A Comparison of the ECMWF, MSC, and NCEP Global Ensemble Prediction Systems

ROBERTO BUIZZA,* P. L. HOUTEKAMER,+ ZOLTAN TOTH,# GERALD PELLERIN,+ MOZHENG WEI,@,&
AND YUEJIAN ZHU#

*European Centre for Medium-Range Weather Forecasts, Reading, United Kingdom
+Meteorological Service of Canada, Dorval, Quebec, Canada
#National Centers for Environmental Prediction, Washington, D.C.
@NCEP/EMC, Washington, D.C.

(Manuscript received 19 November 2003, in final form 27 May 2004)

ABSTRACT

The present paper summarizes the methodologies used at the European Centre for Medium-Range Weather Forecasts (ECMWF), the Meteorological Service of Canada (MSC), and the National Centers for Environmental Prediction (NCEP) to simulate the effect of initial and model uncertainties in ensemble forecasting. The characteristics of the three systems are compared for a 3-month period between May and July 2002. The main conclusions of the study are the following:

- the performance of ensemble prediction systems strongly depends on the quality of the data assimilation system used to create the unperturbed (best) initial condition and the numerical model used to generate the forecasts;
- a successful ensemble prediction system should simulate the effect of both initial and model-related uncertainties on forecast errors; and
- for all three global systems, the spread of ensemble forecasts is insufficient to systematically capture reality, suggesting that none of them is able to simulate all sources of forecast uncertainty.

The relative strengths and weaknesses of the three systems identified in this study can offer guidelines for the future development of ensemble forecasting techniques.

1. The need for ensemble prediction

The weather is a chaotic system: small errors in the initial conditions of a forecast grow rapidly and affect predictability. Furthermore, predictability is limited by model errors linked to the approximate simulation of atmospheric processes of the state-of-the-art numerical models. These two sources of uncertainty limit the skill of single, deterministic forecasts in an unpredictable way, with days of high/poor quality forecasts followed by days of poor/high quality forecasts. Ensemble prediction is a feasible way to complement a single, deterministic forecast with an estimate of the probability density function of forecast states.

Ensemble methods are based on a statistical sampling approach (Leith 1974) in which the forecast probability density function is approximated using a finite sample of forecast scenarios. These forecasts are started from a sample of states drawn from a probability density function of the initial state (which is often implicitly estimated) to sample initial-value-related forecast uncertainty. The ensemble forecasts are often integrated using a variety of different (or modified) numerical models with the aim of capturing model-related forecast uncertainty as well. Methodologies for ensemble-based data assimilation (Evensen 1994) could in principle provide the initial conditions for ensemble systems, but this has not yet been applied operationally at weather prediction centers. Similarly, diagnostics of multimodel ensembles could be used to assess the quality of different aspects of numerical weather prediction models (Houtekamer and Lefaivre 1997).

The focus of this work is on the more mature area of medium-range ensemble prediction. It is thought that the evolution of forecast uncertainty originated from initial condition errors can be described fairly well with available numerical prediction models. This is done in the operational ensemble prediction systems (EPSs) of different centers. Opinions diverge, however, on how to best describe the distribution of the initial errors and on how to subsequently sample that distribution. Three fairly different methods to generate an ensemble of initial conditions are currently in use at operational centers.

& Additional affiliation: UCAR, Boulder, Colorado.

Corresponding author address: Dr. R. Buizza, ECMWF, Shinfield Park, Reading RG2-9AX, United Kingdom.
E-mail: Buizza@ecmwf.int

At the National Centers for Environmental Prediction [NCEP; formerly the National Meteorological Center (NMC)], Toth and Kalnay (1993) introduced the bred-vector (BV) perturbation method. This method, discussed later in more detail, is based on the argument that fast-growing perturbations develop naturally in a data assimilation cycle and will continue to grow as short- and medium-range forecast errors. A similar strategy has been used at the European Centre for Medium-Range Weather Forecasts (ECMWF). Instead of bred vectors, however, ECMWF uses a singular vector (SV)–based method to identify the directions of fastest growth (Buizza and Palmer 1995; Molteni et al. 1996). Singular vectors maximize growth over a finite time interval and are consequently expected to dominate forecast errors at the end of that interval and possibly beyond. Instead of using a selective sampling procedure, the approach developed at the Meteorological Service of Canada (MSC) by Houtekamer et al. (1996a) generates initial conditions by assimilating randomly perturbed observations, using different model versions in a number of independent data assimilation cycles. This Monte Carlo–like procedure is referred to here as the perturbed-observation (PO) approach.

Quantitative comparisons of the bred-vector-, singular-vector-, and perturbed-observation-based ensembles have been so far performed in simplified environments only. Houtekamer and Derome (1995) compared the different strategies of ensemble prediction in a simplified environment consisting of a simulated observational network and the three-level quasigeostrophic T21 model of Marshall and Molteni (1993). They compared the quality of the ensemble mean forecasts and found that, although the basic concepts of the three ensemble prediction methods were rather different, the results were quite comparable. They recommended the use of bred-vector ensembles because of the relative ease of their implementation. The results from this and other simple model experiments (see, e.g., Hamill et al. 2000), however, are difficult to generalize since it is hard to know if all factors that are important for operational forecasts have been accounted for properly. Therefore a comparative analysis of the actual forecasts generated by the three operational systems is desirable for planning future developments.

Forecast errors in real-world applications arise not only because of initial errors, but also because of the use of imperfect models. Representing forecast uncertainty related to the use of imperfect models is thought to be of an even greater challenge than simulating initial-value-related errors. As described in the next section, the three centers follow rather different approaches in this respect as well.

For a better understanding of the differences and similarities between them, the main characteristics of the ensemble systems operational in 2002 at ECMWF, MSC, and NCEP are presented in section 2. The performance of the three ensemble systems are then quantitatively compared in section 3 for a 3-month period (May–June–July 2002), with an attempt to highlight how the different designs lead to different performance characteristics of the ensemble forecasts. It should be noted that for ease of comparison the quantitative analysis is based on a subset of the ensemble systems that includes only 10 perturbed and 1 unperturbed member starting at 0000 UTC. Possible future directions are discussed in section 4 and conclusions are drawn in section 5.

2. Ensemble prediction at ECMWF, MSC, and NCEP

Schematically, the main sources of forecast errors can be classified as follows:

- observations (incomplete data coverage, representativeness errors, measurement errors);
- models (errors due to, e.g., the parameterization of physical processes, the choice of closure approximations, and the effect of unresolved scales);
- data assimilation procedures (errors due to, e.g., the use of a background covariance model that assumes isotropy and the lack of knowledge of the background errors);
- imperfect boundary conditions (e.g., errors due to the imperfect estimation and description of roughness length, soil moisture, snow cover, vegetation properties, and sea surface temperature).

Formally, an ensemble forecast system is represented by a set of numerical integrations

$$e_j(T) = e_j(0) + \int_{t=0}^{T} [P_j(e_j, t) + A_j(e_j, t)]\, dt, \qquad (1)$$

where $P_j(e_j, t)$ denotes the model tendency due to parameterized physical processes (turbulence, moist processes, orographic effect) as used for member j; $A_j(e_j, t)$ denotes the tendency due to the other simulated processes (pressure gradient force, Coriolis, horizontal diffusion); and $e_j(0)$ is the initial state.

In the MSC Monte Carlo approach, initial perturbations are generated by running separate data assimilation cycles:

$$e_j(0) = \Xi[e_j(\tau_1), o(\tau_1, \tau_2) + \delta o_j, P_j, A_j], \qquad (2)$$

where (τ_1, τ_2) is the time spanned during each assimilation cycle, $o(\tau_1, \tau_2)$ and δo_j denote the vector of observations and corresponding random perturbations, and $\Xi[\ldots, \ldots, \ldots]$ denotes the data assimilation process. Note that each assimilation cycle depends on the model used in the assimilation.

In contrast, NCEP and ECMWF initial ensemble states are created by adding either bred or singular vectors $de_j(0)$ to the best estimate of the atmosphere at an initial time $e_0(0)$ that is produced by a high-resolution three- or four-dimensional data assimilation procedure:

TABLE 1. Summary of ensemble characteristics as of Jul 2002.

	MSC	ECMWF	NCEP
P_j (model uncertainty)	2 models + different physical parameterizations	$P_j = P_0$ (single model)	$P_j = P_0$ (single model)
dP_j (random model error)		$dP_j = r_j \times P_j$ (stochastic physics)	$dP_j = 0$
A_j	2 models	$A_j = A_0$ (single model)	$A_j = A_0$ (single model)
o_j (observation error)	Random perturbations	—	—
e_j (initial uncertainty)	e_j from analysis cycles	$e_j = e_0 + de_j$(SV)	$e_j = e_0 + de_j$(BV)
Horizontal resolution HRES forecast	100 km	—	T170(d0–7.5) > T126(d7.5–16)
Horizontal resolution control forecast	T_L149; not available for this study	T_L255 (d0–10)	T126(d0–3.5) > T62(d3.5–16)
Horizontal resolution perturbed members	T_L149	T_L255 (d0–10)	T126(d0–3.5) > T62(d3.5–16)
Vertical levels (control and perturbed members)	23 and 41, 28	40	28
Top of the model	10 hPa	10 hPa	3 hPa
No. of perturbed members	16	50	10
Forecast length	10 days	10 days	16 days
Daily frequency	0000 UTC	1200 UTC	0000 and 1200 UTC
Operational implementation	Feb 1998	Dec 1992	Dec 1992

$$e_0(0) = \Xi[e_0(\tau_1), o(\tau_1, \tau_2), P_0, A_0], \tag{3a}$$

$$e_j(0) = e_0(0) + de_j(0). \tag{3b}$$

The main characteristics of the three global operational systems as of summer 2002 are summarized in Table 1 and are further discussed below.

a. The singular-vector approach at ECMWF

The ECMWF SV approach (Buizza and Palmer 1995; Molteni et al. 1996) is based on the observation that perturbations pointing along different axes of the phase space of the system are characterized by different amplification rates. Given an initial uncertainty, perturbations along the directions of maximum growth amplify more than those along other directions. For defining the SVs used in the ECMWF Ensemble Prediction System (EC-EPS), growth is measured by a metric based on total energy norm. The SVs are computed by solving an eigenvalue problem defined by an operator that is a combination of the tangent forward and adjoint model versions integrated during a time period named the optimization time interval. The advantage of using singular vectors is that if the forecast error evolves linearly and the proper initial norm is used, the resulting ensemble captures the largest amount of forecast-error variance at optimization time (Ehrendorfer and Tribbia 1997).

The EC-EPS has been part of the operational suite since December 1992. The first version, a 33-member T63L19 configuration (spectral triangular truncation T63 with 19 vertical levels; Palmer et al. 1993; Molteni et al. 1996) simulated the effect of initial uncertainties by the introduction of 32 perturbations that grow rapidly during the first 48 h of the forecast range. In 1996 the system was upgraded to a 51-member T_L159L31 system (spectral triangular truncation T159 with linear grid; Buizza et al. 1998). In March 1998, initial uncertainties due to perturbations that had grown during the 48 h prior to the starting time (evolved singular vectors; Barkmeijer et al. 1999) were also introduced. In October 1998, a scheme to simulate model uncertainties due to random model error in the parameterized physical processes was added (Buizza et al. 1999). In October 1999, following the increase of the number of vertical levels in the data assimilation and high-resolution deterministic model from 31 to 60, the number of vertical levels in the EPS was increased from 31 to 40. In November 2000, the EPS resolution was increased to T_L255L40 (Buizza et al. 2003), with initial conditions for the unperturbed forecast (the control) interpolated from the upgraded T_L511L60 analysis. This most recent upgrade coincided with an increase of resolution of the ECMWF data assimilation and high-resolution deterministic forecast from T_L319L60 to T_L511L60.

The 51*T_L255L40 EC-EPS included one "control" forecast started from the unperturbed analysis (interpolated to the lower ensemble resolution), and 50 additional forecasts started from perturbed analysis fields. These perturbed fields were generated by adding to/subtracting from the unperturbed analysis a combination of the dynamically fastest-growing perturbations (defined by the total energy as a measure of growth) computed at T42L40 resolution to optimize growth during the first 48 h of the forecast range, scaled to have an amplitude consistent with analysis error estimates. Three sets of fastest-growing perturbations are used in the ECMWF-EPS, located to have maximum growth in the Northern and Southern Hemisphere extratropics, and in the Tropics. Linear/adjoint moist processes are used to compute the tropical singular vectors (Barkmeijer et al. 2001), but the extratropical fastest-growing perturbations are still computed without linear/adjoint moist processes: work is in progress at ECMWF to change the linear/adjoint models to include them in the extratropical singular-vector computation (Coutinho et al. 2004). Since April 2003 the EPS has been running twice a day, with 0000 and 1200 UTC initial times.

