Assessment of the status of global ensemble prediction

Roberto Buizza[1], Peter L. Houtekamer[2], Zoltan Toth[3], Gerald Pellerin[2], Mozheng Wei[4], Yuejian Zhu[3]

[1] *European Centre for Medium-Range Weather Forecasts, Reading UK (www.ecmwf.int)*
[2] *Meteorological Service of Canada, Dorval, Quebec, Canada (www.msc-smc.ec.gc.ca)*
[3] *National Centers for Environmental Prediction, Washington D. C., US (www.emc.ncep.noaa.gov)*
[4] *UCAR Visiting Scientist, NCEP/EMC, Washington, US*

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 that:

- 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 are 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

Traditionally, the main objective of numerical weather prediction has been the generation of a single forecast that provides the best estimate of the future state of the atmosphere. Though the chaotic nature of the atmosphere that limits predictability and skill for all forecast applications was recognized from early on (*Lorenz* 1963, 1982), it was not until the 1990s that the uncertainty inherent in numerical forecasts became the subject of intense research, eventually leading to changes in operational weather forecasting.

In an environment where forecast skill is limited, a single value (which could be either the expected or the most likely value) of the future state of a system variable contains only limited information. Ideally, the state of the system would be described through a probability distribution function (pdf) specifying the probability that the future state variable will fall in any arbitrary interval. Such a forecast format can also satisfy advanced users, whose applications crucially depend on probability information (*Richardson* 2000, *Smith et al.* 2001, *Zhu et al.* 2002, *Taylor & Buizza* 2003).

Probabilistic forecasts can also be based on statistical post-processing of a single deterministic forecast. Due to the complexity of the atmosphere, and the very limited sample of available pairs of corresponding forecasts and observations (or analyzed fields), such statistical post-processing methods can describe the loss of forecast skill only in an average sense, and are typically unable to resolve case dependent deviations from average forecast uncertainty. Recent results have confirmed earlier indications that such variations are significant (*Toth et al.* 2002), and that probabilistic forecast methods based on statistical port-processing of single forecasts provide inferior forecast performance (*Buizza et al.* 2003).

One solution to the problem of forecasting in the presence of uncertainty is through the use of the Liouville equations (*Ehrendorfer* 1994a, b). This method is based on continuity equations in the probability space, recognizing that the integral of the initial probability density function is preserved throughout the forecast process. The resulting very high computational demand renders the Liouville equations impractical for probabilistic weather forecasting applications. Attempts at finding a satisfying closure scheme with the full probability density functions truncated to their first few moments also turned out to be unsuccessful (*Fleming* 1971a, b, *Pitcher* 1977).

Attention then turned toward practically feasible Monte Carlo methods. These methods are based on a statistical sampling approach (*Leith* 1974) where the forecast probability density function is approximated using a finite sample of forecast scenarios (ensemble of forecasts). These forecasts are started from a sample of states drawn from a probability density function of the initial state (that 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. Note that the concept of ensemble Kalman filter (*Evensen*, 1994) offers a way of unifying ensemble forecasting and data assimilation. Knowledge on model error statistics can potentially also be used to improve the forecast model (*Houtekamer & Lefaivre*, 1997). Methodologies for ensemble-based data assimilation and model improvement, however, have not yet been applied operationally at weather prediction centers. Therefore 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 originating 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.

At the U.S. National Centers for Environmental Prediction (NCEP, formerly NMC) *Toth & 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 Center 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 & 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 by *Houtekamer et al.* (1996b) 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 will be 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 only in simplified environments only. *Houtekamer & Derome* (1995) compared the different strategies of ensemble prediction in a simplified environment consisting of a simulated observational network and the three-level quasi-geostrophic T21 model of *Marshall & 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 due to not only initial errors, but also due to 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 will be presented in section 2. The performance of the three ensemble systems will then be 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 sub-set of the ensemble systems that includes only 10 perturbed and one unperturbed member starting at 00 UTC. Future directions and ongoing research issues are discussed in sections 4 and 5, while section 6 offers some preliminary conclusions.

2. Ensemble prediction at ECMWF, MSC and NCEP

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

- 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);
- 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 parameterised 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:

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

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

In contrast, ECMWF and NCEP initial ensemble states are created by adding either bred or singular vectors to the best estimate of the atmosphere at initial time:

$$e_j(0) = e_0(0) + de_j(0) \qquad (3a)$$

which itself is produced by a high-resolution 3- or 4-dimensional data assimilation procedure:

$$e_0(0) = \Xi[e_0(\tau_1), o_0(\tau_1 \div \tau_2), P_0, A_0] \tag{3b}$$

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

	MSC	ECMWF	NCEP
Pj (model uncert.)	2 mod + Diff. Ph. Par.	Pj=P0 (single model)	Pj=P0 (single model)
dPj (random model error)		dPj=rj*Pj (stoch phys)	dPj=0
Aj	2 models	Aj=A0 (single model)	Aj=A0 (single model)
oj (observ. error)	Random perturbations	-	-
ej (initial uncert.)	ej from Anal. Cycles	ej=e0+dej(SV)	ej=e0+dej(BV)
horiz. resolution HRES forecast	100 km	-	T170(d0-7.5)>T126(d7.5-16)
horiz. resolution control forecast	TL149, not available for this study	TL255 (d0-10)	T126(d0-3.5)>T62(d3.5-16)
horiz. resolution perturbed members	TL149	TL255 (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	10hPa	10hPa	3hPa
number of perturbed members	16	50	10
forecast length	10 days	10 days	16 days
daily frequency	00 UTC	12 UTC (00 UTC exp)	00 and 12 UTC
operational implementation	February 1998	December 1992	December 1992

Table 1. Summary of ensemble characteristics as of July 2002.

2.1. The singular-vector approach at ECMWF

The ECMWF singular-vector (SV) approach (*Buizza & 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 up to the optimization time, and the metric corresponding to the energy-norm. 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 & 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 hours 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 hours *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 parameterised 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 mean of the ensemble interpolated from the 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.

At the time of writing this report (Summer 2003), 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), scaled to have an amplitude consistent with analysis error estimates. Since April 2003 the EPS has been running twice a day, with 00 and 12 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 grid-point longitude and latitude, $\langle \ldots \rangle_{10,6}$ indicates that the same random number r_j is used inside a 10-degree box and a 6-hour time window (see *Buizza et al.* 1999 for more details). The initial perturbations $de_j(0)$ are defined as

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

where for each geographical region (northern and southern hemisphere extra-tropics, NH and SH, 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 4D-Var data-assimilation scheme (see *Molteni et al.* 1996 for more details).

2.2. The MSC perturbed-observation approach

The MSC perturbed-observation approach attempts to obtain a representative ensemble of perturbations by comprehensively simulating the behaviour 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 non-selective, purely random ensemble perturbations are presented in *Houtekamer et al.* (1996a) and by *Anderson* (1997).

In the first version of the MSC-EPS, implemented operationally in February 1998, all 8 members used the Spectral Finite Element model at resolution T_L95 (*Ritchie & 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.

The perturbation of 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 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 parametrizations were of similar quality (*Houtekamer & Lefaivre,* 1997). The selected physical parametrizations 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 8 additional members have been generated using the newly developed Global Environmental Multi-scale model (*Côté et al.* 1998). Furthermore, updated versions of physical parametrizations have been used for the 8 members that use the Global Environmental Multi-scale 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. The spectral truncation of the 8 members that use the Spectral Finite Element model was increased from T_L95 to T_L149, and the resolution of the 8 members that use the Global Environmental Multi-scale model increased from a 2 degree to a 1.2 degree uniform grid. This was possible due to 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 8 additional initial conditions for the medium-range forecasts is obtained by means of a correction (*Houtekamer & Lefaivre*, 1997) towards 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 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 re-evaluation.

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

2.3. 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 & Kalnay* 1993, 1997). For example, based on 4D-VAR data assimilation 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" non-linear perturbation model. When its original form is used with a single global rescaling factor, the bred vectors (BV) represent a non-linear 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 techniques is that while the breeding techniques 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 & 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 spread in the NCEP ensemble slightly hinders performance in the short lead-time ranges, it improves performance in the medium- and extended ranges (*Toth & Kalnay* 1997).

