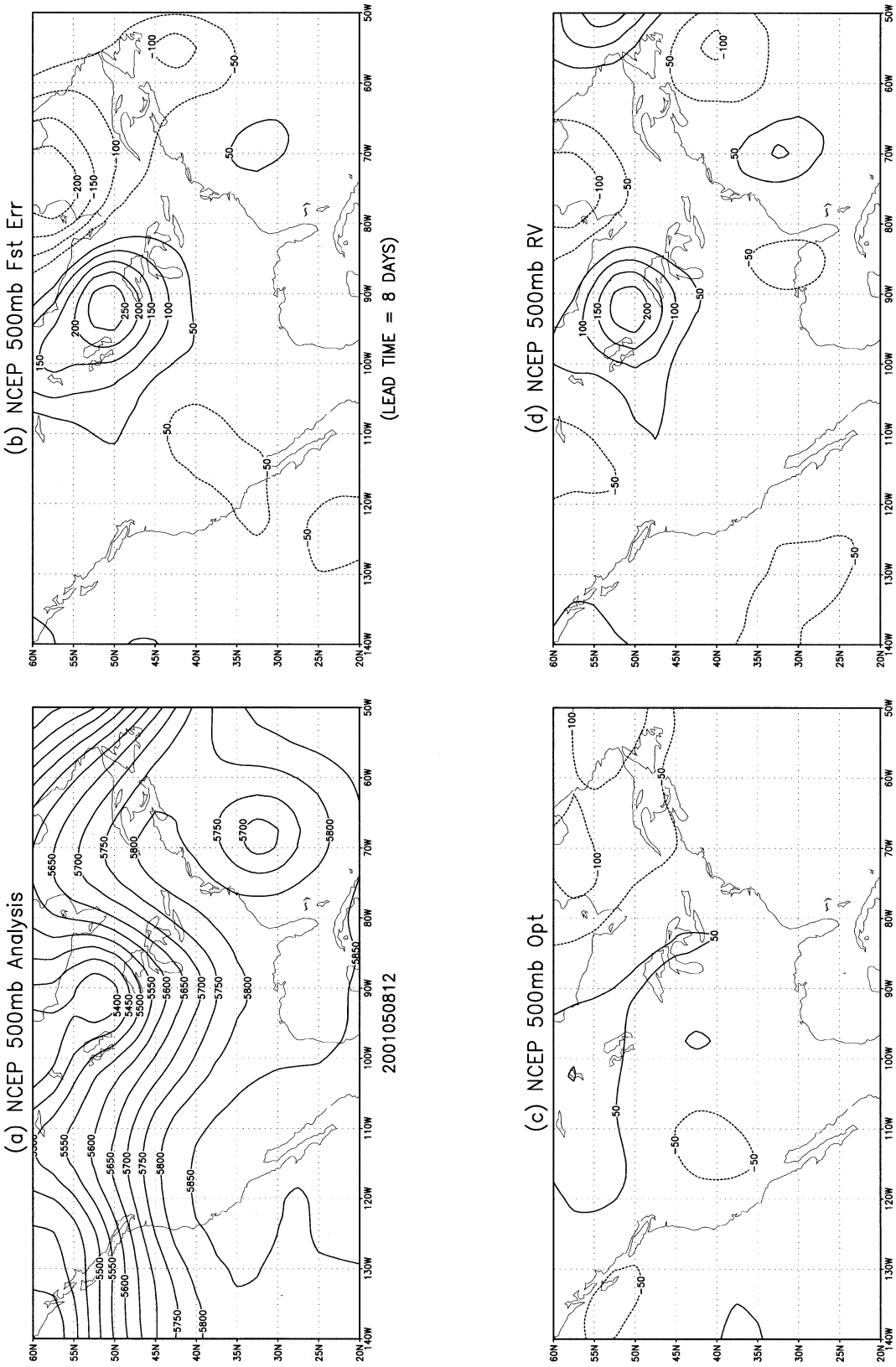

FIG. 1. (a) NCEP-analyzed 500-hPa geopotential height field over North America, (b) corresponding 8-day lead time forecast error field, (c) optimally combined perturbation, and (d) residual error all valid at 1200 UTC 8 May 2001.