Formally, each member of the EC-EPS is defined by Eq. (1) with the same model version

$$e_j(T) = e_j(0) + \int_{t=0}^{T} [P(e_j, t) + dP_j(e_j, t) + A(e_j, t)]\, dt, \tag{4a}$$

with randomly perturbed tendencies

$$dP_j[e_j(\lambda, \phi, t)] = \langle r_j(\lambda, \phi)\rangle_{10,6} \cdot P(e_j, t), \tag{4b}$$

where (λ, ϕ) are the gridpoint longitude and latitude, and $\langle \ldots \rangle_{10,6}$ indicates that the same random number r_j is used inside a 10° box and a 6-h time window (see Buizza et al. 1999 for more details). The initial perturbations $de_j(0)$ are defined as

$$de_j(0) = \underline{\underline{A}} \cdot SV_{\mathrm{NH}} + \underline{\underline{B}} \cdot SV_{\mathrm{SH}} + \underline{\underline{C}} \cdot SV_{\mathrm{TC}}, \tag{4c}$$

where for each geographical region [Northern and Southern Hemisphere extratropics (NH and SH, respectively) and Tropics (TC)] the coefficients of the linear combination matrices are set by comparing the singular vectors with analysis error estimates given by the ECMWF four-dimensional data assimilation (4DVAR) scheme (see Molteni et al. 1996 for more details). The Northern/Southern Hemisphere singular vectors are computed to maximize the day-2 total energy north/south of 30°/−30°N, while the tropical singular vectors are computed to maximize the day-2 total energy inside a region that includes any tropical disturbance present at the analysis time (Barkmeijer et al. 2001).

b. The MSC perturbed-observation approach

The MSC perturbed-observation approach attempts to obtain a representative ensemble of perturbations by comprehensively simulating the behavior of errors in the forecasting system. Sources of uncertainty that are deemed to be significant are sampled by means of random perturbations that are different for each member of the ensemble. Because the analysis and forecast process is repeated several times with different random input, the perturbed-observation method is a classic example of the Monte Carlo approach. Arguments for the use of nonselective, purely random ensemble perturbations are presented in Houtekamer et al. (1996b) and by Anderson (1997).

In the first version of the MSC-EPS, implemented operationally in February 1998, all eight members used the Spectral Finite Element model at resolution T_L95 (Ritchie and Beaudoin 1994) and an optimal interpolation data assimilation system (Mitchell et al. 1996). The members used different sets of perturbed observations, different versions of the model, and different subsets of perturbed surface fields.

Perturbing the observations is straightforward in principle. An estimate of the error statistics is available for each observation that is assimilated with the optimal interpolation method. Random numbers, with Gaussian distribution, can subsequently be obtained from these estimates using a random number generator. Here the Gaussian distribution has zero mean and error (co) variance as specified in the optimal interpolation scheme. It should be noted though that the resulting perturbations have subsequently been multiplied with a factor of 1.8 in order to inflate the ensemble spread and thus compensate for an insufficient representation of model error.

To account for model error, experts on the model physics were consulted as to what physical parameterizations were of similar quality (Houtekamer and Lefaivre 1997). The selected physical parameterizations were state-of-the-art at the time of the implementation of the MSC-EPS. In addition to the models and observations, the surface boundary conditions are also a source of errors, though perhaps less significant than the other two error sources. The associated uncertainty is represented in the MSC-EPS by adding time-constant random perturbation fields to the boundary fields of sea surface temperature, albedo, and roughness length.

In August 1999, the size of the MSC-EPS was doubled to 16 members. Since then, the eight additional members have been generated using the newly developed Global Environmental Multiscale (GEM) model (Côté et al. 1998). Furthermore, updated versions of physical parameterizations have been used for the eight members that use the GEM model. The use of two different dynamical models led to a much better sampling of the model-error component. Improvement was noted in particular in the spread/skill correlation and the rank histograms for 500-hPa geopotential (not shown).

In 2001, it became possible to increase the horizontal resolution (Pellerin et al. 2003). The spectral resolution of the eight members that use the Spectral Finite Element model was increased from T_L95 to T_L149, and the resolution of the eight members that use the GEM model increased from a 2° to a 1.2° uniform grid. This was possible because of an increase in computational resources at MSC.

Note that no additional data assimilation cycles are run for the new members introduced in 1999: instead, the eight additional initial conditions for the medium-range forecasts are obtained by means of a correction (Houtekamer and Lefaivre 1997) toward the operational deterministic high-resolution 3D variational analysis of the Canadian Meteorological Centre (CMC) (Gauthier et al. 1999). Since the high-resolution analysis is of higher quality than the lower-resolution ensemble mean optimal interpolation analysis, the correction is such that the 16-member initial ensemble mean state is a weighted mean of the high-resolution analysis and the original 8-member ensemble mean analysis.

It should be noted that the relative weights of the low-resolution ensemble mean and the high-resolution deterministic analysis were determined at a time when both analyses were performed with an optimal interpolation procedure. Since then the low-resolution ensemble mean analyses have been obtained with a fairly stable configuration whereas the deterministic analysis improved significantly. The way in which these analyses are combined is due for a reevaluation.

One of the difficulties of the MSC-EPS approach is that a significant manpower investment is required to operationally maintain the system at a state-of-the-art level, since this involves a continuous reevaluation, adjustment, correction, and replacement of data assimilation and modeling algorithms by more suitable or acceptable procedures. It is more difficult to maintain a multimodel ensemble, especially during periods of hardware replacements.

c. The NCEP bred-vector approach

The NCEP bred-vector approach is based on the notion that analysis fields generated by data assimilation schemes that use NWP models to dynamically propagate information about the state of the system in space and time will accumulate growing errors by the virtue of perturbation dynamics (Toth and Kalnay 1993, 1997). For example, based on 4DVAR experiments with a simple model, Pires et al. (1996) concluded that in advanced data assimilation systems the errors at the end of the assimilation period are concentrated along the fastest-growing Lyapunov vectors. This is due to the fact that neutral or decaying errors detected by an assimilation scheme in the early part of the assimilation window will be reduced, and what remains of them will decay due to the dynamics of such perturbations by the end of the assimilation window. In contrast, even if growing errors are reduced by the assimilation system, what remains of them will, by definition, amplify by the end of the assimilation window. These findings have been confirmed in a series of studies using assimilation and forecast systems of varying complexity (for a summary, see Toth et al. 1999).

The breeding method involves the maintenance and cycling of perturbation fields that develop between two numerical model integrations, practically amounting to the use of a "virtual" nonlinear perturbation model. When its original form is used with a single global rescaling factor, the BVs represent a nonlinear extension of the Lyapunov vectors (Boffetta et al. 1998). For ensemble applications, the bred vectors are rescaled in a smooth fashion to follow the geographically varying level of estimated analysis uncertainty (Iyengar et al. 1996). In NCEP operations, multiple breeding cycles are used, each initialized at the time of implementation with independent arbitrary perturbation fields ("seeds").

The perturbed-observation and the bred-vector methods are related in that they both aim at providing a random sample of analysis errors. One difference is that while the perturbed-observation method works in the full space of analysis errors, the bred-vector method attempts to sample only the small subspace of the fastest-growing errors. The bred-vector approach is also related to the singular-vector approach followed at ECMWF in that both methods aim at sampling the fastest-growing forecast errors. The difference between these two methods is that while the breeding technique attempts to provide a random sample of growing analysis errors, the singular vectors give a selective sample of perturbations that can produce the fastest linear growth in the future.

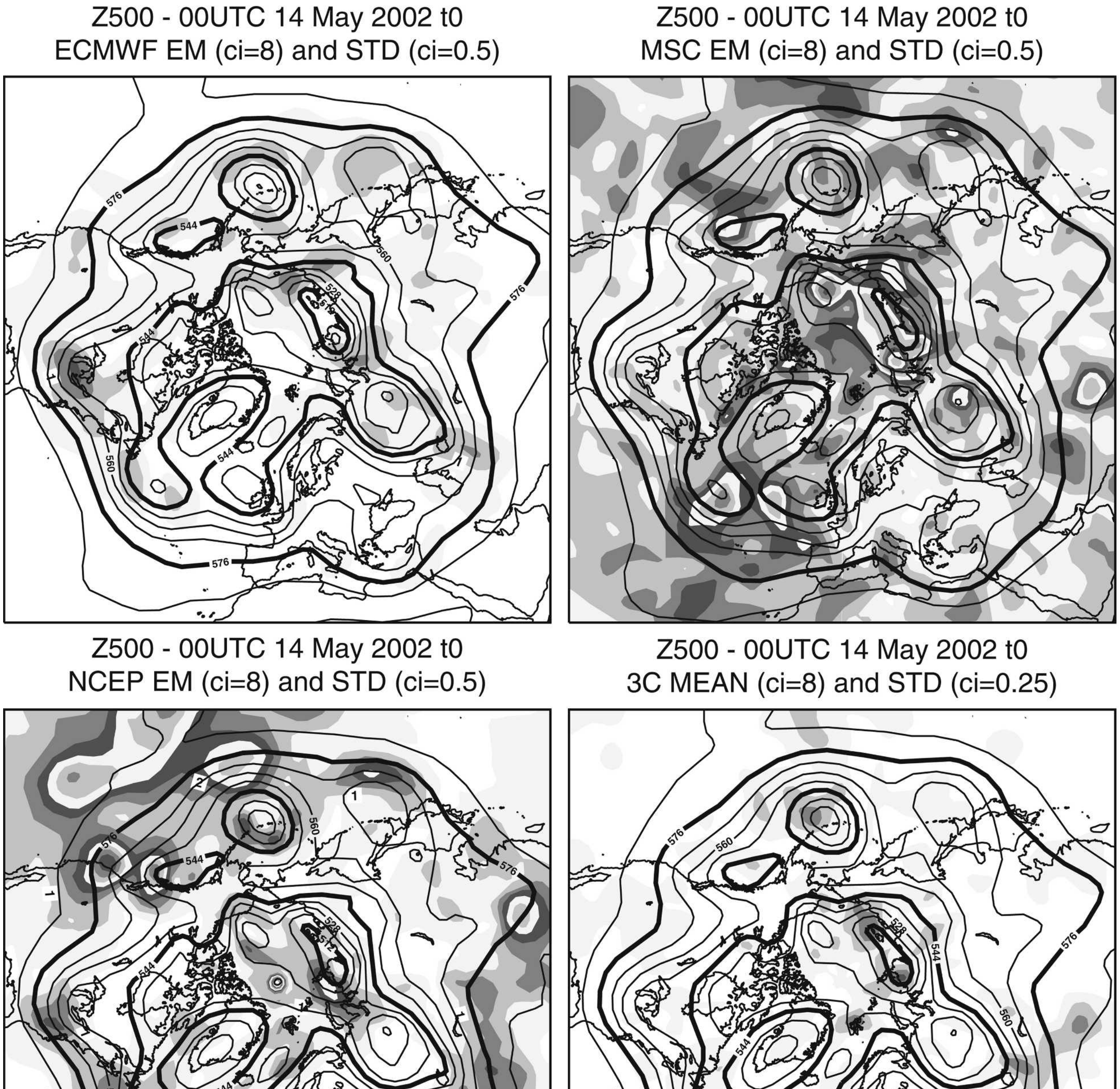
The use of closure schemes in NWP models result in random model errors that behave dynamically like initial-value-related errors (Toth and Vannitsem 2002). These random model errors are simulated in the NCEP ensemble in a crude fashion by setting the size of the initial perturbations at a level somewhat higher than the estimated uncertainty present in the analysis fields. While the larger initial spread in the NCEP ensemble slightly hinders performance in the short lead-time ranges, it improves performance in the medium- and extended ranges (Toth and Kalnay 1997).

NCEP produces 10 perturbed ensemble members both at 0000 and 1200 UTC every day out to 16-days lead time. For both cycles, the generation of the initial perturbations is done in five independently run breeding cycles, originally started with different arbitrary perturbations, using the regional rescaling algorithm. The initial perturbations are centered as positive–negative pairs around the operational high resolution (at the time of the study period, T170L42) NCEP analysis field, truncated to T126L28 (Toth et al. 2002; Caplan et al. 1997). The ensemble forecasts are integrated at this spatial resolution out to 84 h, at which point the forecasts are truncated to, and for computational efficiency integrated at a lower, T62L28 resolution. For both cycles, the ensemble forecasts were complemented by a higher-resolution control forecast (T170L42 up to 180 h, then truncated to T62L28) started from the high-resolution operational analysis. At 0000 UTC, a second control forecast with the same spatial resolution as the perturbed forecasts is also generated.

Formally, each member of the NCEP-EPS is defined by Eq. (1) with the same model version P being used for all members and with initial perturbations $de_j(0)$ defined as

$$de_j(0) = \underline{\underline{\underline{RR}}} \cdot BV_j. \tag{5}$$

The coefficients $\underline{\underline{\underline{RR}}}$ of the linear combination matrices in Eq. (5) are defined by the regional rescaling algorithm (Toth and Kalnay 1997).