NCEP produces 10 perturbed ensemble members both at 00 and 12 UTC every day out to 16 days lead-time. For both cycles, the generation of the initial perturbations is done in 5 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 hours, 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 hours, then truncated to T62L28) starting from the high-resolution operational analysis. At 00 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

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

The coefficients $\underline{\underline{RR}}$ of the linear combination matrices in Eq. (5) are defined by the regional re-scaling algorithm (*Toth & Kalnay* 1997).

2.4. An example: the forecast case of 14 May 2003

To illustrate the effect of the use of different configurations on the characteristics of the three ensembles, a forecast example is presented in Figs. 1, 2 and 3. These figures 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 May 2002), along with the t+48h and the t+120 hour forecasts. The comparison is based on 11-member ensembles started at 00 UTC, and each ensemble is verified against its corresponding analysis.

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 reference 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 30ºN degree is almost zero, since SV perturbations in the tropical region are generated only in the vicinity of tropical cyclones (see section 2.1).

In 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 (Fig. 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+48h (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+120h (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. Note that for this particular forecast case there is a certain degree of agreement between areas of large ensemble spread and large average error, suggesting that the ensembles are able to capture case dependent forecast uncertainty.

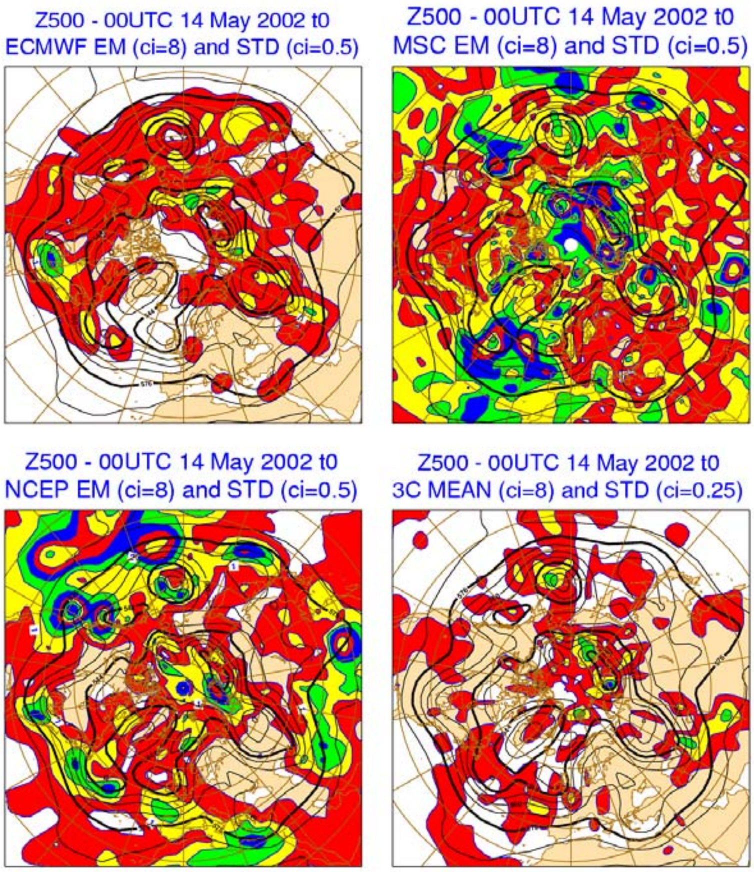

Figure 1. Initial state, 00UTC of 14 May 2002, 500 hPa geopotential height. (a-c) Ensemble mean and standard deviation (shading) of the (a) EC-EPS, (b) the MSC-EPS and (c) the NCEP-EPS. (d) average of the three ensemble-means and standard deviation among the three ensemble means (shading). Contour interval is 8 dam for full field, 0.5 dam for ensemble standard deviation in panels (a-c) and 0.25 dam for standard deviation in panel (d).

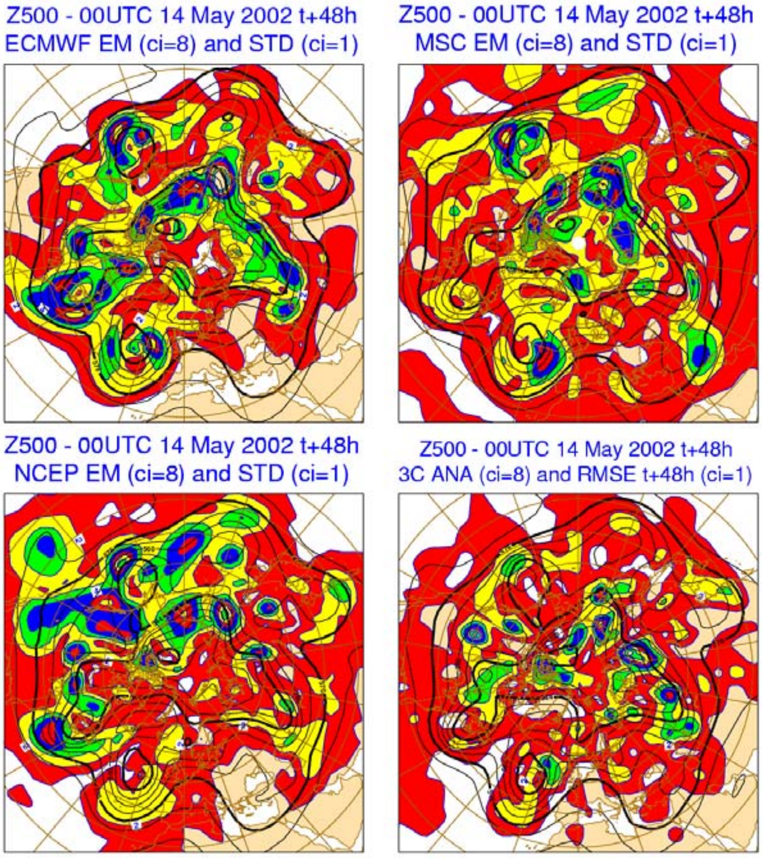

Fig. 2. As Fig 1 but for 48h forecast from 00UTC of 14 May 2002, and with average RMSE of the three control forecasts instead of the analysis STD in panel d, and with 1dam ci for shadings.

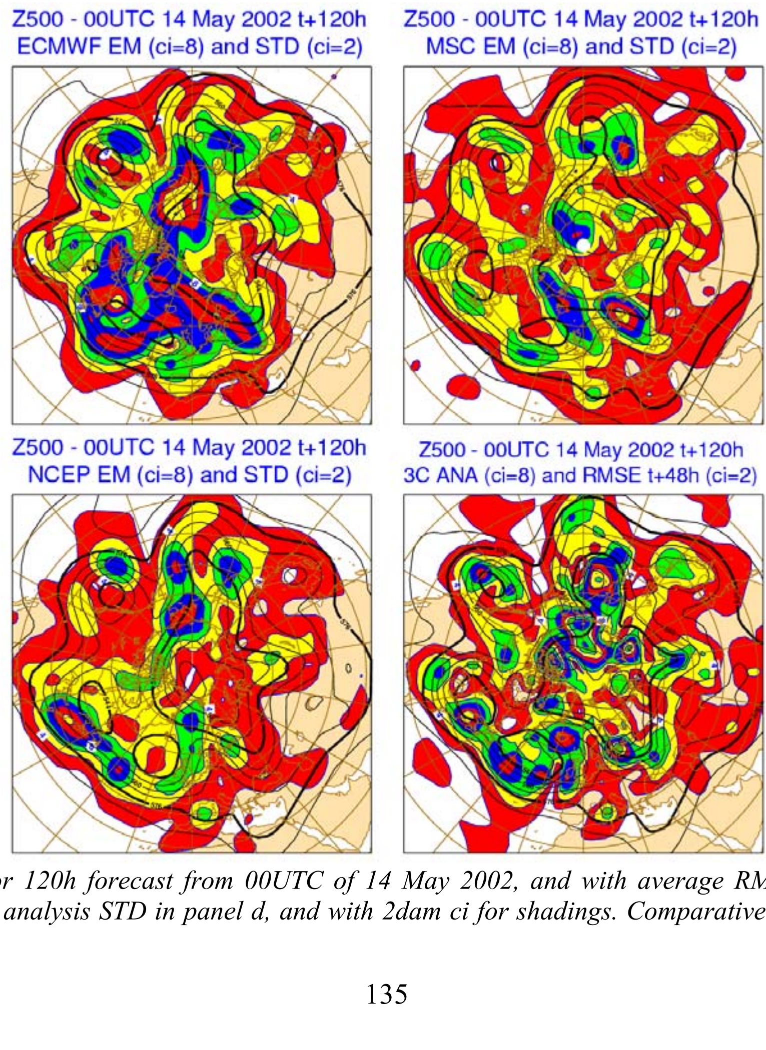

Fig. 3. As Fig 1 but for 120h forecast from 00UTC of 14 May 2002, and with average RMSE of the three control forecasts instead of the analysis STD in panel d, and with 2dam ci for shadings. Comparative verification of the three ensemble systems