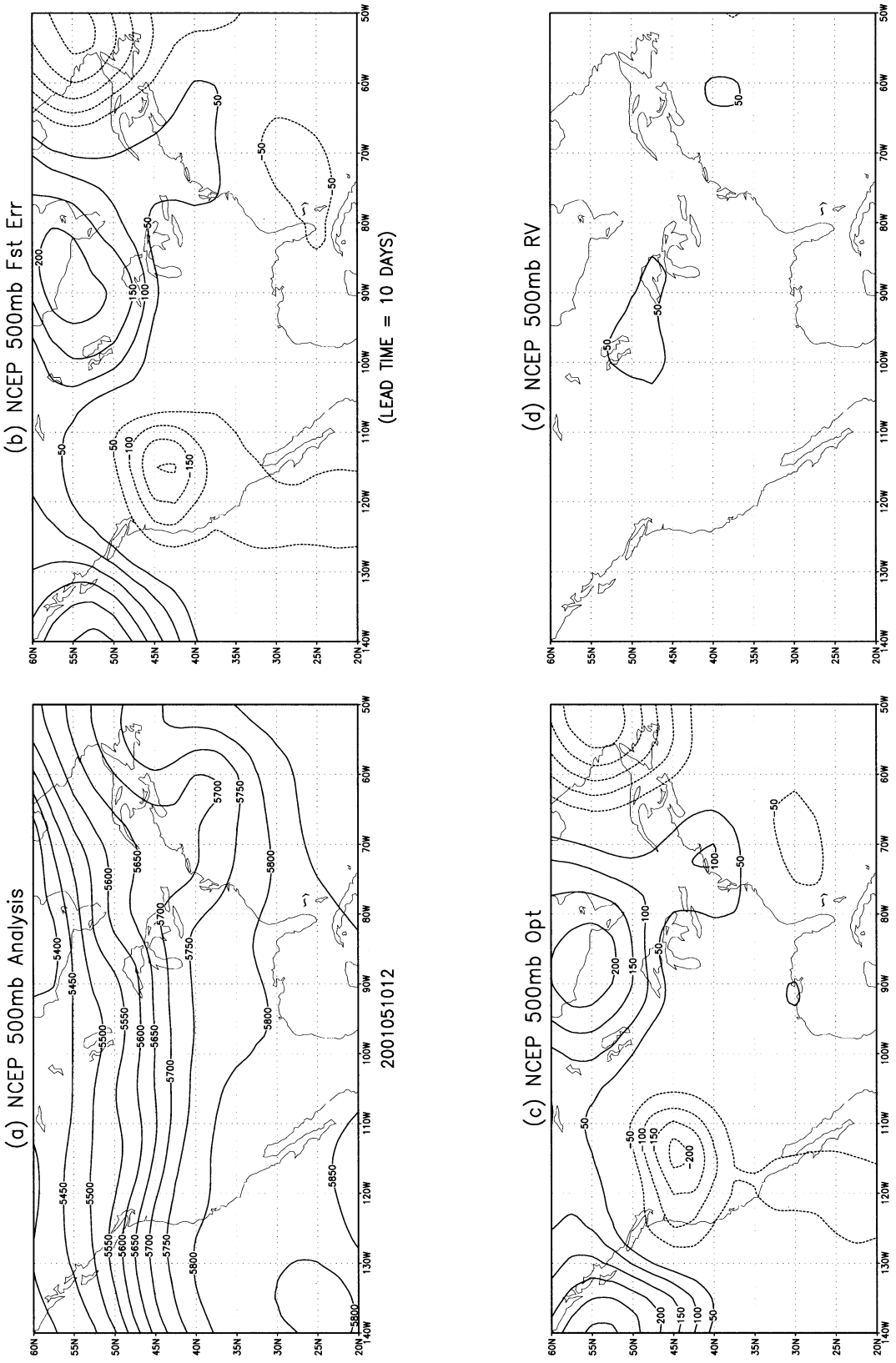

FIG. 2. As in Fig. 1 but for 10-day lead time valid at 1200 UTC 10 May 2001.

5 days), while systematic model-related errors, in a relative sense, may influence the results more at longer (6–10 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.

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 two factors. Perturbations, including the error fields that evolve quasi linearly, are attracted to the fastest growing nonlinear perturbations (Toth and Kalnay 1993, 1997), which are related to the leading Lyapunov vectors (LVs) (e.g., Buizza and Palmer 1995; Szunyogh et al. 1997; Reynolds and Errico 1999; Frederiksen 2000; Wei 2000; Wei and Frederiksen 2002, manuscript submitted to *Nonlinear Processes Geophys.*). 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 ECMWF forecast error fields, and the first 10 ECMWF perturbations (EPs) are used to explain NCEP forecast error fields. The results are shown as dashed and dash–dotted lines, respectively, in Fig. 3.

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 it explains only 10% in the ECMWF forecast error. Over the smaller North American and European areas, however, no such large discrepancy is present (see Figs. 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.

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 an advantage. Over the Southern Hemisphere and the Tropics, however, the optimally combined EPs can explain NCEP forecast error fields better than the optimally combined NPs can explain ECMWF forecast errors.

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.

Over several domains, the inclusion of a few ECMWF members with the NCEP ensemble leads to slightly higher explained error variance for the NCEP forecast error fields (dash–dot–dotted lines) at intermediate lead times. At short and 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-day lead times.

d. The effect of ensemble size

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

Figure 4 shows the PECA values between various number of optimally combined NPs and EPs and the respective NCEP and ECMWF forecast error fields, where 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 0000 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 correlation between the forecast error fields 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 com-