Z500 - 00UTC 14 May 2002 t0
ECMWF EM (ci=8) and STD (ci=0.5)

Z500 - 00UTC 14 May 2002 t0
MSC EM (ci=8) and STD (ci=0.5)

Z500 - 00UTC 14 May 2002 t0
NCEP EM (ci=8) and STD (ci=0.5)

Z500 - 00UTC 14 May 2002 t0
3C MEAN (ci=8) and STD (ci=0.25)

FIG. 1. Initial state, 0000 UTC on 14 May 2002, at 500-hPa geopotential height. Ensemble mean and standard deviation (shading) of the (a) EC-EPS, (b) MSC-EPS, and (c) NCEP-EPS. (d) Average of the three ensemble means and standard deviation among the three ensemble means (shading). Contour interval is 8 dam for full field, 0.5 dam for ensemble standard deviation in (a)–(c), and 0.25 dam for standard deviation in (d).

d. An example: The forecast case of 14 May 2003

Before discussing their average properties, the effect of the use of different configurations on the characteristics of the three ensemble systems is illustrated by discussing a forecast case. Figures 1, 2, and 3 show the ensemble mean and standard deviation (which is a measure of the ensemble spread) for the 500-hPa geopotential height for a randomly selected initial date (14

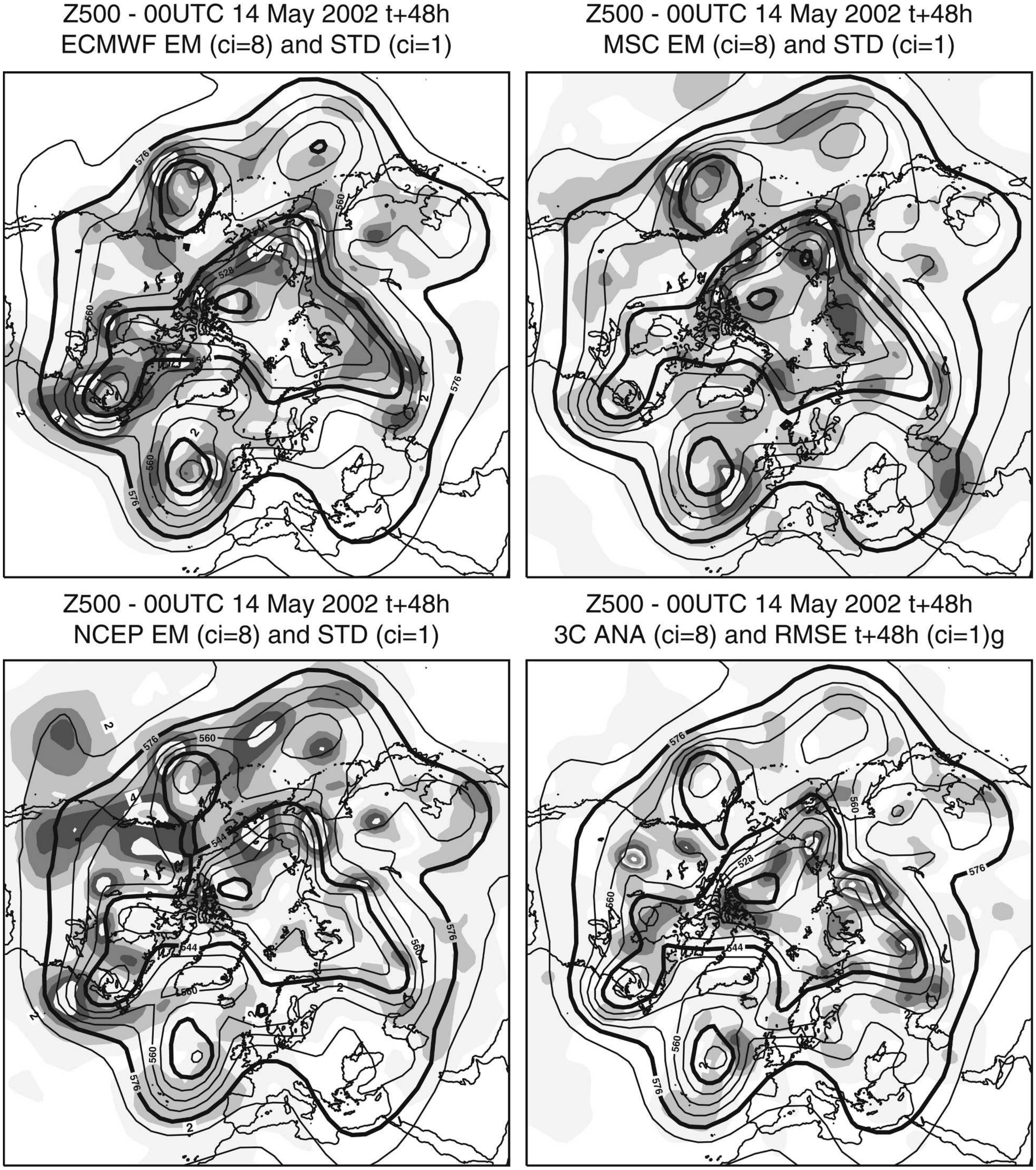
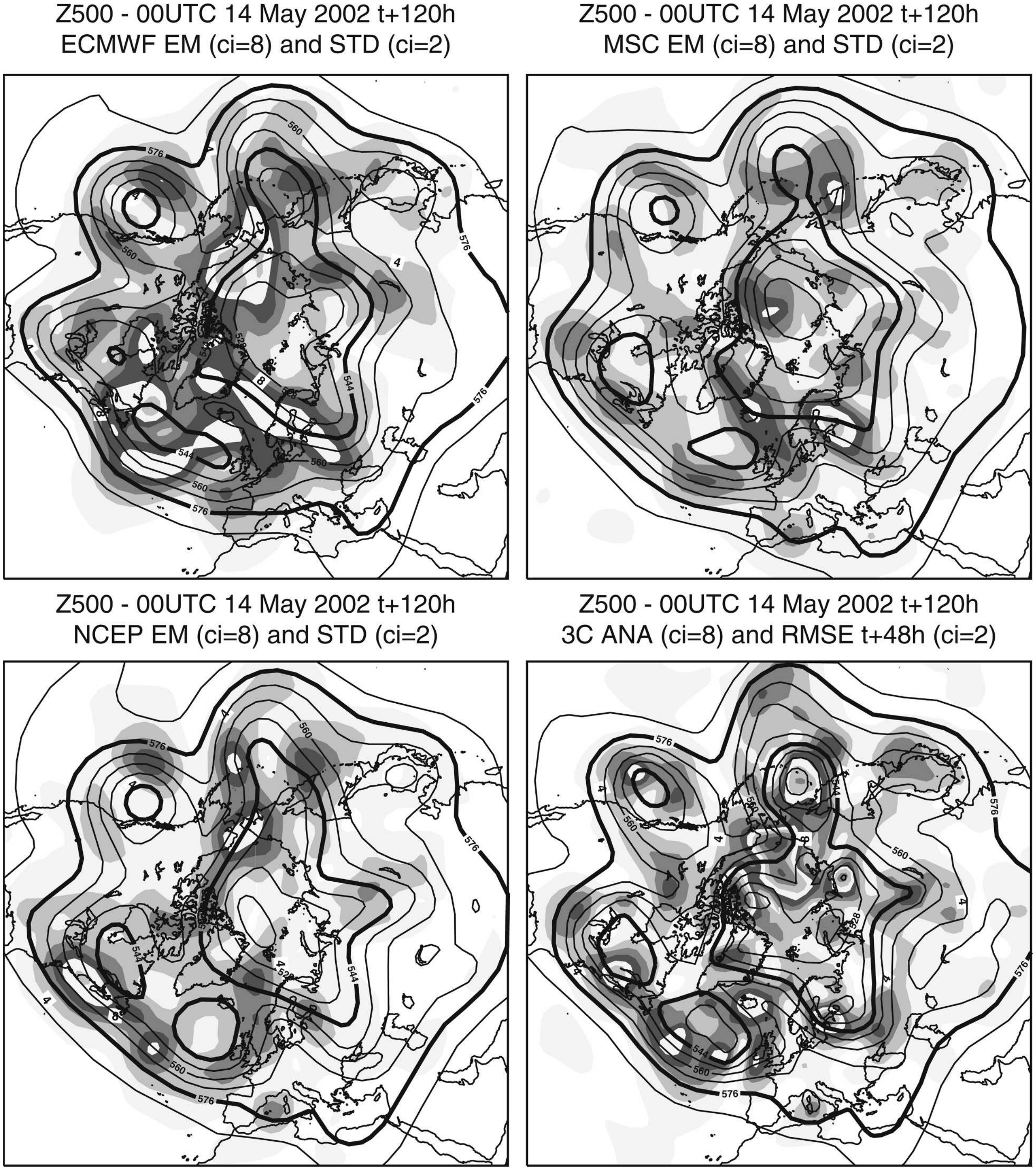

Z500 - 00UTC 14 May 2002 t+48h
ECMWF EM (ci=8) and STD (ci=1)
Z500 - 00UTC 14 May 2002 t+48h
MSC EM (ci=8) and STD (ci=1)
Z500 - 00UTC 14 May 2002 t+48h
NCEP EM (ci=8) and STD (ci=1)
Z500 - 00UTC 14 May 2002 t+48h
3C ANA (ci=8) and RMSE t+48h (ci=1)g
576
560
544
528
2
4

FIG. 2. The 48-h forecast from 0000 UTC on 14 May 2002, at 500-hPa geopotential height. Ensemble mean and standard deviation (shading) of the (a) EC-EPS, (b) MSC-EPS, and (c) NCEP-EPS. (d) Average (of the three centers) ensemble mean and average ensemble mean error (shading). Contour interval is 8 dam for full field and 1 dam for ensemble standard deviations.

May 2002), along with the $t + 48$ h and the $t + 120$ h forecasts. To compare equally populated ensembles, only 10 members from each center are used. Each ensemble is verified against the analysis from the originating center.

At initial time, the spread among the three centers' initial states (measured by the standard deviation of three centers' ensemble means) is also shown (Fig. 1d). This field can be considered as a crude lower-bound estimate of analysis error variance, providing a refer-



Z500 - 00UTC 14 May 2002 t+120h
ECMWF EM (ci=8) and STD (ci=2)

Z500 - 00UTC 14 May 2002 t+120h
MSC EM (ci=8) and STD (ci=2)

Z500 - 00UTC 14 May 2002 t+120h
NCEP EM (ci=8) and STD (ci=2)

Z500 - 00UTC 14 May 2002 t+120h
3C ANA (ci=8) and RMSE t+48h (ci=2)

FIG. 3. The 120-h forecast from 0000 UTC on 14 May 2002, at 500-hPa geopotential height. Ensemble mean and standard deviation (shading) of the (a) EC-EPS, (b) MSC-EPS, and (c) NCEP-EPS. (d) Average (of the three centers) ensemble mean and average ensemble mean error (shading). Contour interval is 8 dam for full field and 2 dam for ensemble standard deviations.

ence for ensemble spread. Since at initial time the ensemble perturbations are designed to represent analysis errors, the ensemble spread should, on average, be similar to analysis error variance. Figure 1 shows that the three ensembles emphasize different geographical regions. The EC-EPS (Fig. 1a) has the smallest initial spread, falling closest in amplitude to the spread among the three centers' initial states (Fig. 1d; note the lower contour interval compared to other panels). Note that the EC-EPS spread over the Northern Hemisphere south of 30°N is almost zero, since SV perturbations in the tropical region are generated only in the vicinity of tropical cyclones (see section 2a).

In the case of the forecasts, the average error of the three centers' ensemble mean forecasts (i.e., average distance of the respective analyses from the ensemble mean forecasts) is used as a reference field (Figs. 2d and 3d). Since the ensemble forecasts are supposed to include the verification as a possible solution, the ensemble spread should, on average, be similar to this field. At t + 48 h (Fig. 2), the spread in the three centers' ensembles is more similar to each other than at initial time, both in terms of amplitude and pattern. At t + 120 h (Fig. 3), the three ensembles continue to have a similar pattern of spread, with a slightly larger spread in the EC-EPS than in the others.

3. Comparative verification of the three ensemble systems

a. Earlier studies

A number of studies have been performed comparing the performance of former versions of the EC- and the NCEP-EPSs. In the first of these studies Zhu et al. (1996) compared the performance of the EC- and NCEP-EPS at a time (winter of 1995/96) when the spatial resolution of the ensemble forecasts (T62 versus T63) and the skill of the control forecasts at the two centers were rather similar. Using a variety of verification measures similar to those used in this study, they concluded that the NCEP-EPS 500-hPa geopotential height forecasts had a 0.5–1-day advantage in skill during the first 5 days, while the EC-EPS became comparable or superior by the end of the 10-day forecast period.

Subsequently, the ECMWF ensemble always had a markedly higher resolution (and superior control forecast performance) than the NCEP ensemble. Despite this difference in horizontal model resolution, in a follow-up study Atger (1999) found the statistical resolution (see, e.g., Wilks 1995) of the NCEP and ECMWF ensembles comparable. In this comparison, the ECMWF ensemble had an advantage in terms of statistical reliability [a reliable forecast is statistically indistinguishable from the corresponding sample of observations (Wilks 1995; Stanski et al. 1989); see also the discussion in section 3c(1)] because of its larger spread, which guaranteed a better agreement between the growth of ensemble spread measured, for example, by the ensemble standard deviation, and the growth of the ensemble mean error. Mullen and Buizza (2001) compared the skill of probabilistic precipitation forecasts based on 10 perturbed members over the United States, using 24-h accumulated precipitation data from the U.S. National Weather Service (NWS) River Forecast Centers. They concluded that during 1998 the limit of skill [measured by the Brier skill score (BSS)] for the EC-EPS was about 2 days longer for the 2 and 10 mm day^{-1} thresholds, while the two systems exhibited similar skill for 20 mm day^{-1} amounts. The results of Atger (2001) confirmed that the EC-EPS performed better than the NCEP-EPS in terms of precipitation forecasts, based on verification against rain gauge observations in France.