3. Comparative verification of the three ensemble systems

3.1. 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 vs. 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 due to its larger spread that guaranteed a better agreement between the growth of ensemble spread measured, e.g., by the ensemble standard deviation, and the growth of the ensemble-mean error. *Mullen & Buizza* (2001) compared the skill of probabilistic precipitation forecasts based on 10 perturbed members over the US, using 24-hour accumulated precipitation data from the US NWS River Forecast Centers. They concluded that during 1998 the limit of skill (measured by the Brier skill score) for the EC-EPS was about 2-days longer for the 2- and 10-mm/d thresholds, while the two systems exhibited similar skill for 20mm/d 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-gage 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 towards establishing a continuous comparison effort based on a fairly extensive set of measures as discussed below.

3.2. The data-set used in the this study: 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: according to *Buizza & Palmer* (1998), increasing the ensemble size from 8 to 32 members in the old T63L19 EC-EPS system increased the skill of the ensemble-mean by ~6h in the medium-range, and the skill of probabilistic predictions by ~12h, with exact numbers function of the accuracy measure. The sub-sampling from the MSC-ensemble will have a negative impact on the quality of its initial ensemble mean fields, and the sub-sampling of the EC-EPS with initial perturbations designed for a 50-perturbed member system will have a negative impact on its skill for the whole forecast period. Furthermore, since in May-June-July 2002 ECMWF had no operational 00 UTC ensemble and MSC had no operational 12 UTC ensemble, for each day 00UTC MSC- and NCEP-EPS and the 12UTC ECMWF ensembles have been considered.

For brevity, only 500 hPa geopotential height forecasts are considered over the middle latitudes of the northern hemisphere (20º-80ºN, except for PECA, see below). Forecast and analysis fields have been interpolated onto a common regular 2.5x2.5 grid, and each ensemble has been verified against its own analysis, i.e. 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-NCAR reanalysis.

3.3. Verification attributes and measures

3.3.1. Attributes of forecast systems

The performance of the ensemble forecast systems will be measured through an analysis of their two main attributes: statistical reliability (or consistency) and resolution. Reliability of a forecast system 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 post-processing techniques. Though important for real world applications, reliability of a forecast system in itself does not guarantee usefulness. For example, a climatological forecast system, by definition, is perfectly reliable, yet has no forecast value. The real value of forecast systems is measured by their other main attribute, statistical resolution. Statistical resolution reflects a forecast system's ability to distinguish between different future events in advance.

3.3.2. Verification measures

There exist a number of verification scores that measure the two main attributes of forecast systems. These measures differ since they emphasize different aspects of forecast performance. In this study, the performance of the three EPS 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, and the Brier Skill Score (*Brier* 1950). All these scores provide a measure that is a function of both the statistical reliability and the resolution. In addition, results for outlier statistics, that measures only reliability, and relative operating characteristics (ROC, *Mason* 1982, *Stanski et al* 1989), that measures only resolution, will also be presented. A brief description of each of these measures is provided in the Appendix, while further details, along with a discussion of the two main attributes of probabilistic forecasts, resolution and reliability, can be found, e. g., in *Wilks* 1995, *Talagrand et al* 1997, and *Toth et al.* 2003.

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 3 EPS systems. To investigate this, *Wei & Toth* (2003) designed a new measure called perturbation vs. error correlation analysis (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 modelling, and ensemble perturbation techniques at the three centers.

3.4. Performance of the three ensemble systems for May-June-July 2002

3.4.1. Quality of data assimilation and numerical modelling 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 modelling at the three centers. Shown in Fig. 4 is the PAC score for each center's control forecast. Note that both ECMWF and NCEP has 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 initial condition equal to the ensemble mean. Due to 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 3D-Var analysis, is shown in Fig. 4. Results indicate that for this period the quality of the control forecast is highest for the EC-EPS and lowest for the MSC-EPS.

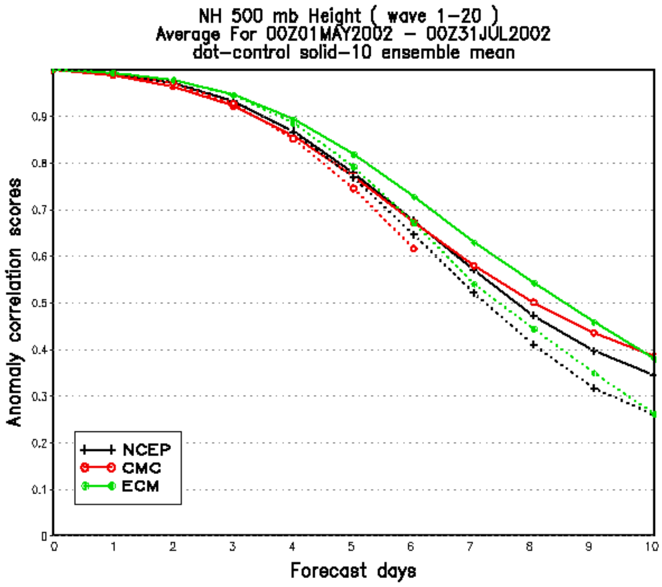

Figure 4. May-June-July 2002 average PAC for the control (dotted lines) and the ensemble-mean (solid lines) of the EC-EPS (green lines), the MSC-EPS (red lines) and the NCEP-EPS (black lines).Values refer to the 500 hPa geopotential height over the northern hemisphere latitudinal band 20º-80ºN.

3.4.2. Overall measures of ensemble performance

RMS error, and the related PAC are influenced by both systematic errors (related to reliability) and random error variance (related to resolution). Therefore these two scores offer good measures of overall forecast performance. In this subsection the value of each ensemble forecast system will be measured by PAC and RMS of the ensemble mean forecasts. For PAC, the ensemble skill will also be compared to that of the control forecasts. These scores will be complemented by the Brier Skill Score computed for probabilistic forecasts based on the three ensembles.

Except as noted below, each ensemble mean forecast is more skilful 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 hours at forecast day 6/9. These gains are due to the non-linear filtering effect that ensemble averaging offers in terms of error growth reduction (*Toth & 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 around an initial state that is inferior to the 3D-VAR analysis and to the sub-sampling to 10 members performed for this study.

Note also that beyond 5 days lead time 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/modelling 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.