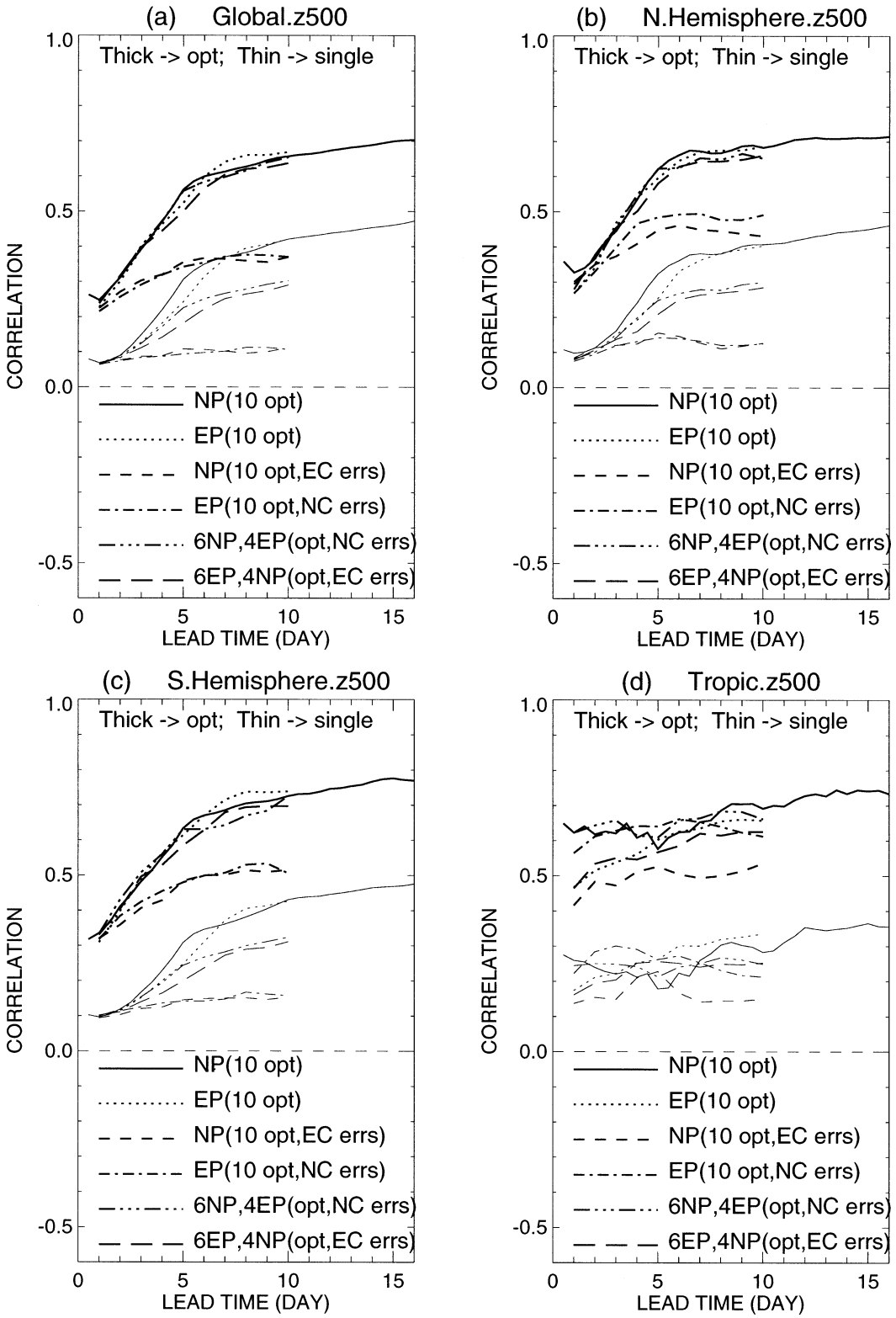

FIG. 3. Correlation between 500-hPa geopotential height in NCEP and ECMWF control forecast error and the corresponding ensemble perturbations (NPs and EPs), averaged over a 30-day period starting at 1200 UTC 1 Apr 2001, over the (a) global, (b) NH, (c) SH, (d) tropical, (e) NA, and (f) EU domains.

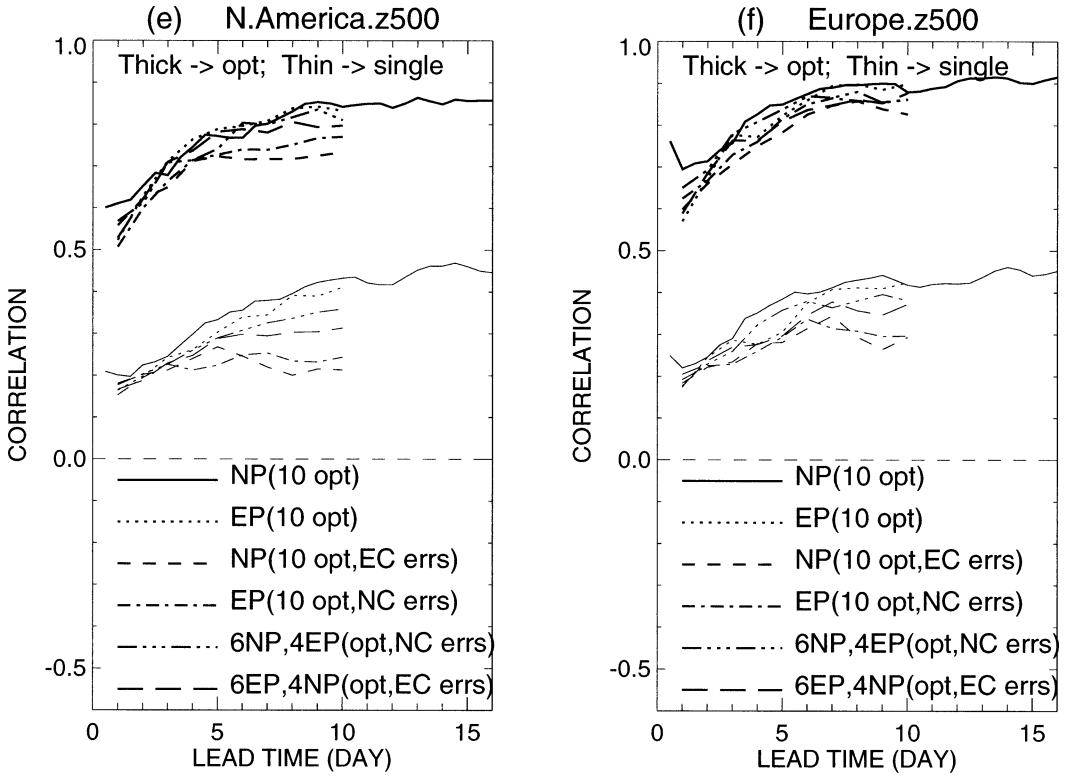

FIG. 3. (*Continued*)

bined 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-day or longer lead times.

It is interesting to note in Figs. 4e and 4f that at short lead times and over smaller areas, an increase in ensemble membership brings significant improvement even beyond 30–40 members. This is also true at larger lead times when the error fields, on average, are rather well explained even by a smaller number of 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 a technique called the NMC method for the construction of forecast error covariance matrices. This technique has frequently been used in data assimilation schemes worldwide. To test how efficient the lagged forecast difference fields are in explaining 1-day forecast error fields, when compared to ensemble perturbations, difference fields between 24- and 48-h lead time forecasts were generated for both the NCEP and ECMWF control forecasts over a period preceding the experimental ensemble date (1200 UTC 5 March–1200 UTC 3 April 2001). In our experiment, we computed 30 consecutive vector fields; that is,

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

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

The NMC perturbation vectors were evaluated in a fashion similar to the ensemble perturbations using the same control forecast error fields from NCEP and ECMWF as in Fig. 3. Different numbers of NMC vectors were optimally combined, like the ensemble perturbations, using equations similar to (3) and (4).