These studies indicate that the relative performance of the different ensemble systems depends on their actual operational configuration and the time period, variables, and verification measures used. The current study offers a recent snapshot of the performance of the three systems compared and has the same limitations as previous work. Results for other seasons have not been carefully studied. We note that work is in progress toward establishing a continuous comparison effort based on a fairly extensive set of measures as discussed below.

b. Verification database: May–June–July 2002

In this study the performance of the three ensemble forecast systems is assessed for a 3-month period for which data from the three ensembles were available and exchanged, May–June–July 2002. Since NCEP generates only 10 perturbed forecasts from each initial time, the comparison has been limited to 10-member ensembles. When considering the quantitative results of this study, the reader should be aware that ensemble size has an impact on ensemble skill: for example, Buizza and Palmer (1998) concluded that increasing the ensemble size from 8 to 32 members in the old T63L19 EC-EPS system increased the skill of the ensemble mean by ~6 h in the medium range and the skill of probabilistic predictions by ~12 h. It should be noted that the initial conditions of the ECMWF 10 perturbed members have been generated using 25 (and not only 5) singular vectors: using only 5 instead of 25 singular vectors would have generated more localized initial perturbations, and thus most probably would have reduced the performance of the ECMWF 10-member ensemble. The subsampling of 10 from the 16-member MSC ensemble could have a negative impact on the MSC results because of a displacing of the ensemble mean from its original position.

Since in May–June–July 2002 ECMWF had no operational 0000 UTC ensemble and MSC had no operational 1200 UTC ensemble, for each day 0000 UTC MSC- and NCEP-EPS and the 1200 UTC ECMWF ensembles have been considered.

For brevity, only 500-hPa geopotential height forecasts are considered over the middle latitudes of the Northern Hemisphere [20°–80°N, except for one measure, perturbation versus error correlation analysis (PECA); see below). This choice has been dictated by three main reasons: the geopotential height at 500 hPa is one of the most used weather fields, it gives a useful view of the synoptic-scale flow, and it was one of the very few fields available for the comparison project.

Forecast and analysis fields have been interpolated onto a common regular 2.5 × 2.5 grid, and each ensemble has been verified against its own analysis, that is, the analysis generated by the same center. Probabilistic forecasts are generated and evaluated in terms of 10 climatologically equally likely intervals determined at each grid point separately (Toth et al. 2002), based on the NCEP–National Center for Atmospheric Research (NCAR) reanalysis.

c. *Verification attributes and measures*

1) ATTRIBUTES OF FORECAST SYSTEMS

The performance of the ensemble forecast systems is assessed considering three attributes of a forecasting system (Murphy 1973): statistical reliability (or consistency), resolution, and discrimination. Statistical reliability implies that a sample of forecasts is statistically indistinguishable from the corresponding sample of observations (or analysis fields). Reliability can often be improved through simple statistical postprocessing techniques. Though important for real-world applications, reliability of a forecast system in itself does not guarantee usefulness (e.g., a climatological forecast system, by definition, is perfectly reliable, yet has no forecast value). Statistical resolution reflects a forecast system's ability to distinguish between different future events in advance. Discrimination, which is the converse of resolution (Wilks 1995), reflects a system's ability to distinguish between the occurrence and nonoccurrence of forecast events. In case observed frequencies of forecast events monotonically increase with increasing forecast probabilities, resolution and discrimination—which are based on two different factorizations of the forecast/observed pair of events—convey the same information about forecast systems.

2) VERIFICATION MEASURES

Different measures, emphasizing different aspects of forecast performance, can be used to assess the statistical reliability, resolution, and discrimination of a forecast system. In this study, the performance of the three EPSs will be compared using a comprehensive set of standard ensemble and probabilistic forecast verification methods, including the pattern anomaly correlation (PAC), root-mean-square (rms) error, the Brier (1950) skill score, the outlier statistics (a measure of reliability), and the area under the relative operating characteristics (ROCs; a measure of discrimination; Mason 1982). The reader is referred to, for example, Stanski et al. (1989), Wilks (1995), Talagrand et al. (1997), and Toth et al. (2003) for a description of these scores.

The above scores measure the quality of probabilistic forecasts of scalar quantities. In the context of this study, one would also like to evaluate the relevance of perturbation patterns. The characteristics of the patterns could be very different for the three EPS systems. To investigate this, Wei and Toth (2003) designed a new measure called PECA. By evaluating how much of the error in a forecast can be explained by a single, or an optimal combination of ensemble perturbations, PECA ignores the magnitude of forecast errors that may dominate other verification measures. Therefore the PECA values shown in the next subsection may be helpful in attributing the ensemble performance results to differences in the quality of data assimilation, NWP modeling, and ensemble perturbation techniques at the three centers.

d. *Performance of the three ensemble systems for May–June–July 2002*

1) QUALITY OF DATA ASSIMILATION AND NUMERICAL MODELING SYSTEMS

Since the performance of the ensemble forecast systems is affected not only by the ensemble generation schemes but also by the quality of the data assimilation and forecast procedures used, it will be useful to first compare the performance of single forecasts started from the best analysis available at each center ("control" forecasts). This can serve as a reference reflecting the quality of data assimilation and NWP modeling at the three centers. Shown in Fig. 4 is the PAC score for each center's control forecast. Note that both ECMWF and NCEP have a control forecast that is run at the same model resolution as the respective perturbed ensemble members (note that this resolution is different at the three centers), started from the unperturbed initial condition. Because of communication problems, such an equal resolution control forecast from the MSC-EPS was not available for this comparison. In its place the skill of the MSC high-resolution control forecast, started from the operational three-dimensional variational data assimilation (3DVAR) analysis, is shown in Fig. 4. For the period under investigation, results indicate that the quality of the control forecast is highest for the EC-EPS and lowest for the MSC-EPS.

2) OVERALL MEASURES OF ENSEMBLE PERFORMANCE

Rms error, and the related PAC, are influenced by both systematic errors (such as a low bias in ensemble spread, degrading reliability) and random error variance (reducing a forecast system's ability to distinguish among different events, leading to reduced resolution). Therefore these two scores offer good measures of overall forecast performance. In these subsections the accuracy of each ensemble forecast system is measured by PAC and rms of the ensemble mean forecasts. For PAC, the ensemble skill is also compared to that of the control forecasts. These scores are complemented by

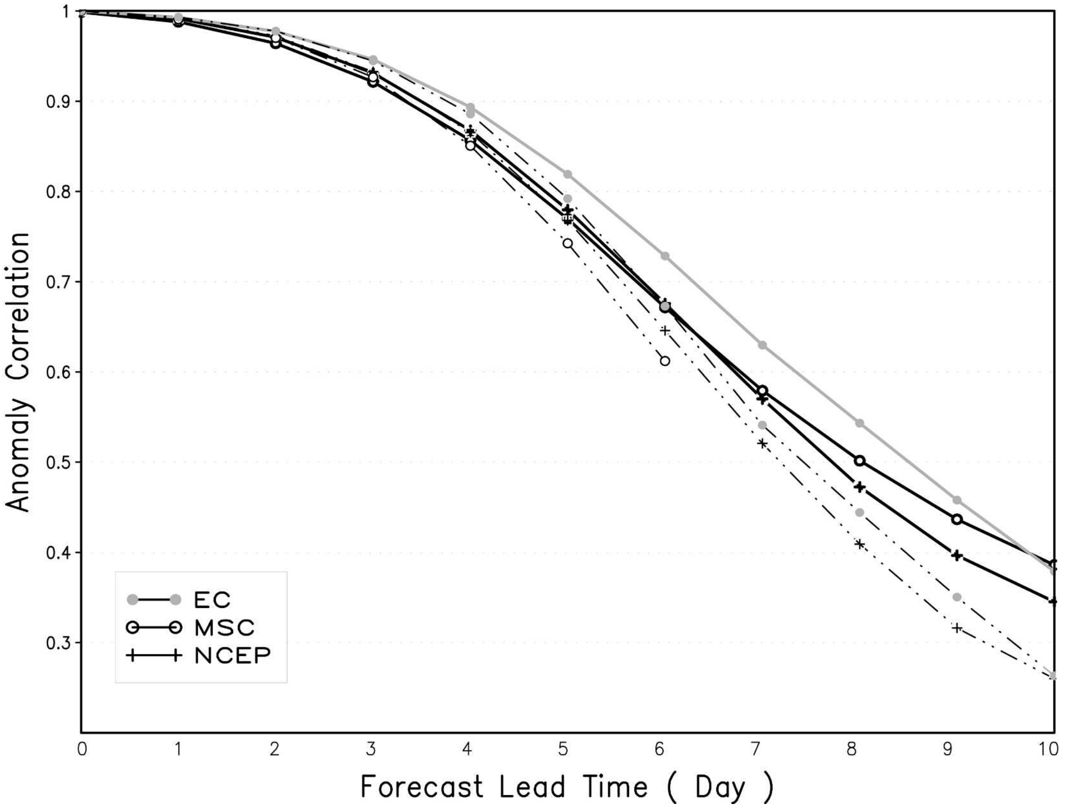


FIG. 4. May–Jun–Jul 2002 average PAC for the control (dotted lines) and the ensemble mean (solid lines) of the EC-EPS (gray lines with full circles), the MSC-EPS (black lines with open circles), and the NCEP-EPS (black lines with crosses). Values refer to the 500-hPa geopotential height over the Northern Hemisphere latitudinal band 20°–80°N.

the Brier skill score computed for probabilistic forecasts based on the three ensembles.

Except as noted below, each ensemble mean forecast is more skillful than its control in terms of PAC (see Fig. 4). The gain in predictability from running an ensemble (instead of a single control forecast) is about 12/24 h at forecast day 6/9. These gains are due to the nonlinear filtering effect that ensemble averaging offers in terms of error growth reduction (Toth and Kalnay 1997). For the first few days, the MSC control forecast has higher skill than the MSC-EPS mean. Most likely this is due to the MSC-EPS being centered on an initial state that is inferior to the 3DVAR analysis and to the subsampling to 10 members performed for this study (see section 3b). Note also that beyond day 5, the gain from ensemble averaging is smallest in the NCEP-EPS: this may be related to the lack of explicit representation of model errors in that ensemble.

Given the earlier finding that the ECMWF forecast system has the best overall data assimilation/modeling components, it is not surprising that the ensemble mean for the EC-EPS also performs better than those for the other centers, both in terms of PAC (Fig. 4) and rms error (Fig. 5). Note also that by the end of the 10-day forecast period the performance of the EC- and the MSC-EPS become very similar. This may be due to the beneficial effect of using different model versions in the MSC-EPS in terms of the rms error and PAC measures.

The BSS, shown in the top panel of Fig. 6, is computed by averaging the BSS for 10 climatologically equally likely events, considering climatology as a reference forecast. Just like the rms error and PAC, the BSS reflects both the reliability and resolution of ensemble forecast systems. The BSS can be decomposed into its reliability and resolution components (Murphy 1973; Toth et al 2003; see the appendix for details).