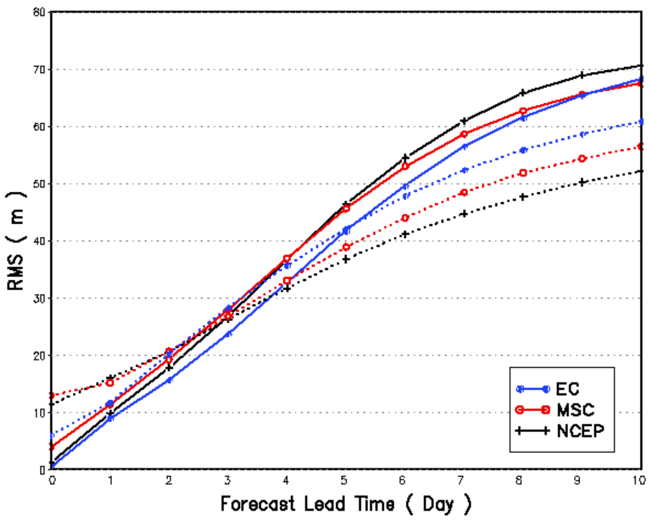

Figure 5. May-June-July 2002 average RMS error of the ensemble-mean (solid lines) and ensemble standard deviation (dotted lines) of the EC-EPS (blue lines), the MSC-EPS (red lines) and the NCEP-EPS (black lines). Values refer to the 500 hPa geopotential height over the northern hemisphere latitudinal band 20º-80ºN.

The BSS, shown in Fig. 6, is computed by averaging the BSS for 10 climatologically equally likely events, considering climatology as a reference forecast. Just as the RMS error and PAC, the Brier Skill Score (BSS) also reflects both the reliability and resolution of ensemble forecast systems. 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 due to the lack of model perturbations. The relatively good performance of the MSC system at long (8-10 days) lead time again may be due to the use of multiple model versions in that ensemble.

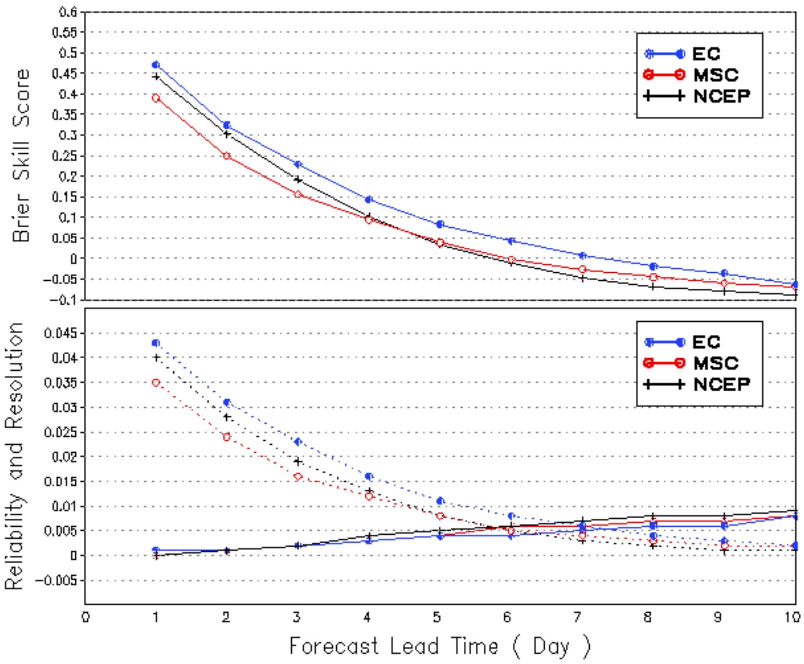

Figure 6. Top: May-June-July 2002 average Brier skill score for the EC-EPS (blue lines), the MSC-EPS (red lines) and the NCEP-EPS (black lines). Bottom: resolution and reliability contributions to the Brier skill score. Values refer to the 500 hPa geopotential height over the northern hemisphere latitudinal band 20º-80ºN, and have been computed considering 10 equally-climatologically-likely intervals.

The BSS can be decomposed into reliability and resolution components (*Murphy* 1973). Results from the lower panel of Fig. 6 indicate that at shorter lead times (before day 6) it is the resolution, while at longer lead times (beyond day 7) it is the reliability part of the BSS that dominates the overall results.

3.4.3. Measures of reliability

In this subsection statistical reliability will be assessed in three different ways. The first measure used is the discrepancy between the ensemble spread and the ensemble mean error, both shown for all three systems in Fig. 5. One should note here that the very small initial ensemble mean errors are a consequence of validating against analyses. 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 (ensemble mean error) should match that between the ensemble mean and a randomly selected ensemble member (ensemble standard deviation or spread). A large difference between the ensemble mean error and spread 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. During the first two days the EC-EPS perturbation growth even exceeds error growth. This is due to the use of SVs that are optimised for maximum growth. Due to 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 large growth rate of the EC-EPS in the 3-10 day range is mostly due to the sustained growth 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 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 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. An important observation based on Fig. 5 is that the perturbation growth is lower than the error growth even in the ensemble with the most realistic behavior (EC-EPS), indicating that none of the techniques currently used for representing model related uncertainty provides satisfactory performance. In the MSC- and NCEP-EPS, the deficiency in perturbation growth is partially compensated by initial perturbation amplitudes that are larger than the level of estimated initial errors.

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 June 2002. As a further reference, a linear measure of atmospheric instability, the Eady index (*Hoskins & Valdes* 1990), is also shown Figs. 7 and 8.

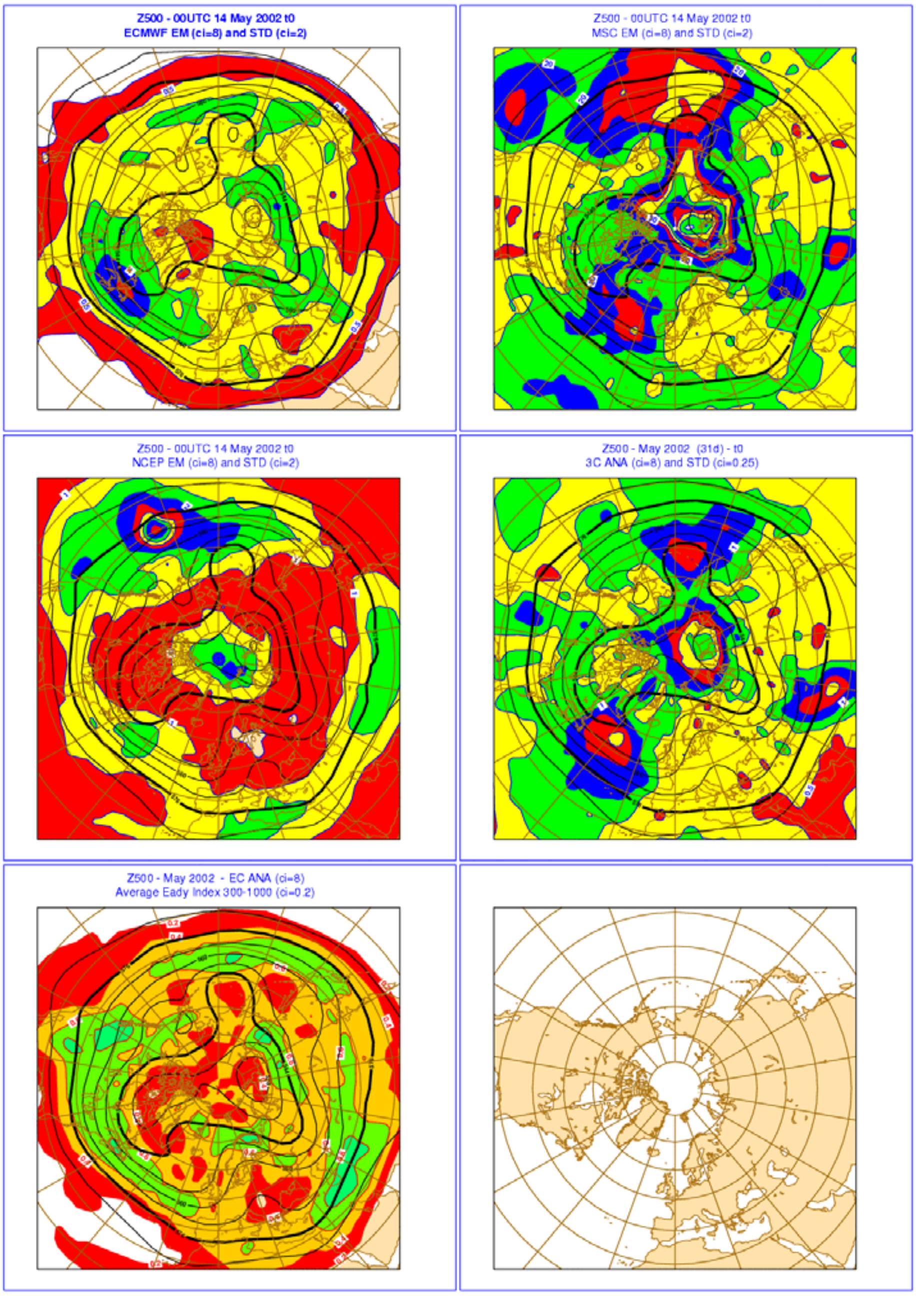

Figure 7. May-2002 initial-time average, 500 hPa geopotential height. (a-c) Ensemble mean and standard deviation (shading) of the (a) EC-EPS, (b) the MSC-EPS and (c) the NCEP-EPS, (d) average of the three ensemble-means and standard deviation among the three ensemble means (shading) and (e) average of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