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 should be mentioned that there are some differences between the way the NMC vectors are generated in this paper and in practical implementations of the NMC method in three-dimensional variational (3DVAR) data assimilation schemes. For instance, in this paper the NMC vectors are generated for a period just prior to their use, while in practical data assimilation imple-

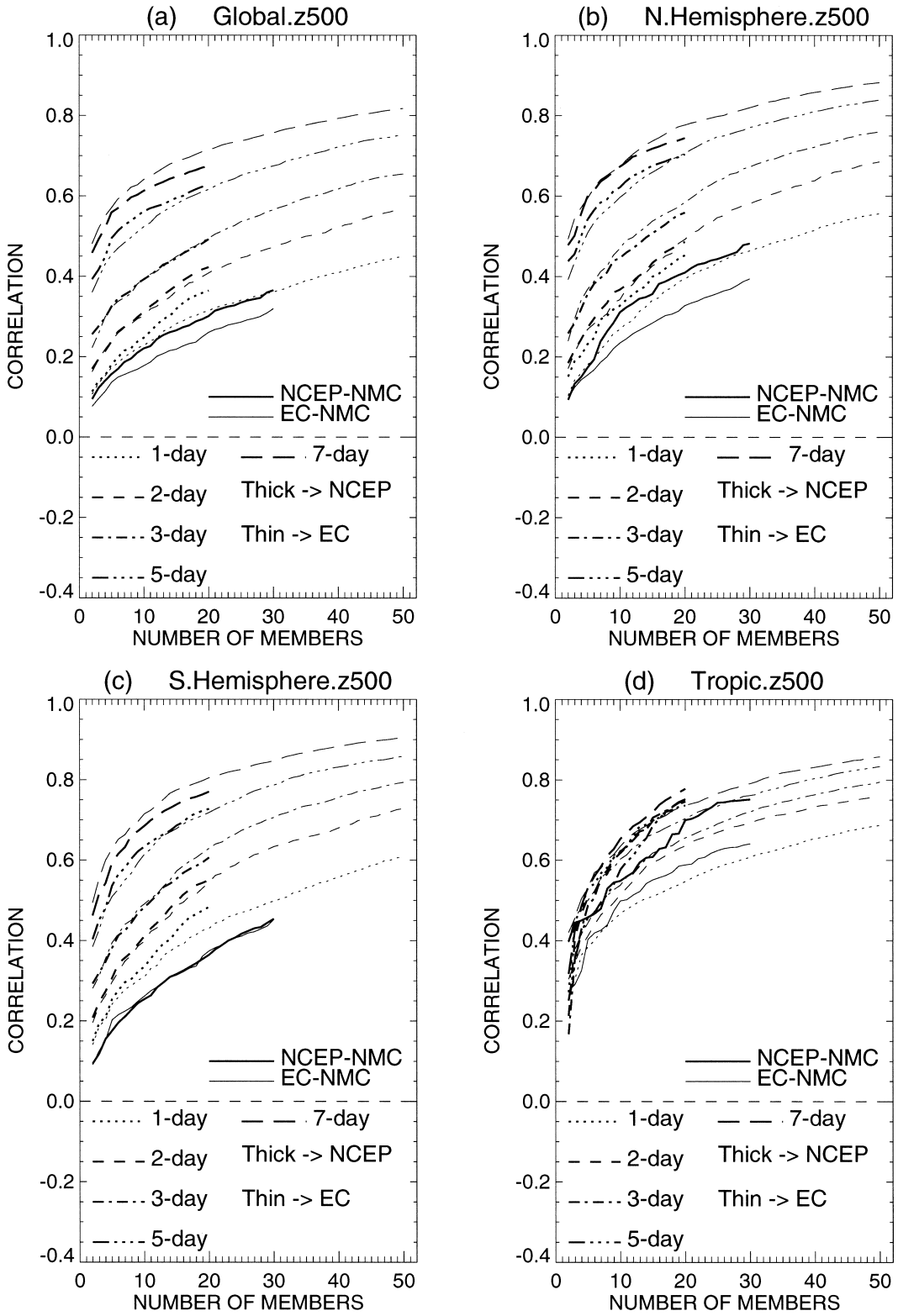

FIG. 4. Correlation between various numbers of optimally combined NPs, EPs, and NMC perturbations, and the respective forecast error for lead times of 1, 2, 3, 5, and 7 days over the same domains as in Fig. 3.

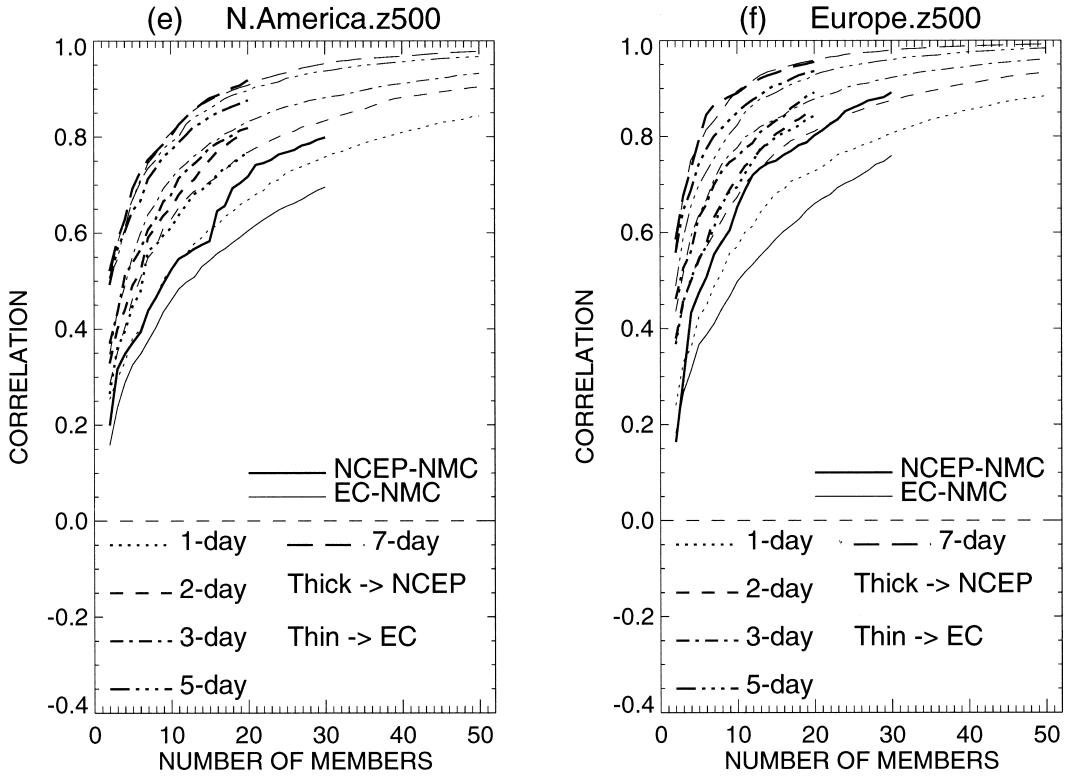

FIG. 4. (*Continued*)

mentations, the statistics are generated over a fixed period in the past. The NMC vectors are averaged over space and possibly over the wavenumber spectrum as well. Therefore the results presented may overestimate the value of NMC vectors in explaining flow-dependent error variance.