Results from the lower panel of Fig. 6 indicate that at shorter times (before day 6) it is the resolution, while at longer times it is the reliability term of the BSS that dominates the overall result. Not surprisingly, the BSS results are somewhat similar to those presented for the rms error in Fig. 5. Overall, the best performance is obtained by the ECMWF ensemble. During the first few days, the NCEP system remains competitive, suggesting perhaps a positive effect of the initial perturbations (bred vectors). At longer lead times the performance of the NCEP system slips, probably because of the lack of model perturbations. The relatively good performance of the MSC system at long (8–10 days)

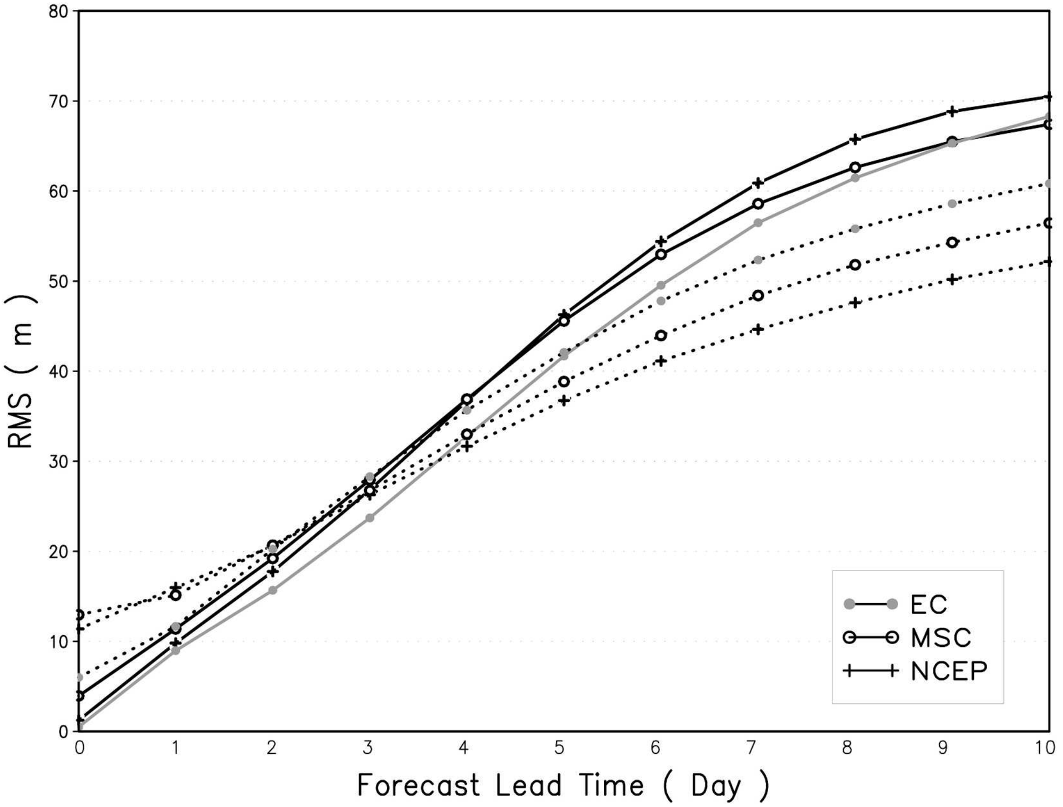


FIG. 5. May–Jun–Jul 2002 average rms error of the ensemble mean (solid lines) and ensemble standard deviation (dotted lines) of the EC-EPS (gray lines with full circles), the MSC-EPS (black lines with open circles), and the NCEP-EPS (black lines with crosses). Values refer to the 500-hPa geopotential height over the Northern Hemisphere latitudinal band 20°–80°N.

lead time again may be due to the use of multiple model versions in that ensemble.

3) MEASURES OF RELIABILITY

In this subsection statistical reliability is assessed in three different ways. The first measure used is the discrepancy between the ensemble spread and the error of the ensemble mean, both shown for all three systems in Fig. 5. For a statistically reliable ensemble system, reality should statistically be indistinguishable from the ensemble forecasts. It follows that the distance between the ensemble mean and the verifying analysis (error of the ensemble mean) should match that between the ensemble mean and a randomly selected ensemble member (ensemble standard deviation or spread). A large difference between the error of the ensemble mean and the ensemble standard deviation is therefore an indication of statistical inconsistency.

As seen from Fig. 5, the growth of the rms error exceeds that of the spread for all three systems (except as noted below). The growth of ensemble perturbations (spread) in the three systems is affected by two factors: the initial ensemble perturbations, and the characteristics of the model (or model versions) used. While the initial perturbations are important during the first few days, their influence diminishes with increasing lead time since the perturbations rotate toward directions that expand most rapidly due to the dynamics of the atmospheric flow (as represented in a somewhat different manner in each model), as discussed in relation with Figs. 1–3.

Out of the three systems the EC-EPS exhibits the largest (and therefore most realistic) perturbation growth. An important observation based on Fig. 5 is that the perturbations' growth is lower than the error growth in the MSC- and NCEP-EPS: this deficiency in perturbation growth is partially compensated by initial perturbation amplitudes that are larger than the level of estimated initial errors. Because of the use of a purely Monte Carlo perturbation technique that generates initial perturbations containing neutral and decaying modes, the MSC-EPS exhibits the lowest perturbation growth during the first day of integration. After the first day, the NCEP-EPS exhibits the lowest (and least realistic) perturbation growth. Most likely this is due to the lack of model perturbations in that ensemble.

The relatively larger growth rate of the EC-EPS in the 3–10-day range is due partly to the sustained growth

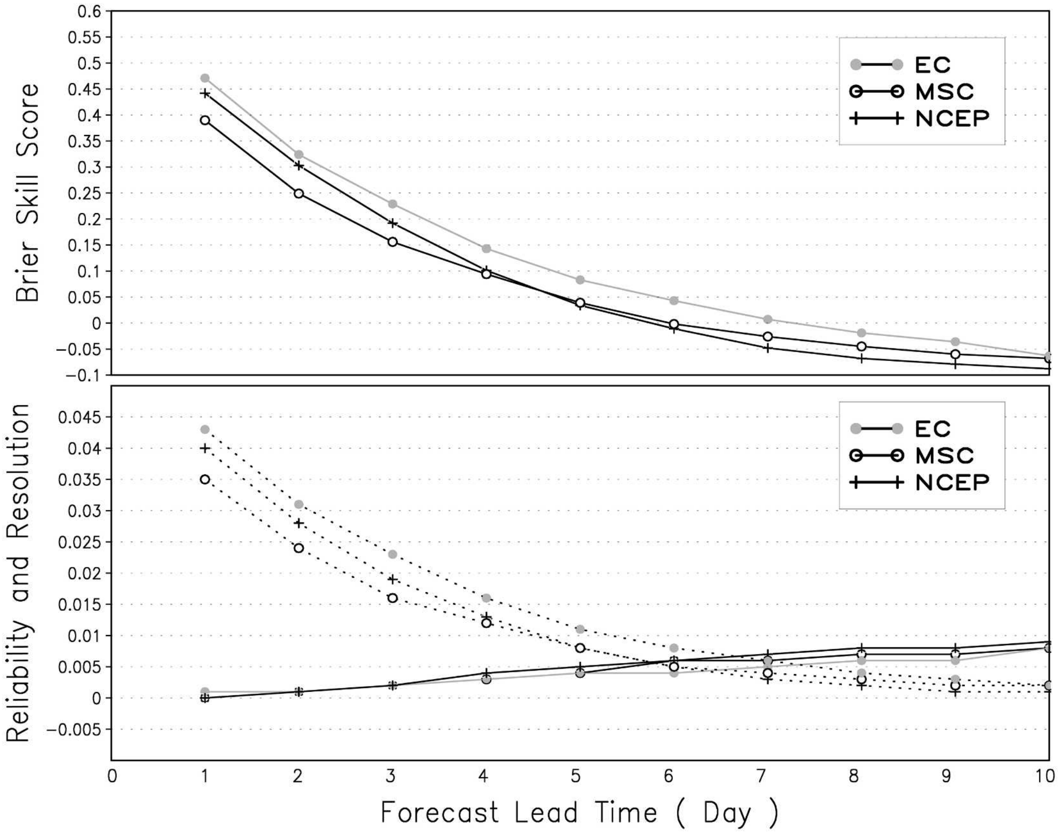


FIG. 6. (top) May–Jun–Jul 2002 average Brier skill score for the EC-EPS (gray lines with full circles), the MSC-EPS (black lines with open circles), and the NCEP-EPS (black lines with crosses). (bottom) Resolution (dotted) and reliability (solid) contributions to the Brier skill score. Values refer to the 500-hPa geopotential height over the Northern Hemisphere latitudinal band 20°–80°N and have been computed considering 10 equally climatologically likely intervals.

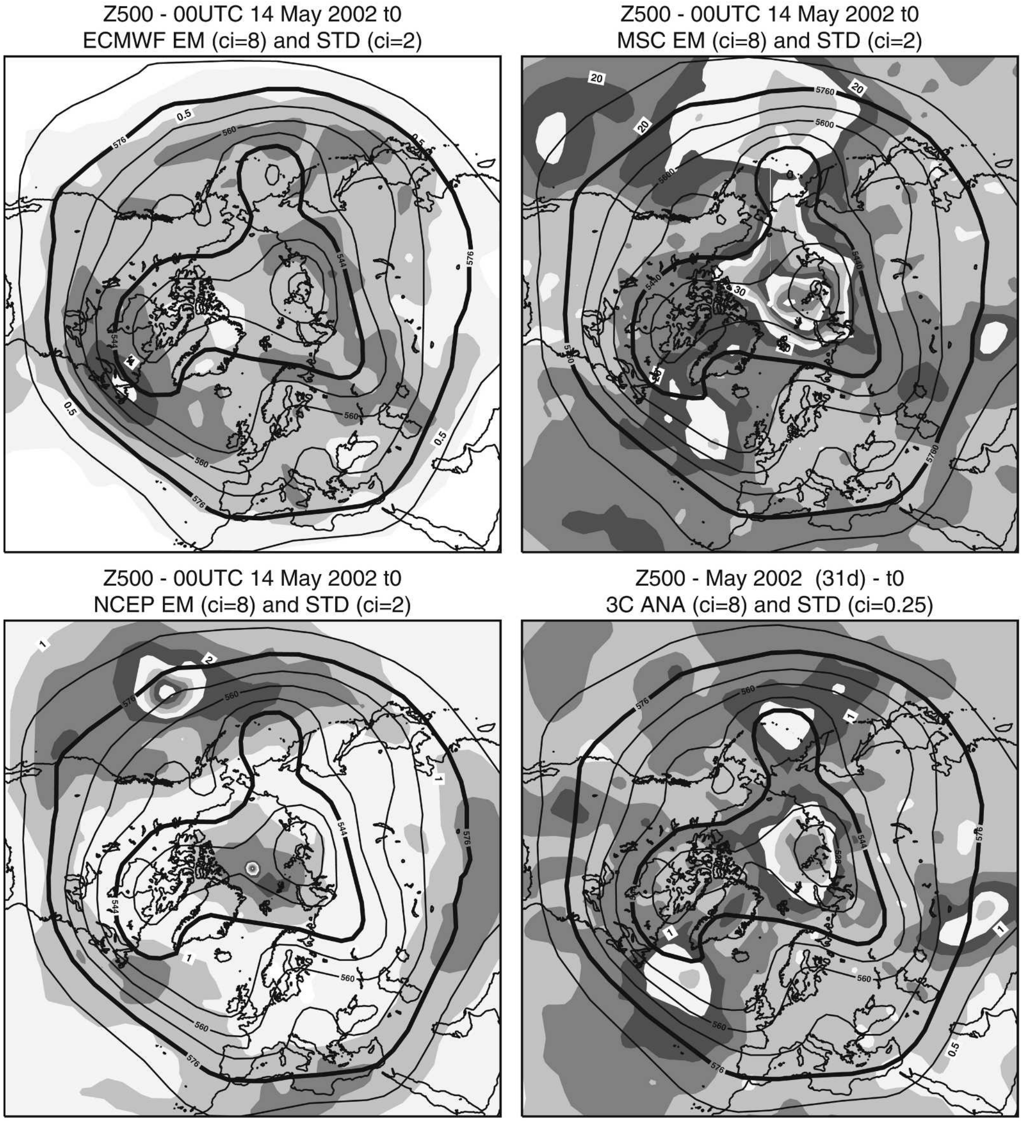
of the SV-based perturbations, and partly to the stochastic simulation of random model errors [Buizza et al. (1999) documented that the introduction of the stochastic simulation of random model errors increased the spread of the old $T_L159L31$ EC-EPS by ~6% at forecast day 7]. This suggests that the introduction of random model perturbations may be at least as effective in increasing ensemble spread as the use of different model versions in the MSC-EPS. This explanation, however, is not definitive, since some models, especially at higher resolution, may be more active than others, contributing to differences in perturbation growth rates.

To provide further insight into the statistical behavior of the forecast systems, the geographical distribution of spread in the three ensembles is contrasted in Figs. 7 and 8 with a crude estimate of uncertainty at initial and 2-day forecast lead times in a manner similar to Figs. 1–3, except averaged for the month of May 2002. As a further reference, a linear measure of atmospheric instability, the Eady index (Hoskins and Valdes 1990), is also shown in Fig. 7:

$$\sigma_E = 0.31\frac{f}{N}\frac{du}{dz}, \tag{6}$$

where N is the static stability and the wind shear is computed using the 300–1000-hPa potential temperature and wind provided by a T63 truncated version of ECMWF analyses; u is the magnitude of the vector wind; and f is the Coriolis parameter.