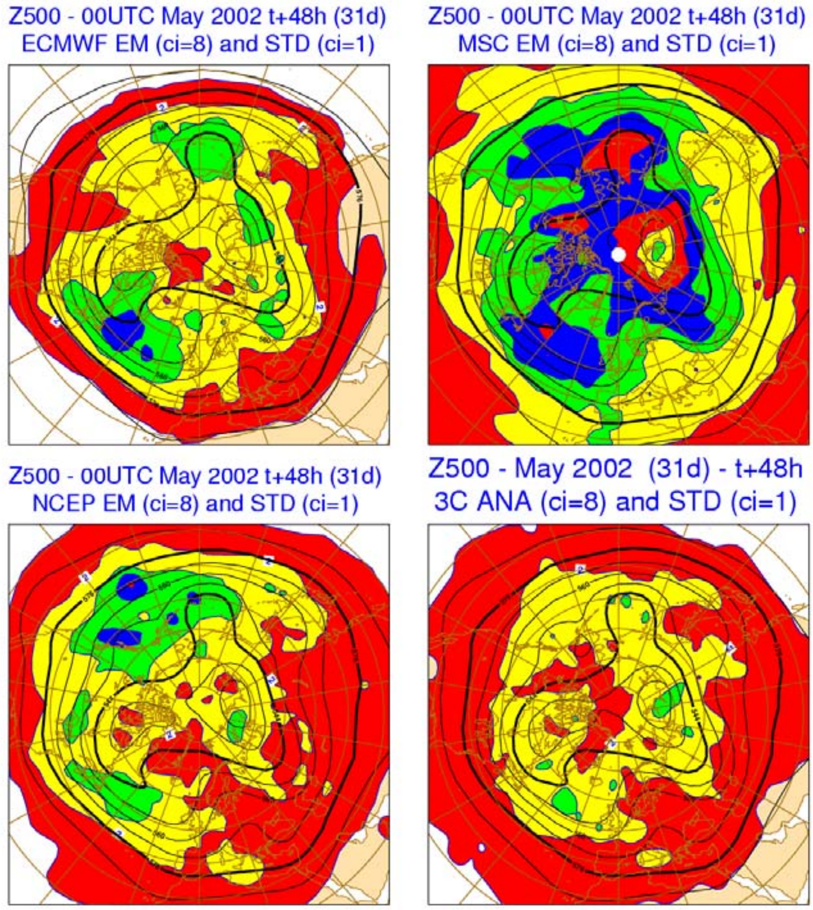

Figure 8. May-2002 +48-hour average, 500 hPa geopotential height. (a-c) Ensemble mean and standard deviation (shading) of the (a) EC-EPS, (b) the MSC-EPS and (c) the NCEP-EPS. (d) average of the three ensemble-means and standard deviation among the three ensemble means (shading). Contour interval is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

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

where N is the static stability and the wind shear is computed using the 300-1000 hPa potential temperature and wind.

As already noted in Figs. 5 and 1, 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, with absolute (relative) maxima over the Atlantic, Arctic (Pacific, Atlantic), and Atlantic (Arctic) regions in the EC-, MSC-, and NCEP-EPS respectively. The characteristics of the SV-based EC-EPS, and the BV-based NCEP-EPS perturbations are further discussed by *Buizza & Palmer* 1995, and *Toth & Kalnay* 1997, respectively. We only note here that the distribution of EC-EPS initial spread exhibits some similarity with the Eady Index (Fig. 7e). 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 in this subsection is the percentage of the number of cases when the verifying analysis 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 is used, since this measure *aggregates* (and not averages) results obtained at different grid points.

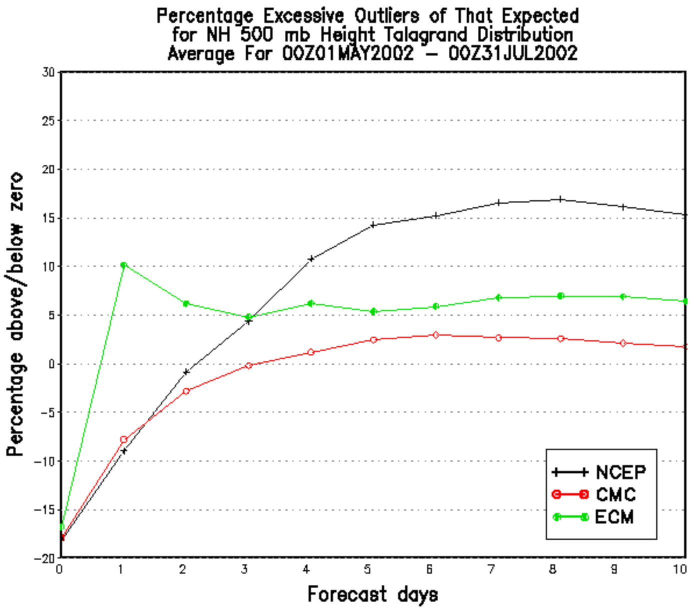

Figure 9. May-June-July 2002 average percentage of excessive outliers for the EC-EPS (green lines), the MSC-EPS (red lines) and the NCEP-EPS (black lines). Values refer to the 500 hPa geopotential height over the northern hemisphere latitudinal band 20º-80ºN.

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 due to 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 by considering that the MSC-EPS outlier results benefit from the use of ensemble members with distinctly different time mean errors (biases). Alternatively, one could argue that the MSC-sampling of model error is very appropriate but insufficient in the sense that not all model weaknesses are actually sampled. Perturbation growth may be influenced more 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 has contributed to the model-error terms.

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 score indicates that the EC-EPS is the least reliable at short, while the most reliable at longer lead times. These results are consistent with the outlier statistics at short, and the rms spread results at longer lead times respectively.

3.4.4. Measures of resolution

The resolution component of the BSS (lower panel of Fig. 6) and the Relative Operating Characteristics (ROC) Score (Fig. 10) provide two different quantitative measures of the inherent skill of the ensemble forecast systems, the statistical resolution. Though there are some quantitative differences between these two scores, they agree that the best resolution is obtained by the EC-EPS. At short, up to 2 days lead time the NCEP-EPS is competitive, probably due to the beneficial effects of initial perturbations (bred vectors). With increasing lead time, the resolution 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 due to the use of multiple model versions.