By comparing the dotted and solid lines in Fig. 4, it becomes clear that for most domains, the ensemble perturbations can better explain their respective 1-day forecast error than the NMC perturbations. 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. 4d). As we mentioned above, the introduction of TSVs on 22 January 2002 in the ECMWF Ensemble Prediction System (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. Note that systematic, lead-time-dependent errors (i.e., model drift) may be captured by lagged forecast differences, but not by the ensemble forecasts. Possibly larger regional biases in the NCEP ensemble may contribute to the difference found between the performance of the two centers' NMC vectors.

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. Three-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 three variables, temperature (T) and velocity (U, V) at three levels. Based on data at 850, 500, and 250 hPa (200 hPa for ECMWF), we define a new variable p: $p = (U, V, \alpha T)$, where $\alpha = \sqrt{C_p/T_r}$, $C_p = 1004.0$ J kg^{-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 *U.S. Standard Atmospheric, 1976.* Thus the inner product $\langle p, p \rangle$ has the form of total energy as defined

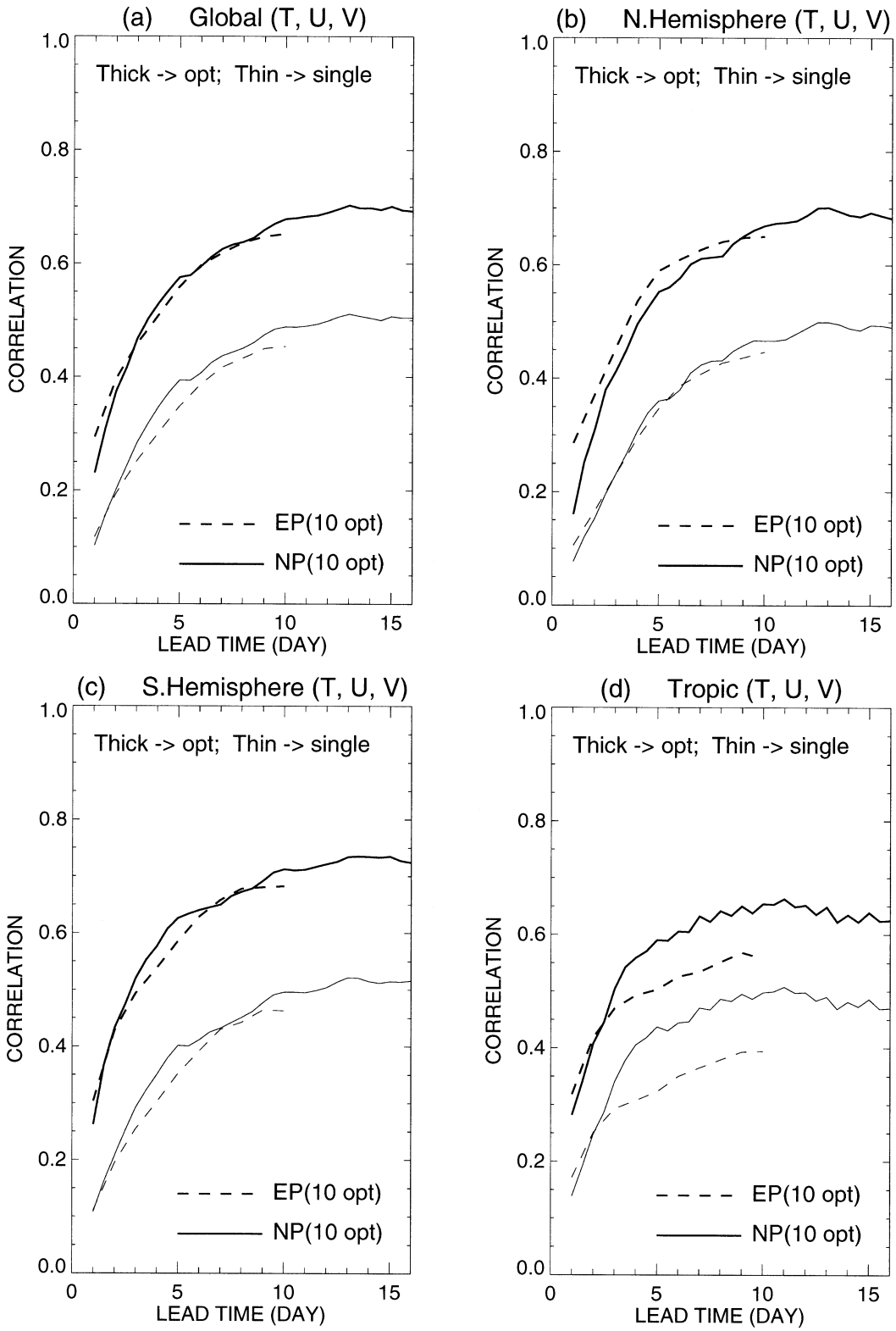

FIG. 5. Correlation between NCEP (*U, V, T* at 250, 500, and 850 hPa) and ECMWF (*U, V, T* at 200, 500, and 850 hPa) forecast errors and the corresponding ensemble perturbations (NPs and EPs) over the same domains as in Fig. 3.