As already noted in connection with Figs. 1 and 5, the magnitudes of the initial perturbations in the EC-EPS (note use of half-size contour interval) is on average half of that in the other two systems and is comparable to the uncertainty estimate in Fig. 7d. More interesting here are the differences in the geographical distribution of the initial perturbations between the three ensembles: the EC-EPS shows an absolute maximum over the Atlantic, the MSC-EPS over the Arctic, and the NCEP-EPS over the Pacific. The characteristics of the SV-based EC-EPS, and the BV-based NCEP-EPS perturbations are further discussed by Buizza and Palmer (1995) and Toth and Kalnay (1997), respectively. We only note here that the distribution of EC-EPS initial



Z500 - 00UTC 14 May 2002 t0
ECMWF EM (ci=8) and STD (ci=2)

Z500 - 00UTC 14 May 2002 t0
MSC EM (ci=8) and STD (ci=2)

Z500 - 00UTC 14 May 2002 t0
NCEP EM (ci=8) and STD (ci=2)

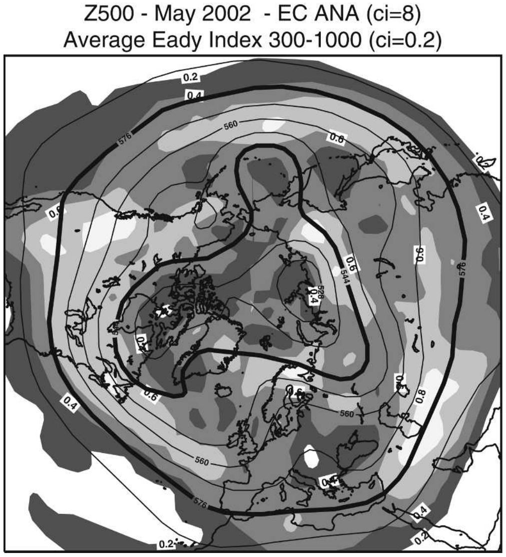
Z500 - May 2002 (31d) - t0
3C ANA (ci=8) and STD (ci=0.25)

FIG. 7. May 2002 initial-time average, at 500-hPa geopotential height. Ensemble mean and standard deviation (shading) of the (a) EC-EPS, (b) MSC-EPS, and (c) NCEP-EPS. (d) Average of the three ensemble means and standard deviation among the three ensemble means (shading), and (e) average of the three ensemble means and Eady index (shading). Contour interval is 8 dam for full field, 0.25 dam for ensemble standard deviations in (a), and 0.5 dam in (b)–(c), 0.25 dam in (d), and 0.2 day^{-1} for Eady index in (e).

spread exhibits some similarity with the Eady index (Fig. 7e), as expected because of the singular vectors' characteristics (Buizza and Palmer 1995). In contrast with the EC-EPS perturbations that are generated by pure linear dynamics, the MSC-EPS perturbations are more representative of the characteristics of the observational network, with more pronounced maxima over the data-sparse regions of the globe. Since it attempts to capture flow-dependent growing analysis errors, it is not surprising that the results from the NCEP-EPS in Fig. 7 appear to fall in between the results generated by pure linear dynamics (EC-EPS) and Monte Carlo analysis error simulation (MSC-EPS).

Interestingly, the geographical distribution of perturbations from the three systems develop considerable similarity even after just 2 days of integrations (Fig. 8). Despite the large discrepancies at initial time, the absolute and relative maxima in the three 2-day forecast ensemble spread charts are reasonably aligned with each other and also with those in the estimated forecast uncertainty (Fig. 8d). Again, this is a reflection of the convergence of initial perturbation and error patterns into a small subspace of perturbations that can grow in a sustainable manner based on the flow-dependent dynamics of the atmosphere.

The second measure of statistical reliability discussed



Z500 - May 2002 - EC ANA (ci=8)
Average Eady Index 300-1000 (ci=0.2)

FIG. 7. (*Continued*)

in this subsection is the percentage of the number of cases when the verifying analysis at any grid point lies outside the cloud of the ensemble in excess to what is expected by chance (Fig. 9). A reliable ensemble will have a score of zero in this measure, whereas larger positive (negative) absolute values indicate more (fewer) outlier verifying analysis cases than expected by chance. Despite an adequate level of spatially averaged spread indicated at day 1 in Fig. 5, the ECMWF ensemble has too many outliers at short lead times. The apparent discrepancy between the results in Figs. 5 and 9 can be reconciled by considering that too small spread in certain areas can be easily compensated by too large spread in other areas when reliability is evaluated using rms standard variation. This is not the case, however, when the outlier statistics are used, since this measure *aggregates* (and not averages) results obtained at different grid points.

In contrast with the EC-EPS results, the MSC-EPS (and to a lesser degree, the NCEP-EPS) rms spread and the outlier statistics results are consistent with each other. This suggests that the initially too large ensemble spread in these two ensembles becomes adequate around days 2–3, before it turns deficient at later lead times. The largest deficiency at later lead times is observed for the NCEP-EPS, probably because of the lack of any model perturbations in that ensemble. Best reliability in terms of outlier statistics is indicated for the MSC-EPS. This is in contrast with the rms spread results (Fig. 5) that suggest the EC-EPS as the most reliable of the three ensembles. The apparent contradiction between the outlier and rms spread results might be explained, on the one hand, by considering that the MSC-EPS outlier results benefit from the use of ensemble members with distinctly different systematic errors due to the use of different models and/or model versions. This MSC sampling of model error is appropriate but insufficient because not all model weaknesses are actually sampled. On the other hand, perturbation growth is known to be influenced by the addition of random noise during model integrations as done in the EC-EPS, where the entire tendency vector obtained from the model physics is affected by the stochastic model-error simulation scheme (Buizza et al 1999).

The third measure of statistical consistency is the reliability component of the BSS (lower panel of Fig. 6). Interestingly, the reliability component of the BSS indicates that the EC-EPS is the least reliable at short lead times and the most reliable at longer lead times. These results are consistent with the outlier statistics at short lead times and the rms spread results at longer lead times.

4) MEASURES OF RESOLUTION AND DISCRIMINATION

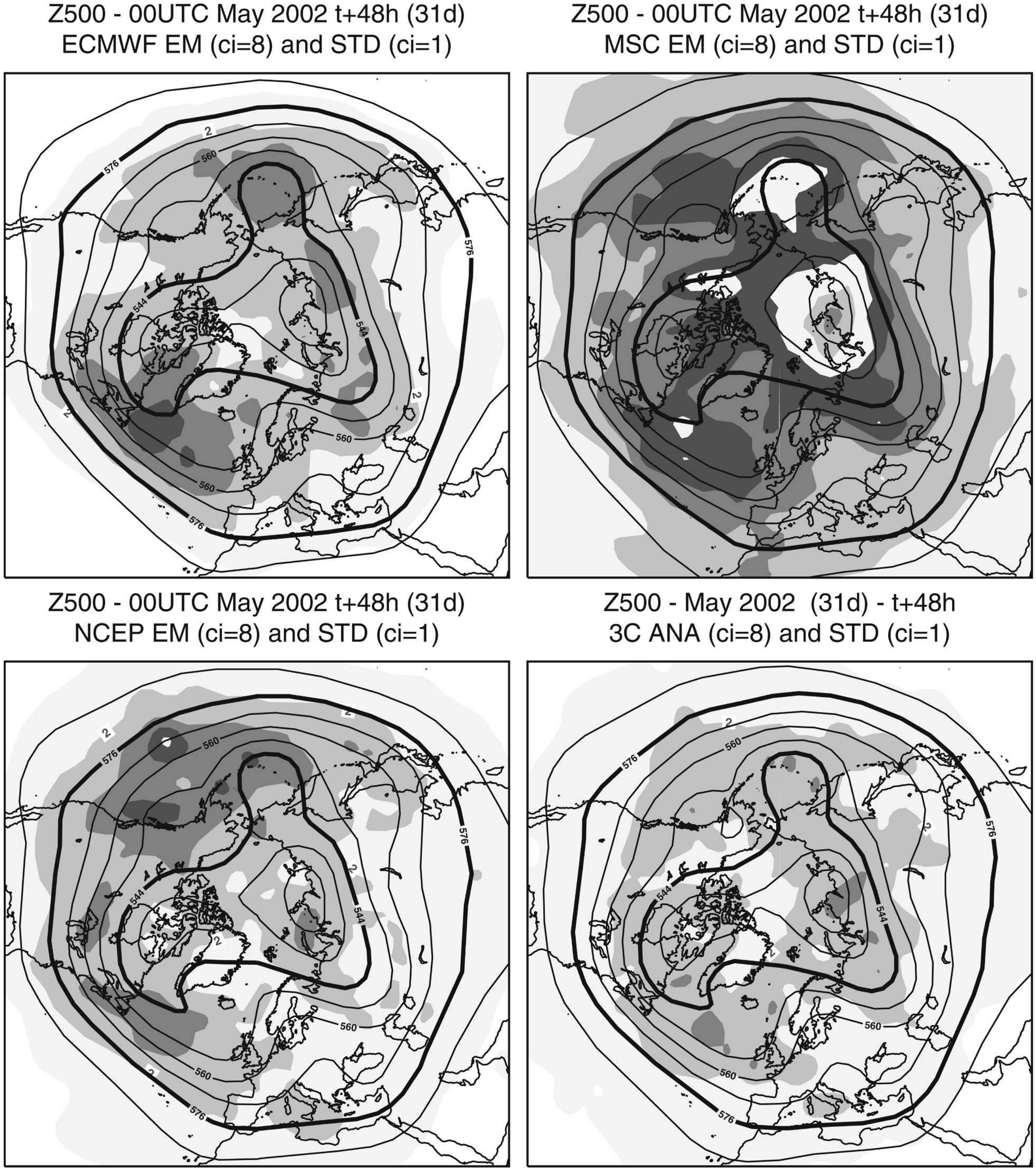
The resolution component of the BSS (lower panel of Fig. 6) provides a quantitative measure of the statistical resolution of the ensemble systems, and the ROC score (Fig. 10) provides a measure of the statistical discrimination capability of the ensemble systems. Note that while statistical postprocessing can enhance reliability, the same does not apply to resolution and discrimination. Therefore, measures of these two characteristics assess more directly the inherent value of a forecasting system.

As was the case for the BSS (Fig. 6), the ROC area shown in Fig. 10 has been computed by averaging the ROC area for 10 climatologically equally likely events. Both measures indicate that the best resolution and discrimination are obtained by the EC-EPS. At short—up to 2 days—lead time the NCEP-EPS is competitive, probably because of the beneficial effects of initial perturbations (bred vectors). With increasing lead time, the resolution and discrimination of the NCEP-EPS, just as its reliability, suffer from the lack of model perturbations. On the other hand, the MSC-EPS becomes competitive even with the EC-EPS near the end of the 10-day forecast period, probably because of the use of multiple model versions.

5) PERTURBATION PATTERN ANALYSIS

The PECA (Wei and Toth 2003), a score designed to be insensitive to the quality of the deterministic prediction system, is used to evaluate directly the quality of ensemble perturbation patterns. The higher the correlation of individual (or of optimally combined) ensemble perturbations with the error in the control forecast, the more successful the ensemble is in encompassing the verifying analysis.

The most visible feature in the results presented in Fig. 11 is that for all three ensemble systems the PECA values increase with increasing lead time. This is related to the convergence of both the perturbation and the error



Z500 - 00UTC May 2002 t+48h (31d)
ECMWF EM (ci=8) and STD (ci=1)

Z500 - 00UTC May 2002 t+48h (31d)
MSC EM (ci=8) and STD (ci=1)

Z500 - 00UTC May 2002 t+48h (31d)
NCEP EM (ci=8) and STD (ci=1)

Z500 - May 2002 (31d) - t+48h
3C ANA (ci=8) and STD (ci=1)

FIG. 8. May 2002 + 48-h average, at 500-hPa geopotential height. Ensemble mean and standard deviation (shading) of the (a) EC-EPS, (b) MSC-EPS, and (c) NCEP-EPS. (d) Average of the three ensemble means and standard deviation among the three ensemble means (shading). Contour interval is 8 dam for full field, and 1 dam for ensemble standard deviations.

patterns to a small subspace of growing patterns, characterized by the leading Lyapunov vectors in a linear setting or by the fastest-growing nonlinear perturbations [Toth and Kalnay (1997) and Boffetta et al. (1998) showed that these coincide with bred vectors or nonlinear Lyapunov vectors]. Note also that the PECA values also increase when the number of degrees of freedom is reduced because of the use of smaller domain size.

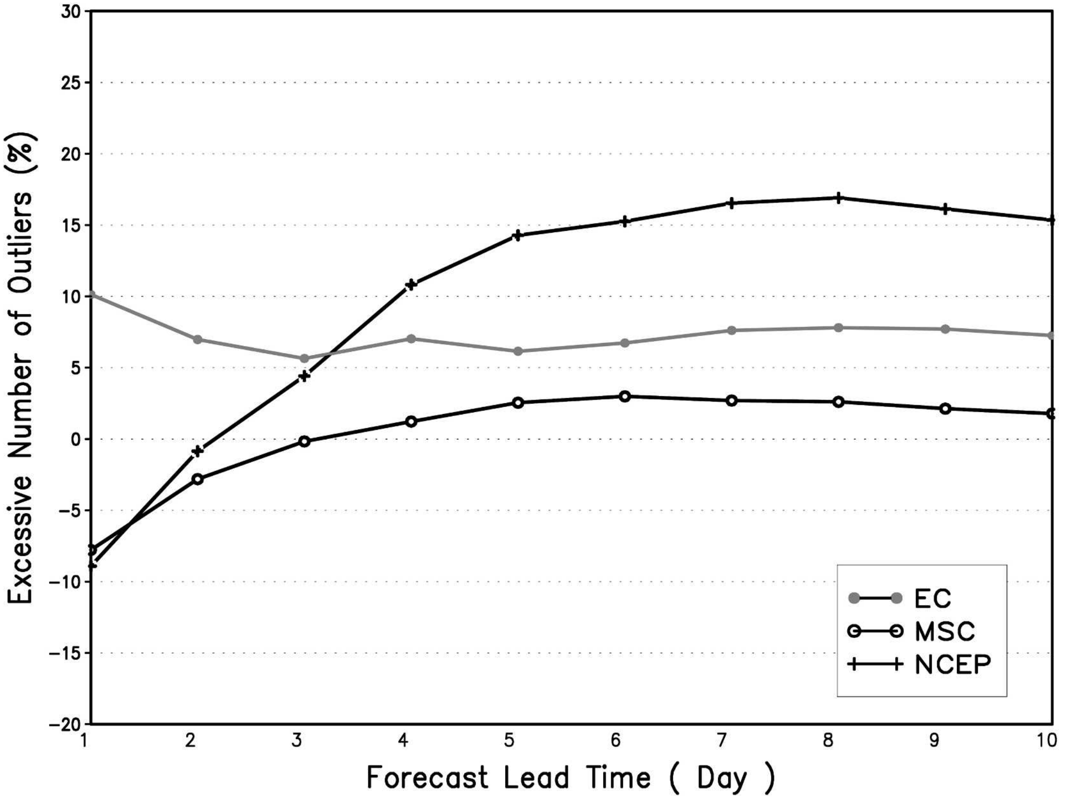


FIG. 9. May–Jun–Jul 2002 average percentage of excessive outliers for the EC-EPS (gray lines with full circles), the MSC-EPS (black lines with open circles), and the NCEP-EPS (black lines with crosses). Values refer to the 500-hPa geopotential height over the Northern Hemisphere latitudinal band 20°–80°N.