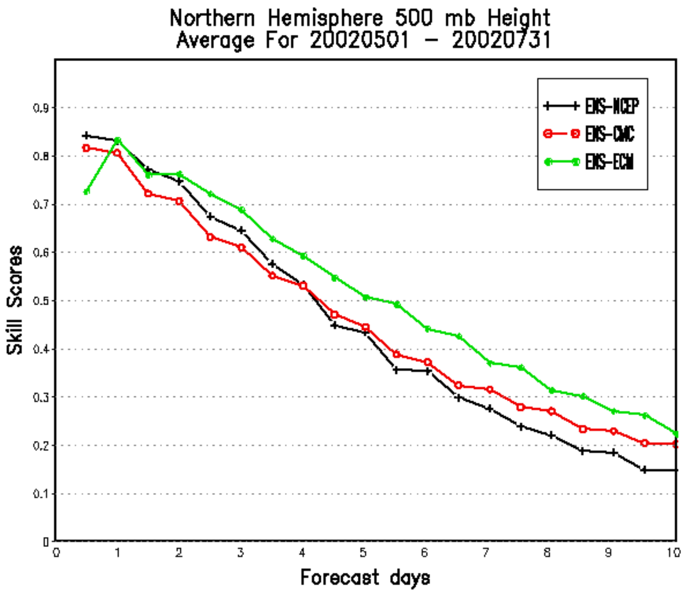

Figure 10. May-June-July 2002 area under the relative operating characteristics for the EC-EPS (green lines), the MSC-EPS (red lines) and the NCEP-EPS (black lines). Values refer to the 500 hPa geopotential height over the northern hemisphere latitudinal band 20º-80ºN, and have been computed considering 10 equally-cliamatologically-likely intervals.

3.4.5. Perturbation pattern analysis

All the previously discussed verification measures of reliability and resolution are used to evaluate ensemble forecasts for scalars. In contrast, the Perturbation vs. Error Correlation Analysis (PECA) evaluates directly the quality of ensemble *perturbation* patterns (Appendix A). This score is thought to be insensitive to the quality of the deterministic prediction system. The higher the correlation of individual, or optimally combined ensemble perturbations with the error in the control forecast, the closer the structure of the ensemble perturbations are to the forecast error structure.

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

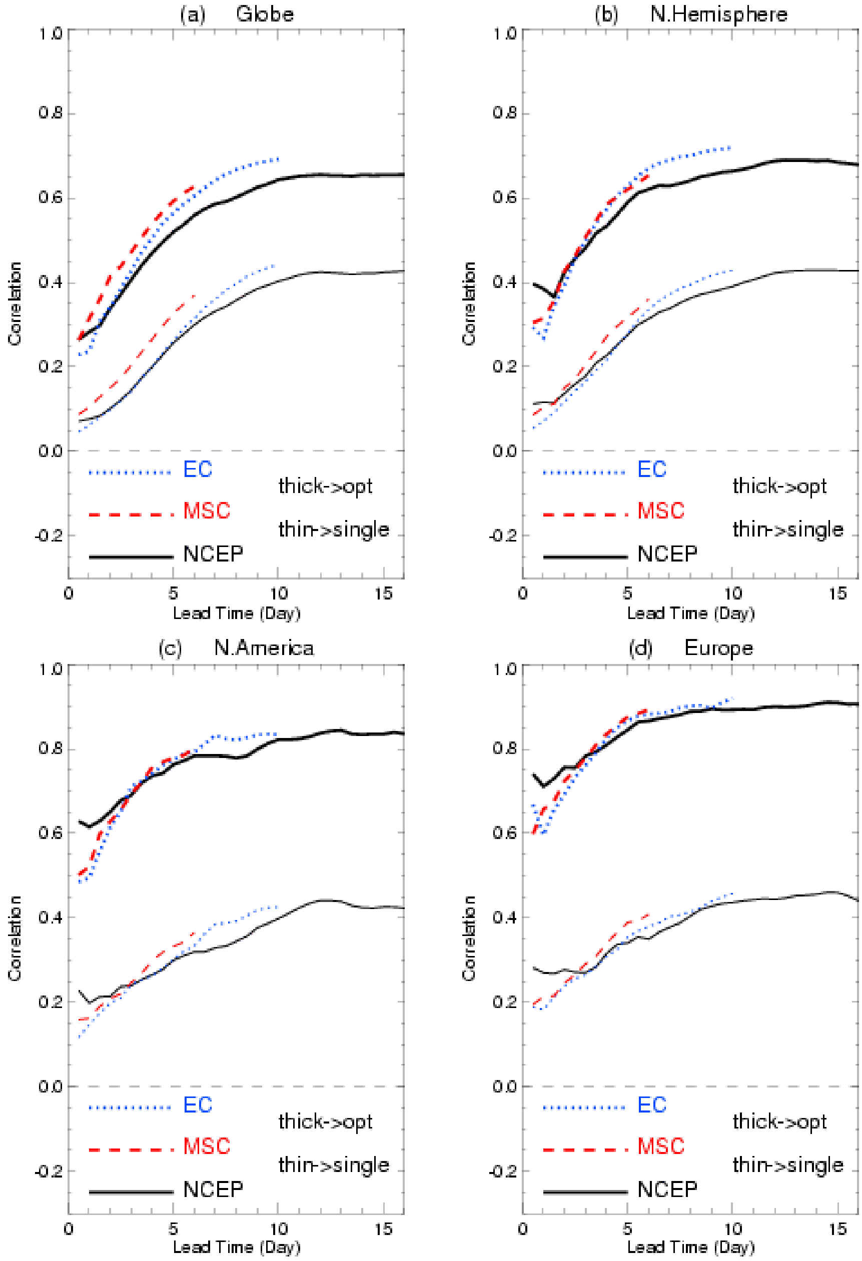

Figure 11. Perturbation vs. Error Correlation Analysis (PECA) for the global (top left), NH (top right), North American (bottom left) and European (bottom right) regions for individual (thin) and optimally combined (heavy) ensemble perturbations. For further details on PECA, see text and Appendix A.

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,c,d), the NCEP-EPS has a clear advantage for the short, 1 day forecast range. Over the North American/European region, the optimally combined NCEP-EPS perturbations, for example, can explain around 38/53% of the 12-24 hour 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 5th day, the differences between the NCEP-EPS, and 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 hour) 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 in the computation of the singular vector perturbations, total energy, 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.

4. Future directions

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

4.1. 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 non-growing directions, or only in the fast growing sub-space 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.

4.2. 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?

4.3. Unification of data assimilation and ensemble generation techniques

Ensemble forecasting and data assimilation efforts can mutually benefit from each other and the two systems can be jointly designed (*Houtekamer et al.* 1996b). 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 modelling applications (*Toth et al.* 2003).

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

5. Ongoing research

Out of a wider array of activities, the following is a list of the most critical developmental tasks carried out at the three centers.

At ECMWF, ensemble developmental efforts are focused on:

- Improving the description of initial uncertainties through the use of more physical processes in the computation of extratropical singular vectors (Coutinho et al. 2003), and by the use of different norms in the singular vector computation;
- Increasing the resolution of the ensemble system especially for the prediction of severe weather events;
- Revising the scheme used to simulate random errors due to physical processes, and developing a more complex scheme to simulate model error, e.g., by using forcing singular vectors (Barkmeijer et al. 2003).

At MSC, the main areas of ensemble related research are:

- The development of an ensemble Kalman filter that will provide the initial conditions for the MSC-EPS (*Mitchell et al*. 2002);
- The identification of the main uncertain parameters of the physical parametrizations, so that different members can use different values for these parameters;
- The development of products based on an NCEP-MSC superensemble.

NCEP focuses its ensemble related activities on:

- Increasing the horizontal resolution (to T126 resolution out to 180 hours) and the frequency (4 times a day) of ensemble forecasts for increased forecast skill;
- Enhancing the generation of initial perturbations through modifications to the breeding algorithm (6-hour cycling, use of the Ensemble Transform Kalman Filter for rescaling the perturbations, *Wang & Bishop* 2003);
- Long-term efforts aimed at the development of an NWP model capable of simulating different types of model related uncertainty in ensemble applications;
- Statistical post-processing of ensemble forecasts for the reduction of systematic errors; and the combination of bias-corrected ensemble forecasts originating from different NWP centers (including ECMWF and MSC), for the generation of probabilistic weather forecast products.

6. 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 global modelling suite. Beyond ECMWF, MSC, and NCEP, global ensemble forecasting is currently operational at the Fleet Numerical Meteorology and Oceanography Center (FNMOC, *Rennick* 1995); the Korean Meteorological Agency (*KMA* 2001), the Japan Meteorological Agency (*Kyouda* 2002) and the Bureau of Meteorology in Australia (*Bourke et al.* 2003). In addition to the centers mentioned above, global ensemble forecasting systems have also been adapted and tested at the South African Weather Service (*Tennant* 1997, personal communication), the National Centre for Medium-Range Weather Forecasting in India (Iyengar 1998, personal communication) and the UK Meteorological Office (*Harrison et al.* 1999).