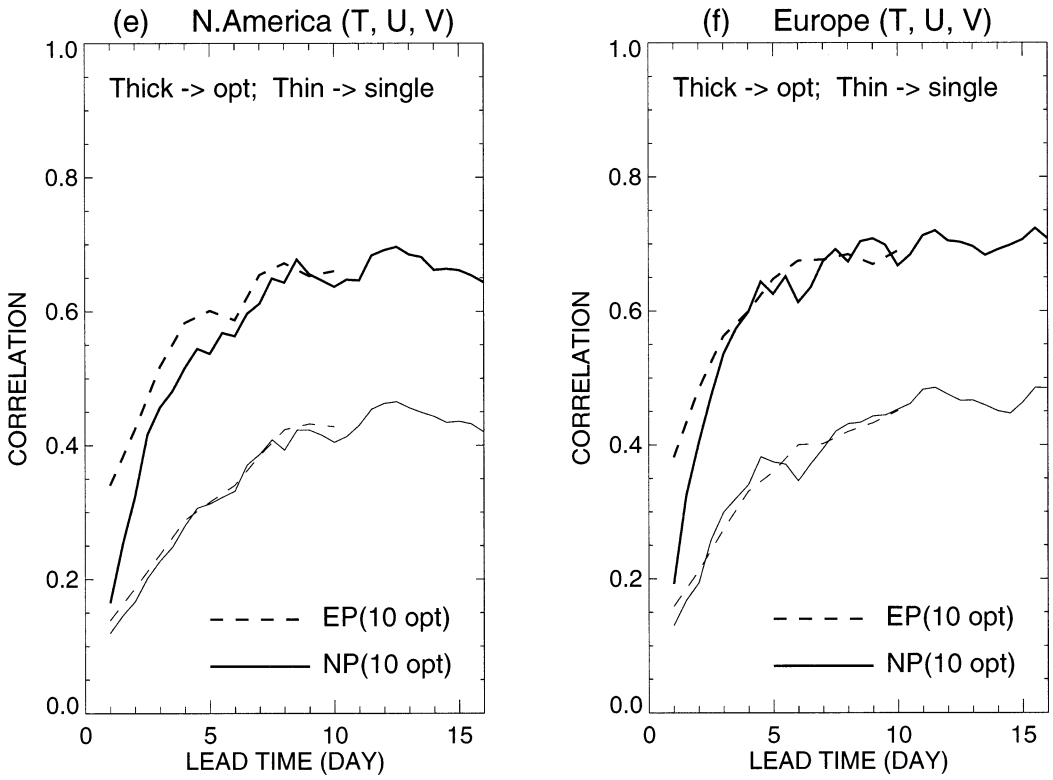

FIG. 5. (*Continued*)

by Rabier et al. (1996) and Barkmeijer et al. (1999). The dimension of p is 9 m, 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. 5 in a manner similar to Fig. 3. Note first that due to an increase in the degrees of freedom, PECA values for the optimally combined ensemble perturbations in Fig. 5 are considerably lower than in Fig. 3. While the individual NCEP ensemble perturbations performed better than the ECMWF for the case with one variable at one level, the ECMWF ensemble becomes more efficient when the multilevel/multivariable 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 on the smaller scales 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 also supported by the results of D. 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-day lead time, the NCEP individual multilayer/multivariable perturbations perform better than the ECMWF perturbations over the larger domains (global, NH, SH, and Tropics). For short lead times, 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 remains better over the Tropics. The optimal perturbations from the two systems perform similarly over the global and SH 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. Results for perfect and random ensembles

In the above section, the ability of ensemble perturbations in explaining forecast error fields 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. 6 are identical to the solid lines in Fig. 3, except here only eight perturbations are combined optimally to estimate the actual skill of