When comparing the PECA values for the three ensemble systems, we note first that the EC-EPS scores are not above those from the MSC-EPS or NCEP-EPS. Since PECA is insensitive to the quality of the initial analysis, this result suggests that the main reason for the better performance of the EC-EPS in terms of the rms, PAC, ROC, and Brier skill score measures is the superiority of the ECMWF data assimilation (and perhaps numerical forecast modeling) system, and not necessarily the strategy used to simulate initial value and model-related uncertainties in the EC-EPS.

Since the PECA results reflect more directly the performance of the three ensemble generation schemes, the advantages of the different ensemble systems can be more easily detected. When the PECA analysis is restricted to the hemispheric and smaller scales (Figs. 11b–d), the NCEP-EPS has a clear advantage for the short forecast ranges. Over the North American/European region, the optimally combined NCEP-EPS perturbations, for example, can explain around 38%–53% of the 12–24-h forecast-error variance (with PECA values around 0.62%–0.72%), compared with around 25%–40% explained error variance (associated with 0.5%–0.63% PECA values) by the other two EPS systems. Assuming that PECA values are independent only every fifth day, the differences among the NCEP-EPS, the ECMWF-EPS, and the MSC-EPS are statistically significant at the 0.1%–0.5% level. Statistically significant results are also found for day-2 (36–48 h) lead times. The relatively good performance of the NCEP-EPS at short lead times may be due to the ability of the breeding method to efficiently sample analysis errors on the synoptic and smaller scales. When larger, global scales are also included in the analysis (Fig. 11a), the MSC ensemble becomes superior, especially at longer lead times. This may be due to the value of model diversity in capturing forecast-error patterns that are potentially affected by large-scale model systematic errors, especially at longer lead times.

During the first 1–2 days, PECA values for the EC-EPS tend to be lower when compared to the other ensembles. This may be due to the use of a norm (total energy) in the computation of the singular-vector perturbations that is not directly connected with analysis uncertainty. It is also interesting to note that on the hemispheric and global domains, it is the ECMWF ensemble that shows the largest gain when individual ensemble perturbations are optimally combined to maximize the explained forecast-error variance. Likely this is an advantage related to the orthogonalization inherent in the calculation of the singular-vector perturbations.

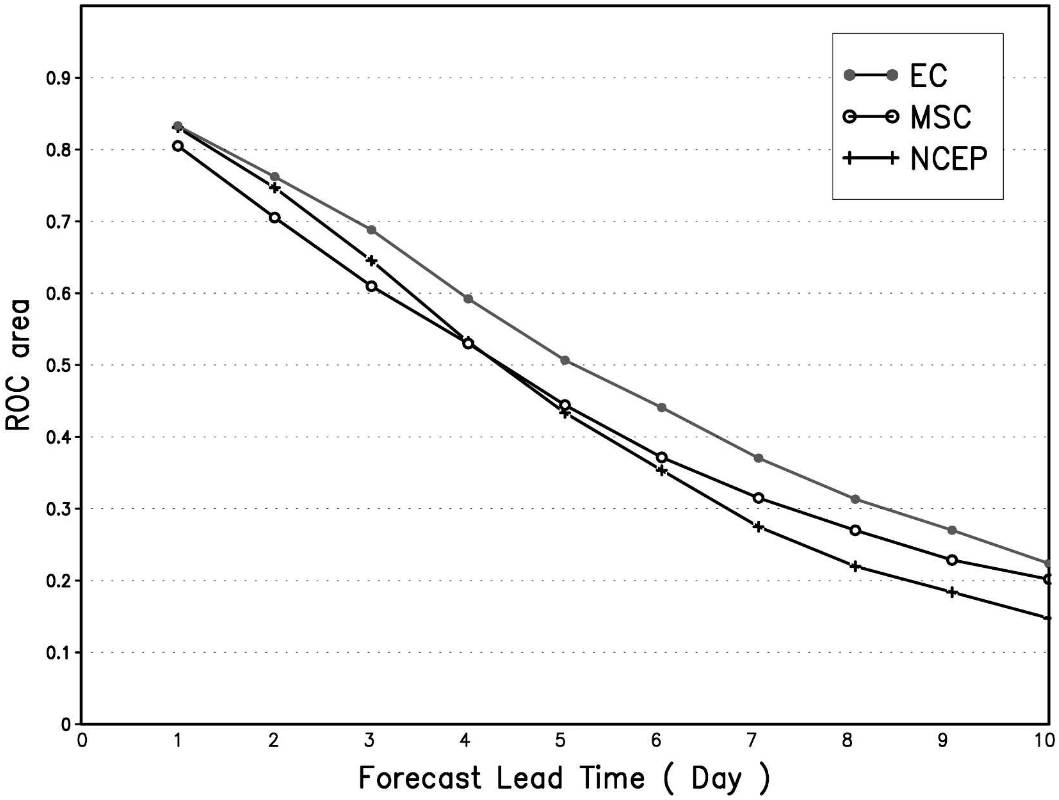


FIG. 10. May–Jun–Jul 2002 area under the relative operating characteristics for the EC-EPS (gray lines with full circles), the MSC-EPS (black lines with open circles), and the NCEP-EPS (black lines with crosses). Values refer to the 500-hPa geopotential height over the Northern Hemisphere latitudinal band 20°–80°N and have been computed considering 10 equally cliamatologically likely intervals.

4. Future directions

After more than a decade of intense research, a number of open science questions related to ensemble generation methods still remain.

a. Random versus selective sampling

Ensemble forecasting involves Monte Carlo sampling. However, there is a disagreement on whether random sampling should occur in the full space of analysis errors, including nongrowing directions, or only in the fast-growing subspace of errors. It is likely that the next several years will see more research in this area of ensemble forecasting, yielding quantitative results that will allow improvements in operational procedures.

b. Significance of transient errors

Another open question is related to the role of transient behavior in the evolution of forecast errors. If one chooses to explore only the fast-growing subspace of possible analysis errors for the generation of ensemble perturbations, should one use the leading Lyapunov or bred vectors, or alternatively should one use singular vectors that can produce super-Lyapunov error growth?

c. Exploring the links between data assimilation and ensemble forecasting

Ensemble forecasting and data assimilation efforts can mutually benefit from each other and the two systems can be jointly designed (Houtekamer et al. 1996a). Such an approach is pursued at MSC and is considered at several other centers. In these efforts an appropriate sampling of model error (Dee 1995) and an optimal use of a limited number of ensemble members are of critical importance. This is a very complex, yet potentially promising, area of research that many in the field view with great expectations not only for global but also for limited area modeling applications (Toth 2003).

d. Representation of model uncertainties

The representation of forecast uncertainty related to the use of imperfect models will be another area of intense research. Do the currently used techniques capture flow-dependent variations in skill linked with model-related errors, or only improve statistical reliability of the forecasts that can potentially be achieved

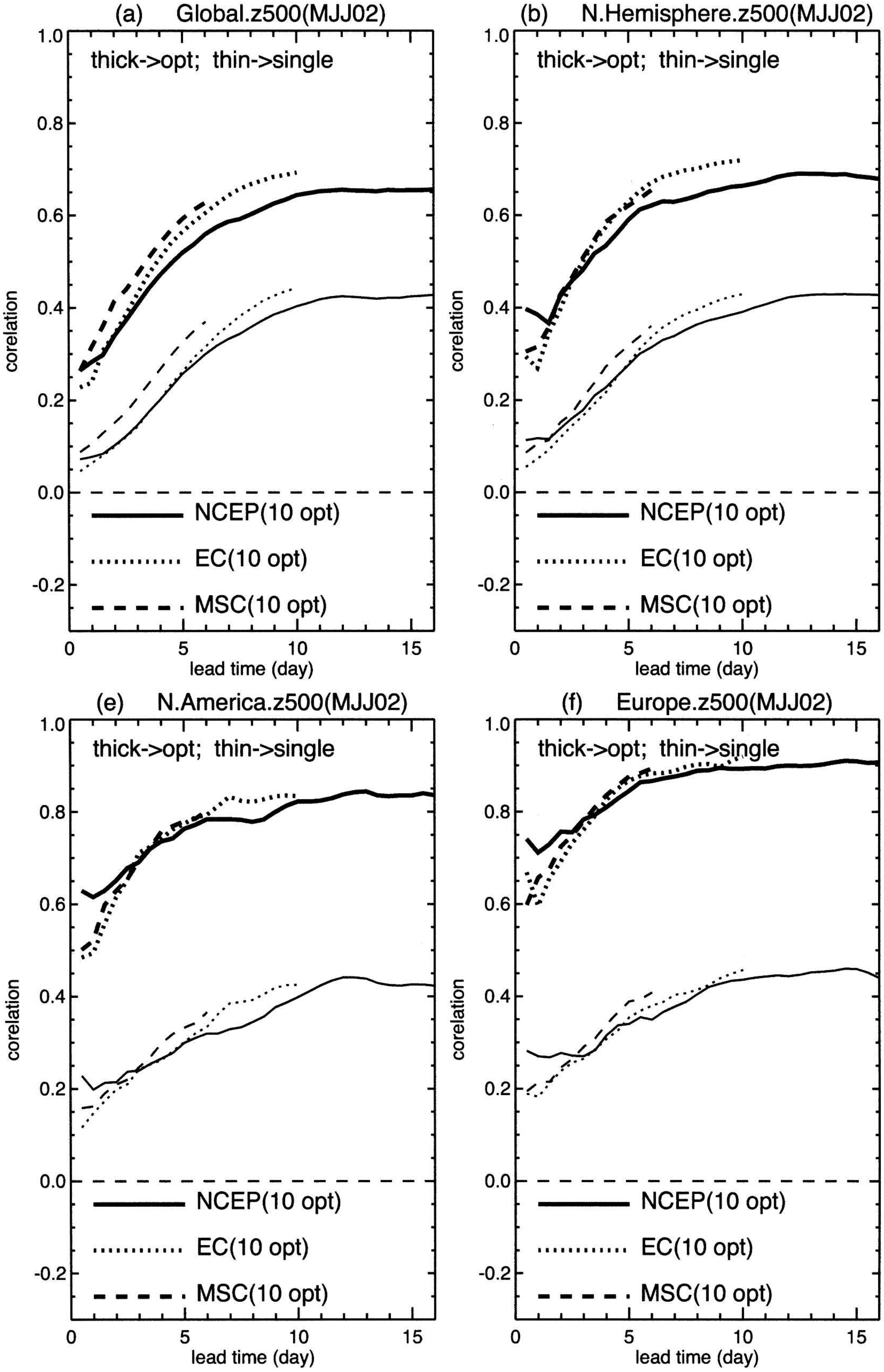


FIG. 11. PECA for the (a) global, (b) NH, (c) North American, and (d) European regions for individual (thin) and optimally combined (heavy) ensemble perturbations. See text for further details on the definition of PECA.

equally well through statistical postprocessing? Can a new generation of NWP models be developed that offer a comprehensive approach to capturing both random and systematic model errors (Toth and Vannitsem 2002)? Research on these issues will be pursued at ECMWF, NCEP, and MSC in the coming years, with the hope that one day case-dependent model-related errors can be better captured by the ensemble systems.

5. Conclusions

In a chaotic system like the atmosphere, probabilistic information is recognized as the optimum format for weather forecasts both from a scientific and a user perspective. Ensemble forecasts are well suited to support the provision of such probabilistic information. In fact, ensembles not only improve forecast accuracy in a traditional sense (by reducing errors in the estimate of the first moment of the forecast probability distribution), but also offer a practical way of measuring case-dependent variations in forecast uncertainty (by providing an estimate of the higher moments of the forecast probability density function).

Ensemble forecasting has gained substantial ground in numerical weather prediction in the past decade. Today, many numerical weather prediction centers use ensemble methods in their modeling suite (WMO 2003).