Since the mid 1990s, the ensemble techniques have also gained ground in limited area modelling applications both in research (Meteo-France, *Nicolau* 2001; the Royal Dutch Meteorological Institute, *Hershbach et al.* 2000) and operations (at NCEP, *Du & Tracton* 2001, and at the Regional Meteorological Service of Emilia Romagna, *Marsigli et al.* 2001 and *Molteni et al.* 2001). Various ensemble products based on global or regional ensemble applications established themselves as indispensable tools for a large group of users in public and private weather forecasting operations.

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 under-estimation of the actual skill of the ensemble systems, especially for systems with a large member-ship, *Buizza & 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 value of ensemble forecasts influenced by the quality of the data assimilation, numerical weather prediction modelling, and ensemble generation schemes. The results therefore are not directly indicative of the value 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 patters, 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 emerge on the following aspects of the systems:

i. The EC-EPS gave overall the most skilful performance when measured using all (RMS, PAC, BSS and ROC-area) but PECA and the outlier statistic as performance measure.

ii. Using PECA to measure the correlation between perturbation and forecast-error patters, the EC-EPS does not show any superior performance. By contrast, the breeding technique at NCEP describes better error patterns for small areas at the short-range, and the MSC-EPS describes them better in the medium forecast range (say for forecast day 3 to 5).

iii. These two 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.

iv. The multiple model versions technique used at MSC provides added value at longer lead times by capturing large-scale model related errors in the forecasts. This could explain the superior outlier statistics of the MSC-EPS.

v. After forecast day 3, which is the time when the three systems have approximately a comparable ensemble spread, the spread of the multi-model MSC-EPS does not grow as fast as forecast error grow. By contrast, the spread in the single-model EC-EPS grows faster, due to a combine effect of sustained SV-based perturbations' growth and of the stochastic simulation of random model errors. This indicates that the addition of stochastic model perturbations at ECMWF improves the statistical reliability of the forecasts.

vi. The relatively low quality of the ensemble of data-assimilation cycles in the MSC-EPS has a negative impact on the quality of the ensemble mean of that system.

During the past decade different ensemble generation techniques received significant attention and underwent substantial refinements. Yet a number of open questions still remain. On-going ensemble related research in the coming years is expected to provide a better understanding of the still remaining scientific issues. The inter-comparison 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 further useful information, contributing to improved forecast operations.

Acknowledgements

The emergence of ensemble forecasting, as of 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.

Appendix A: performance measures

Root Mean Square error (RMS). The RMS is one of the most commonly used measures of forecast error. It is defined as the root of the area mean squared difference between an (ensemble mean) forecast field *e(t)* and the corresponding verification field (either an observed or analysed field) *o(t)*:

$$RMS = \sqrt{< e(t) - o(t); e(t) - o(t) >} = \left\| e(t) - o(t) \right\| \tag{A.1}$$

As a norm (and corresponding inner product <..;..>), the Euclidean inner product is used. Higher RMS errors indicate poorer forecast quality.

Pattern Anomaly Correlation (PAC). The PAC measures the correlation between predicted and observed fields. It is defined as:

$$PAC = \frac{< (e(t) - c); (o(t) - c) >}{\left\| e(t) - c \right\| \cdot \left\| o(t) - c \right\|} \tag{A.2}$$

where *c* denotes the climatology. The more the forecast anomaly taken from the climate mean correlates with the observed anomaly, the higher the PAC is. A forecast with a perfect anomaly pattern would have PAC=1.

Outlier statistics. An ordered series of N ensemble values define N+1 intervals, including two open ended bins at the extremes. With an N-member ensemble one may define N + 1 intervals at every gridpoint. For a reliable ensemble the likelihood of the verifying analysis falling into either of the two extreme intervals is

2/(N+1). The excessive number of outliers measures the percentage of cases (averaged over all gridpoints and forecast cases) where the verifying analysis falls in the two extreme categories *in excess* of the value 2/(N+1) expected for a reliable ensemble. Larger/smaller than zero values indicate too many/few outliers (too small/large ensemble spread).

Brier Skill Score (BSS). The Brier Score (BS) averages the squared difference between pairs of forecast probabilities and a corresponding binary observation variable with a value of 1 (the forecast event occurred) or 0 (did not occur). The Brier Skill Score is defined with respect to the Brier Score of a reference (often climatology based) forecast system (BS_{ref}):

$$BSS = \frac{BS_{ref} - BS}{BS_{ref}} \tag{A.3}$$

A BSS value of 1 (higher/lower than zero) indicates perfect (better/worse than reference) performance.

Relative Operating Characteristics (ROC). The ROC is defined using a two-category contingency table. The hits are defined as the percentage of forecasts of an event that verify and the false alarms are the percentage of forecasts of the event that did not verify. In case of an ensemble, forecasts for the occurrence of an event can be issued depending on the number of the total of N ensemble members predicting the event. This yields N progressively more stringent decision criteria. The *N* pairs of hit and false alarm rates define the ROC points. An ROC curve is generated by connecting these points to each other and to the points (0,0) and (1,1). The area on the right of the curve defines the ROC area (ROC_{area}). Perfect/random/no skill corresponds to a ROC area of 1/0.5/0. The ROC Skill Score (RSS) is defined as $RSS=(ROC_{area}-0.5)*2$.

PECA (Perturbation versus Error Correlation Analysis). PECA is defined as the correlation between perturbation and forecast error fields:

$$A_c(X,Y) = \frac{<X,Y>}{\|X\|\cdot\|Y\|}, \quad X = P_i^c \quad or \quad X = P_{optimal}^c, \quad Y = E^c \tag{A.4}$$

where:

$$\begin{aligned} E^c(t) &= F_{control}^c(t) - A^c(t) \\ P_i^c(t) &= F_i^c(t) - F_{control}^c(t) \\ P_{optimal}^c &= \sum_{j=1}^{n} P_j^c \end{aligned} \tag{A.5}$$

where $F_{control}^c(t)$, $F_i^c(t)$ and $A^c(t)$ are the control and perturbed forecasts, and corresponding verifying analysis fields at time t, C is the originating forecast center, and $i = 1,...,n$ is the number of ensemble members. For the a posteriori determined combination of ensemble perturbations optimized to explain the maximum error variance in the control forecast, α_j is defined by the solution of the least-square problem:

$$\min\left| E^c - \sum_{j=1}^{n} \alpha_j P_j^c \right|_{L2} \tag{A.6}$$

PECA provides a measure of how well individual or optimally combined ensemble perturbations can explain forecast error variance. As defined, this measure evaluates the performance of ensemble perturbations, not the full forecast fields. By reducing the influence of the magnitude of initial errors (that reflect the quality of the analysis scheme), PECA offers a more direct measure of ensemble performance. The higher the PECA values, the more closely ensemble perturbations, on average, are correlated with forecast error.

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.

Atger, F., 2001: Verification of intense precipitation forecasts from single models and ensemble prediction systems. Nonlinear Processes in Geophysics, 8, 401-417.

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

Barkmeijer, J., Iversen, T., & Palmer, T. N., 2003 : Forcing singular vectors and other sensitivity model perturbations. *Q. J. R. Meteor. Soc.*, in press.

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

Bourke, W., R. Buizza, & M. Naughton, 2003: Performance of the ECMWF and the BoM Ensemble Systems in the Southern Hemisphere. *Mon. Wea. Rev.*, submitted.

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

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

Buizza, R., & T.N. Palmer, 1998: Impact of ensemble size on ensemble prediction. *Mon. Wea. Rev.*, **126**, 2503-2518.

Buizza, R., T. Petroliagis, T. N. Palmer, J. Barkmeijer, M. Hamrud, A. Hollingsworth, A. Simmons, & N. Wedi, 1998: Impact of model resolution and ensemble size on the performance of an ensemble prediction system. *Q. J. R. Meteorol. Soc.*, 1**24**, 1935-1960.