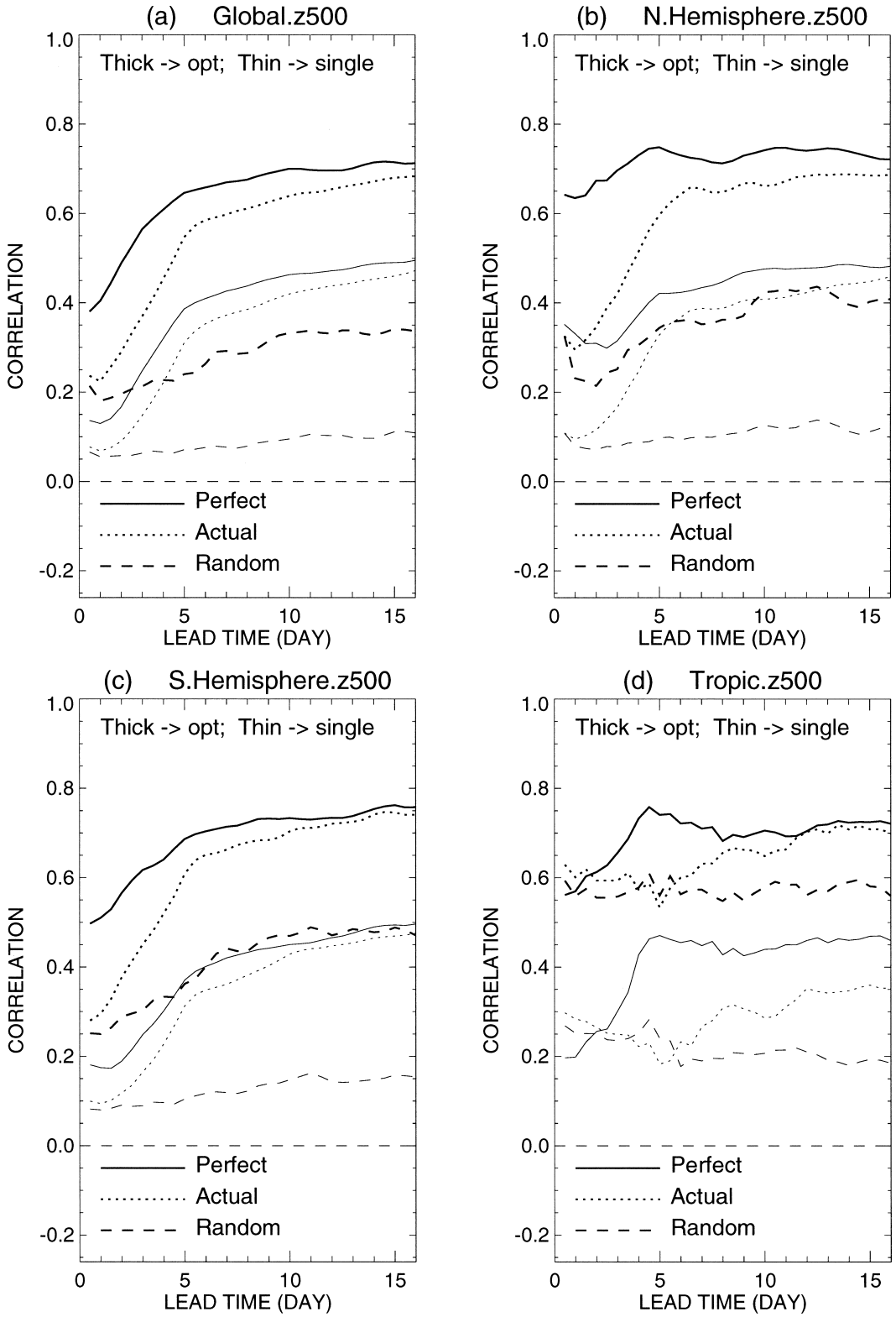

FIG. 6. Correlation between forecast error and eight randomly chosen (dashed) ''perfect'' and actual (dotted) ensemble perturbations over the same domain as in Fig. 3.

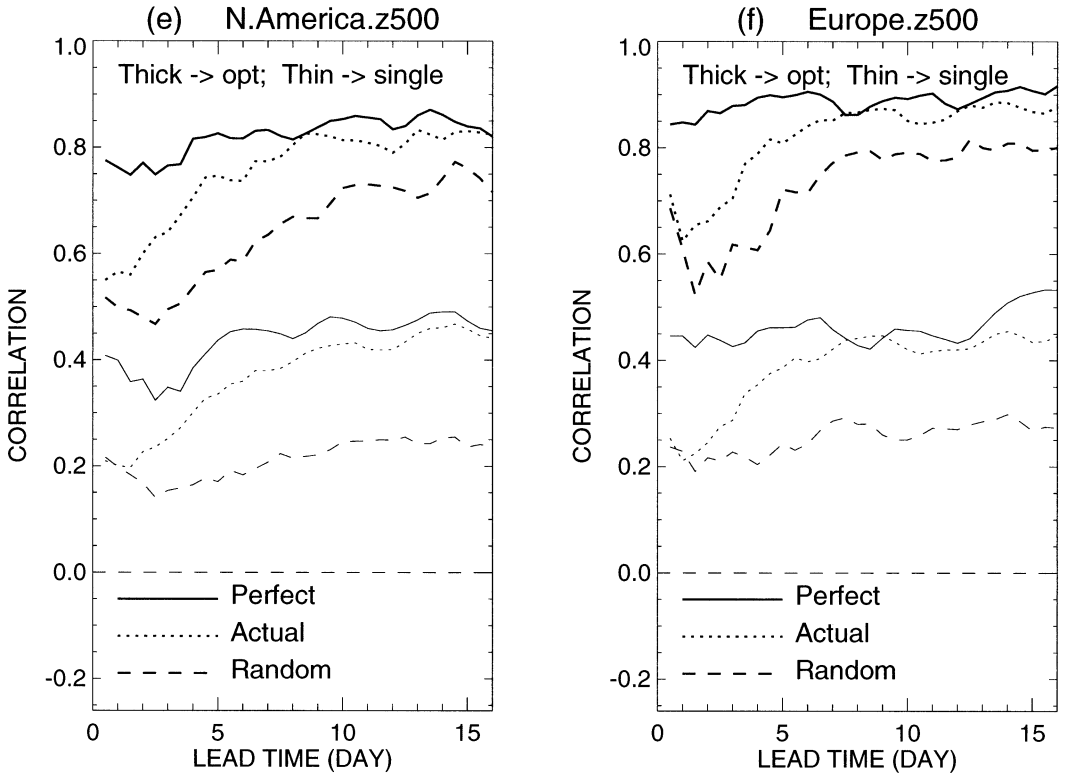

FIG. 6. (*Continued*)

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 eight perturbations are presented as the dashed lines in Fig. 6. 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 for both individual and optimally combined perturbations. The results indicate that the randomly chosen perturbations are dynamically not relevant and cannot explain flow-dependent forecast errors.

When comparing PECA values for the real and random ensemble cases, one should also note that the score lowering effect of the inclusion of analysis errors [see Eq. (2) and associated discussion] is expected to affect the results for the dynamically relevant real ensemble more than for the random ensemble. Therefore PECA results based on the true forecast error may show more advantage for the real ensemble than seen in Fig. 6, especially at shorter lead times.

If both the model and ensemble generation techniques 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 the truth and the remaining four pairs of members will be used to explain the ''error.'' In this case, $\mathbf{NP}_i(t) = \mathbf{F}_i^{\mathrm{NCEP}}(t) - \mathbf{F}_{\mathrm{control}}^{\mathrm{NCEP}}(t)$, where $i = 1, 2, \ldots, 8$. The forecast error in this case is defined as

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

where $j \neq i$. Note that the perfect PECA values defined above and shown as the solid lines in Fig. 6 are only a function of correlation among the members of the ensemble.

In Fig. 6 we first 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 contraction similar to 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 the initial time, given their low level of skill in explaining forecast errors (cf. low actual PECA values; dotted curve in Fig. 6).

For the global domain, for example, the initial correlation value for the perfect case (thin solid line in Fig. 6a) is as high as the correlation between the error and individual ensemble perturbations at around 3-day lead time (thin dotted lines in Fig. 6a). The problem is exacerbated over the smaller areas. The results suggest that by imposing more diversity among the ensemble members on the smaller scales, the introduction of re-

gional orthogonalization in the rescaling of the bred vectors (Toth and Kalnay 1997) may alleviate the problem and potentially lead to improved ensemble performance. This conjecture is corroborated by the results of Wang and Bishop (2002), who found that the use of the ensemble transform Kalman filter technique (ETKF; Bishop et al. 2001) for rescaling and orthogonalizing ensemble perturbations improved the performance of a bred ensemble.