In this paper ensemble techniques such as the singular vector, multiple analysis cycle, and breeding methods for the generation of initial perturbations and the stochastic perturbation and multiple model version techniques for representing model-related uncertainty were reviewed and compared. To assess the merit of the different existing approaches, operational ensemble forecasts generated at three operational numerical weather prediction centers were comparatively verified over a 3-month period, May–June–July 2002. Since NCEP generates only 10 perturbed forecasts from each initial time, for ease of comparison and interpretation the quantitative analysis has been limited to 10-member ensembles (the reader should be aware that this induces an underestimation of the actual skill of the ensemble systems, especially for systems with a large membership; Buizza and Palmer 1998).

Most verification measures indicate that the ECMWF ensemble forecast system has the best overall performance, with the NCEP system being competitive during the first, and the MSC system during the last, few days of the 10-day forecast period. These verification methods, however, measure the overall accuracy of ensemble forecasts influenced by the quality of the data assimilation, numerical weather prediction modeling, and ensemble generation schemes. The results therefore are not directly indicative of the strengths/weaknesses of the different ensemble generation schemes. When the forecasts are evaluated using a new technique (PECA) that measures the correlation between ensemble perturbations (instead of the full forecasts, thus eliminating the effect of the quality of the analysis on the scores) and forecast-error patterns, the three ensemble systems are found to perform rather similarly.

From a careful analysis of the results based on small-size (10 perturbed members only) ensemble systems for May–June–July 2002, consensus emerges on the following aspects of the systems:

- Overall, the EC-EPS exhibits the most skillful performance when measured using rms, PAC, BSS, and ROC-area measures.
- When PECA is used to measure the correlation between perturbation and forecast-error patterns, the EC-EPS does not show any superior performance. At short lead times, the error patterns are best described by the NCEP-EPS if one considers the small scales and by the MSC-EPS if one considers the large scales.
- Results suggest that the superior skill of the EC-EPS may be mostly due to its superior model and data assimilation systems and should not be considered as a proof of a superior performance of SV-based initial perturbations. In other words, at MSC and NCEP ensemble performance is negatively affected in the short range by the relatively low quality of the ensemble of data assimilation systems, and in the long range by the relatively low model resolution.
- As for statistical reliability, the superior outlier statistics of the MSC-EPS may be due to the use of multiple model versions. This technique may capture large-scale model-related errors in longer lead times.
- The spread in the (single model) EC-EPS grows faster than that in the other two systems because of a combined effect of sustained SV-based perturbations' growth and the stochastic simulation of random model errors. This is due to the combined effect of sustained SV-based perturbations' growth and the stochastic simulation of random model errors.
- There are indications that the stochastic simulation of the random model-error scheme implemented in the ECMWF-EPS improves the forecast statistical reliability.

During the past decade different ensemble generation techniques received significant attention and underwent substantial refinements. Yet a number of open questions still remain. Ongoing ensemble-related research in the coming years is expected to provide a better understanding of the still remaining scientific issues. The intercomparison of the performance of the ECMWF, MSC, and NCEP ensemble forecast systems reported in this paper can be considered as a first necessary step toward answering some of the open questions. Continued future collaboration, where in a controlled experiment initial ensemble perturbations from the three different systems are introduced in the analysis/forecast system of a selected center, could potentially provide additional useful information, contributing to improved forecast operations.

Acknowledgments. The emergence of ensemble forecasting, as with any major scientific and technological development, is the result of the dedicated, high quality work of an entire community. The development, operational implementation, and continuous improvement of ensemble predictions would not have been possible without the contribution of many staff members and consultants; their work, as well as the enthusiastic support of the management at ECMWF, MSC, and NCEP is gratefully acknowledged here. Drs. Laurie Wilson, Dingchen Hou, and Henry Juang are thanked for providing very useful comments that helped improve the quality of this manuscript. Three anonymous referees and the journal's editor are also thanked for their constructive criticism and comments to the first submitted version of this manuscript.

APPENDIX

Brier Score Decomposition

The Brier score can be decomposed into its reliability, resolution, and uncertainty components:

$$\begin{aligned}
\mathrm{BS} &= \mathrm{BS}_{\mathrm{rel}} - \mathrm{BS}_{\mathrm{resol}} + \mathrm{BS}_{\mathrm{unc}},\\
\mathrm{BS}_{\mathrm{rel}} &= \frac{1}{n}\sum_{i=1}^{I} N_i(y_i - \overline{o}_i)^2,\\
\mathrm{BS}_{\mathrm{resol}} &= \frac{1}{n}\sum_{i=1}^{I} N_i(\overline{o}_i - \overline{o})^2,\\
\mathrm{BS}_{\mathrm{unc}} &= \overline{o}(1 - \overline{o}),
\end{aligned} \qquad \text{(A1)}$$

where y_i is the forecast probability, o_i is the observed probability, N_i is the relative frequency of the forecast event in each subsample i, and

$$\begin{aligned}
n &= \sum_{i=1}^{I} N_i,\\
\overline{o}_i &= \frac{1}{N_i}\sum_{k\in N_i} o_k,\\
\overline{o} &= \frac{1}{n}\sum_{k=1}^{n} o_k.
\end{aligned} \qquad \text{(A2)}$$

The reliability term summarizes the calibration, or conditional bias, of the forecast. It consists of a weighted average of squared differences between the forecast probabilities and relative frequencies of the forecast event in each subsample. The resolution term summarizes the ability of the forecast to discern subsample forecast periods with different relative frequencies of the event. The forecast probabilities do not appear explicitly in this term, yet it still depends on the forecasts through the sorting of the events making up the subsample relative frequencies. The uncertainty term depends only on the sample climatological relative frequency and is unaffected by forecasts.

The Brier skill score is defined as

$$\mathrm{BSS} = \frac{\mathrm{BS} - \mathrm{BS}_{\mathrm{ref}}}{\mathrm{BS}_{\mathrm{perf}} - \mathrm{BS}_{\mathrm{ref}}} = 1 - \frac{\mathrm{BS}}{\mathrm{BS}_{\mathrm{ref}}}. \qquad \text{(A3)}$$

If one considers $\mathrm{BS}_{\mathrm{ref}} = \mathrm{BS}_{\mathrm{unc}}$, then

$$\mathrm{BSS} = \frac{\mathrm{BS}_{\mathrm{resol}} - \mathrm{BS}_{\mathrm{rel}}}{\mathrm{BS}_{\mathrm{unc}}}. \qquad \text{(A4)}$$

REFERENCES

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

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

——, 2001: Verification of intense precipitation forecasts from single models and ensemble prediction systems. *Nonlinear Processes Geophys.,* **8,** 401–417.

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.

Boffetta, G., P. Guliani, G. Paladin, and A. Vulpiani, 1998: An extension of the Lyapunov analysis for the predictability problem. *J. Atmos. Sci.,* **55,** 3409–3416.

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

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

——, and ——, 1998: Impact of ensemble size on ensemble prediction. *Mon. Wea. Rev.,* **126,** 2503–2518.

——, T. Petroliagis, T. N. Palmer, J. Barkmeijer, M. Hamrud, A. Hollingsworth, A. Simmons, and N. Wedi, 1998: Impact of model resolution and ensemble size on the performance of an ensemble prediction system. *Quart. J. Roy. Meteor. Soc.,* **124,** 1935–1960.

——, M. Miller, and T. N. Palmer, 1999: Stochastic representation of model uncertainties in the ECMWF ensemble prediction system. *Quart. J. Roy. Meteor. Soc.,* **125,** 2887–2908.

——, D. S. Richardson, and T. N. Palmer, 2003: Benefits of increased resolution in the ECMWF ensemble system and comparison with poor-man's ensembles. *Quart. J. Roy. Meteor. Soc.,* **129,** 1269–1288.

Caplan, P., J. Derber, W. Gemmill, S. Hong, H. Pan, and D. Parrish, 1997: Changes to the 1995 NCEP Operational Medium-Range Forecast Model Analysis–Forecast System. *Wea. Forecasting,* **12,** 581–594.

Côté, J., S. Gravel, A. Méthot, A. Patoine, M. Roch, and A. Staniforth, 1998: The operational CMC/MRB Global Environmental Multiscale (GEM) model. Part I: Design considerations and formulation. *Mon. Wea. Rev.,* **126,** 1373–1395.

Coutinho, M. M., B. J. Hoskins, and R. Buizza, 2004: The influence of physical processes on extratropical singular vectors. *J. Atmos. Sci.,* **61,** 195–209.

Dee, D. P., 1995: On-line estimation of error covariance parameters for atmospheric data assimilation. *Mon. Wea. Rev.,* **123,** 1128–1145.

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

Evensen, G., 1994: Sequential data assimilation with a nonlinear quasi-geostrophic model using Monte Carlo methods to forecast error statistics. *J. Geophys. Res.,* **99** (C5), 10 143–10 162.

Gauthier, P., C. Charette, L. Fillion, P. Koclas, and S. Laroche, 1999: Implementation of a 3d variational data assimilation system at the Canadian Meteorological Centre. Part 1: The global analysis. *Atmos.–Ocean,* **37,** 103–156.

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

Hoskins, B. J., and P. J. Valdes, 1990: On the existence of storm tracks. *J. Atmos. Sci.,* **47,** 1854–1864.

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

——, and L. Lefaivre, 1997: Using ensemble forecasts for model validation. *Mon. Wea. Rev.,* **125,** 2416–2426.

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

——, ——, and ——, 1996b: The RPN ensemble prediction system. *Proc. ECMWF Seminar on Predictability,* Vol. II, Reading, United Kingdom, ECMWF, 121–146.

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

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

Marshall, J., and F. Molteni, 1993: Toward a dynamical understanding of planetary-scale flow regimes. *J. Atmos. Sci.,* **50,** 1792–1818.

Mason, I., 1982: A model for assessment of weather forecasts. *Aust. Meteor. Mag.,* **30,** 291–303.

Mitchell, H. L., C. Chouinard, C. Charette, R. Hogue, and S. J. Lambert, 1996: Impact of a revised analysis algorithm on an operational data assimilation system. *Mon. Wea. Rev.,* **124,** 1243–1255.

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

Mullen, S. L., and R. Buizza, 2001: Quantitative precipitation forecasts over the United States by the ECMWF Ensemble Prediction System. *Mon. Wea. Rev.,* **129,** 638–663.

Murphy, A. H., 1973: A new vector partition of the probability score. *J. Appl. Meteor.,* **12,** 595–600.

Palmer, T. N., F. Molteni, R. Mureau, and R. Buizza, 1993: Ensemble prediction. *Proc. ECMWF Seminar on Validation of Models over Europe,* Vol. I, Reading, United Kingdom, ECMWF, 21–66.

Pellerin, G., L. Lefaivre, P. Houtekamer, and C. Girard, 2003: Increasing the horizontal resolution of ensemble forecasts at CMC. *Nonlinear Processes Geophys.,* **10,** 463–488.

Pires, C., R. Vautard, and O. Talagrand, 1996: On extending the limits of variational assimilation in nonlinear chaotic systems. *Tellus,* **48A,** 96–121.

Ritchie, H., and C. Beaudoin, 1994: Approximations and sensitivity experiments with a baroclinic semi-Lagrangian spectral model. *Mon. Wea. Rev.,* **122,** 2391–2399.

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 Tech. Doc. 358, World Meteorological Organization, 114 pp.

Talagrand, O., R. Vautard, and B. Strauss, 1997: Evaluation of probabilistic prediction systems. *Proc. ECMWF Workshop on Predictability*, Reading, United Kingdom, ECMWF, 1–26.

Toth, Z., cited 2003: Report on the 2nd Ensemble Based Data Assimilation Workshop. NCEP, Camp Springs, MD. [Available online at http://wwwt.emc.ncep.noaa.gov/gmb/ens/enswksh.html.]

——, 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 Workshop on Meteorological Operational Systems,* Reading, United Kingdom, ECMWF, 146–154.

——, I. Szunyogh, E. Kalnay, and G. Iyengar, 1999: Comments on: "Notes on the appropriateness of 'bred modes' for generating initial perturbations." *Tellus,* **51A,** 442–449.

——, Y. Zhu, I. Szunyogh, M. Iredell, and R. Wobus, 2002: Does increased model resolution enhance predictability? Preprints, *Symp. on Observations, Data Assimilation, and Probabilistic Prediction,* Orlando, FL, Amer. Meteor. Soc., J18–J23.

——, O. Talagrand, G. Candille, and Y. Zhu, 2003: Probability and ensemble forecasts. *Forecast Verification: A Practitioner's Guide in Atmospheric Science,* I. T. Jolliffe and D. B. Stephenson, Eds., Wiley, 137–163.

Wei, M., and Z. Toth, 2003: A new measure of ensemble performance: Perturbation versus error correlation analysis (PECA). *Mon. Wea. Rev.,* **131**, 1549–1565.

Wilks, D. S., 1995: *Statistical Methods in the Atmospheric Sciences*. Academic Press, 467 pp.

WMO, 2003: 21st status report on the implementation of the World Weather Watch—2003. WMO 957, 61 pp.

Zhu, Y., G. Iyengar, Z. Toth, M. S. Tracton, and T. Marchok, 1996: Objective evaluation of the NCEP global ensemble forecasting system. Preprints, *15th Conf. on Weather Analysis and Forecasting,* Norfolk, VA, Amer. Meteor. Soc., J79–J82.