5. Discussion and conclusions

One of the goals of ensemble forecasting is to generate a set of forecast scenarios that encompasses truth. The success of ensemble forecasts can be measured in a number of ways. Most existing verification tools measure the overall skill of an ensemble forecast system. The verification results from such methods are strongly influenced by the quality of the analysis around which the ensemble is initialized and the forecast model used, which reduces their value in assessing ensemble techniques. In this study a new metric is introduced that measures how well individual or optimally combined ensemble perturbations can explain forecast error variance (perturbation versus error correlation analysis, or PECA). This measure evaluates the performance of ensemble perturbations and not the full forecast fields. The more closely ensemble perturbations, on average, are correlated with forecast error, the better the ensemble represents the truth. By evaluating ensemble perturbations instead of full forecast fields, PECA reduces the influence of the magnitude of initial errors (which reflects the quality of the analysis scheme) and offers a more direct measure of ensemble performance.

Like all other common measures of forecast performance in numerical weather forecasts, such as root-mean-square and anomaly correlation, PECA values also depend on the regions and variables. When one compares the performances of two different forecasts from different models or centers, it is implied that the measures will be taken over the same region and using the same set of variables so as to have a fair comparison.

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 local Lyapunov vectors (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 the 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 (which are used at several NWP centers to construct forecast error covariance matrices in data assimilation schemes, using 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 error variance explained by a posteriori optimally combined perturbations increases with ensemble membership. The extrapolation of the results suggests that at least 100/(200) members are needed to explain most of the short-range (1 day) forecast error on continental/(global) scales. These numbers set a minimum requirement for the size of an ensemble to be used in fully ensemble-based data assimilation studies.
4) The NCEP and ECMWF ensembles generally exhibit a similar level of skill. The following, relatively minor differences were noted: (i) Individual NCEP perturbations were found to be more skillful in explaining errors in a single variable (500-hPa geopotential height) over the first 5–7 days of lead time. This may be an indication of more efficient initial ensemble perturbations in the NCEP ensemble. (ii) The ECMWF ensemble was found to be better in explaining multiple level/variable error fields in the short range (up to 3 days), especially on the smaller scales. This result, as shown in the finding in section 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. (iii) 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.
5) 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.

6) The experiments where the two centers ensembles were combined gave mixed results when compared to the use of a single center's ensemble in explaining that center's ensemble error fields. The PECA results do not favor a multimodel approach to ensemble forecasting.
7) 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 effective and lead to improved forecast performance. The perturbation versus 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 earlier experiments carried out by Jun Du (SAIC at 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 (University 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 Hennessy, for participating in an exchange of ensemble forecast data between ECMWF and NCEP. We would also like to thank Kenneth Campana, Glenn White from NCEP, and three anonymous reviewers for comments on an early version of this paper.

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.

——, ——, ——, K. Puri, and J.-F. Mahfouf, 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, 2001: Adaptive sampling with the ensemble transform Kalman filter. Part I: Theoretical aspects. *Mon. Wea. Rev.*, **129,** 420–436.

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

——, and ——, 1999: Ensemble data assimilation. Preprints, *17th Conf. on Weather Analysis and Forecasting*, Denver, CO, Amer. Meteor. Soc., 231–234.

Frederiksen, J. S., 2000: Singular vectors, finite-time normal modes and error growth during blocking. *J. Atmos. Sci.*, **57,** 312–333.

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

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.

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

——, ——, T. Palmer, and T. Petroliagis, 1996: The ECMWF ensemble prediction system: Methodology and validation. *Quart. J. Roy. Meteor. Soc.*, **122,** 73–119.

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.

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

Richardson, D. S., 2000: Skill and economic value of the ECMWF Ensemble Prediction System. *Quart. J. Roy. Meteor. Soc.*, **126,** 649–668.

Saha, S., cited 2001: Empirical orthogonal teleconnections (EOT). An analysis of NCEP and ECMWF forecast error patterns. [Available online at http://www.emc.ncep.noaa.gov/gmb/ssaha.]

Stanski, H. R., L. J. Wilson, and W. R. Burrows, 1989: Survey of common verification methods in meteorology. World Weather Watch Tech. Rep. 8, WMO/TD 358, 144 pp.

Stephenson, D. B., and F. J. Doblas-Reyes, 2000: Statistical methods for interpreting Monte Carlo ensemble forecasts. *Tellus*, **52A,** 300–322.

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.*, **74,** 2317–2330.

——, and ——, 1997: Ensemble forecasting at NCEP and the breeding method. *Mon. Wea. Rev.*, **125,** 3297–3319.

——, and S. Vannitsem, 2002: Model errors and ensemble forecasting. *Proc. Eighth ECMWF Workshop on Meteorological Operational Systems*, Reading, United kingdom, ECMWF, 146–154.

——, O. Talagrand, G. Candille, and Y. Zhu, 2002: Probability and ensemble forecasts. *Environmental Forecast Verification: A Practitioner's Guide in Atmospheric Science*, I. T. Jolliffe and D. B. Stephenson, Eds., John Wiley, in press.

Van den Dool, H. M., and L. Rukhovets, 1994: On the weights for an ensemble averaged 6–10-day forecast at NMC. *Wea. Forecasting*, **9,** 457–465.

——, S. Saha, and A. Johansson, 2000: Empirical orthogonal teleconnections. *J. Climate*, **13,** 1421–1435.

Wang, X., and C. H. Bishop, 2002: A comparison of breeding and ensemble transform Kalman filter ensemble forecast schemes. Preprints, *Symp. on Observations, Data Assimilation, and Probabilistic Prediction*, Orlando, FL, J28–J31.

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

Zhu, Y., Z. Toth, R. Wobus, D. Richardson, and K. Mylne, 2002: The economic value of ensemble-based weather forecasts. *Bull. Amer. Meteor. Soc.*, **83,** 73–83.