Buizza, R., M. Miller, & T. N. Palmer, 1999: Stochastic representation of model uncertainties in the ECMWF ensemble prediction system. *Q. J. R. Meteorol. Soc.*, **125**, 2887-2908.

Buizza, R., D. S. Richardson, & T. N. Palmer, 2003: Benefits of increased resolution in the ECMWF ensemble system and comparison with poor-man's ensembles. *Q. J. R. Meteorol. Soc.*, **129**, 1269-1288.

Caplan, Peter, John Derber, William Gemmill, Song-You Hong, Hua-Lu Pan, David Parrish, 1997: Changes to the 1995 NCEP Operational Medium-Range Forecast Model Analysis-Forecast System. Weather and Forecasting: Vol. 12, No. 3, pp. 581-594.

Côté, J., S. Gravel, A. Méthot, A. Patoine, M. Roch & 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, & R. Buizza, 2003: The influence of physical processes on extra-tropical singular vectors. *J. Atmos. Sci.*, accepted for publication.

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

Du, J., & M. S. Tracton, 2001: Implementation of a real-time short-range ensemble forecasting system at NCEP: an update. Preprints, 9th Conference on Mesoscale Processes, Ft. Lauderdale, Florida, Amer. Meteor. Soc., 355-356.

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

Ehrendorfer, M., 1994b: The Liouville equation and its potential usefulness for the prediction of forecast skill. Part II: applications. *Mon. Wea. Rev*., **122**, 714-728.

Ehrendorfer, M., & 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), 10143 –10162.

Fleming, R. J., 1971a: On stochastic dynamic prediction. I: the energetics of uncertainty and the question of closure. *Mon. Wea. Rev.*, **99**, 851-872.

Fleming, R. J., 1971b: On stochastic dynamic prediction. II: predictability and utility. *Mon. Wea. Rev.*, **99**, 927-938.

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

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

Harrison, M. S. J., T. N. Palmer, D. S. Richardson, & R. Buizza, 1999: Analysis and model dependencies in medium-range ensembles: two transplant case-studies. *Q. J. R. Meteorol. Soc.*, **125**, 2487-2515.

Hersbach, H., R. Mureau, J. D. Opsteegh, & J. Barkmeijer, 2000: A short-range to early-medium-range Ensemble Prediction System for the European Area. *Mon. Wea. Rev*., **128**, 3501 – 3519.

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

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

Houtekamer, P. L., L. Lefaivre & J. Derome, 1996a: The RPN ensemble prediction system. *Proc. ECMWF seminar on predictability, Vol II*, Reading, U.K., ECMWF, 121-146.

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

Houtekamer, P L & L. Lefaivre, 1997: Using ensemble forecasts for model validation. *Mon. Wea. Rev*, **125**, 2416-2426.

lyengar, G., Z. Toth, E. Kalnay, & J. Woollen, 1996: Are the bred vectors representative of analysis errors? Preprints of the 11th AMS Conference on Numerical Weather Prediction,19-23 August 1996, Norfolk, Virginia, p. J64-J66.

Kyouda, M., 2002: Ensemble prediction system. In: Outline of the operational NWP at the JMA. Japan Meteorological Agency, Ed.: T. Fujita, p. 59-63. [Available from JMA]

KMA, 2001: Numerical Weather Prediction. Korea Meteorological Administration, Newsletter, Vol. 2, No. 1.

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

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

Lorenz, E N, 1982: Atmospheric predictability experiments with a large numerical model. *Tellus*, **34**, 505-513.

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

Marsigli, C., A. Montani, F. Nerozzi, T. Paccagnella, S. Tibaldi, F. Molteni, & R. Buizza, 2001. A strategy for high-resolution ensemble prediction. Part II: limited-area experiments in four Alpine flood events. *Q. J. R. Meteorol. Soc.*, **127**, 2095-2115.

Mason, I., 1982. A model for assessment of weather forecasts. *Aust. Met. Mag.*, **30**, 291-303.

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

Mitchell, H. L., P. L. Houtekamer & G. Pellerin, 2002: Ensemble size, balance and model-error representation in an ensemble Kalman filter. *Mon. Wea. Rev*., **130**, 2791-2808.

Molteni, F., R. Buizza, T. N. Palmer, & T. Petroliagis, 1996: The ECMWF ensemble prediction system: methodology and validation. *Q. J. R. Meteorol. Soc.*, **122**, 73-119.

Molteni, F., R. Buizza, C. Marsigli, A. Montani, F. Nerozzi, & T. Paccagnella, 2001. A strategy for high-resolution ensemble prediction. I: definition of representative members and global-model experiments. *Q. J. R. Meteorol. Soc.*, **127**, 2069-2094.

Mullen, S. L., & 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. Apll. Met., 12, 534-537.

Nicolau, J., 2001: Short-range ensemble forecasting at Meteo-France: a preliminary study. European Geophysical Society XXVI General Assembly, Nice 25-30 March 2001, EGS Newsletter n. 78, pg 251.

Palmer, T. N., F. Molteni, R. Mureau, & R. Buizza, 1993: Ensemble prediction. *ECMWF Seminar Proceedings 'Validation of models over Europe: Vol. I'*, available from ECMWF, Shinfield Park, Reading RG2-9AX, UK.

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

Pitcher, E. J., 1977: Application of stochastic dynamic prediction to real data. *J. Atmos. Sci.*, **34**, 3-21.

Rennick, M. A., 1995: The ensemble forecast system (EFS). Models Department Technical Note 2-95, Fleet Numerical Meteorology and Oceanography Center. p. 19. [Available from: Models Department, FLENUMMETOCCEN, 7 Grace Hopper Ave., Monterey, CA 93943.]

Richardson, D. S., 2000: Skill and relative economic value of the ECMWF Ensemble Prediction System. *Q. J. R. Meteorol. Soc.*, **126**, 649-668.

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

Smith, L. A., M. S. Roulston, & J. von Hardenberg, 2001: End to end ensemble forecasting: Towards evaluating the economic value of the Ensemble Prediction System. *ECMWF Research department Technical Memorandum n. 336*, available from ECMWF, Shinfield Park, Reading RG2-9AX, UK.

Stanski, H. R., L. J. Wilson, & W. R. Burrows, 1989. Survey of common verification methods in meteorology. *World Weather Watch Technical Report No. 8*, WMO/TD. No. 358, World Meteorological Organization. 114 pp.

Talagrand, O., R. Vautard, & B. Strauss, 1997: Evaluation of probabilistic prediction systems. Proceedings of the ECMWF *Workshop on Predictability*, 20-22 October 1997, ECMWF, Shinfield Park, Reading RG2-9AX, 1-26.

Taylor, J. W., & R. Buizza, 2003: Using weather ensemble predictions in electricity demand forecasting. *International Journal of Forecasting*, **19**, 57-70. Also published in 2001 as *ECMWF Research department Technical Memorandum n. 312*, available from ECMWF, Shinfield Park, Reading RG2-9AX, UK.

Toth, Z. & E. Kalnay, 1993: Ensemble forecasting at NMC: The generation of perturbations. *Bull. Amer. Meteor. Soc.*, **74**, 2317-2330.

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

Toth, Z., I. Szunyogh, E. Kalnay, & G. Iyengar, 1999: Comments on: "Notes on the appropriateness of 'bred modes' for generating initial perturbations". Tellus, 51A, 442-449.

Toth, Z., & S. Vannitsem, 2002: Model errors and ensemble forecasting. Proceedings of the 8th ECMWF Workshop on Meteorological Operational Systems. November 12-16, 2001, Reading, England, 146-154.

Toth, Z., Y. Zhu, I. Szunyogh, M. Iredell, & R. Wobus, 2002: Does increased model resolution enhance predictability? Preprints for the Symposium on Observations, Data Assimilation, and Probabilistic Prediction, 13-17 January 2002, Orlando, FL, J18-23.

Toth, Z., O. Talagrand, G. Candille, & Y. Zhu, 2003: Probability and ensemble forecasts. In: Forecast Verification: A practitioner's Guide in Atmospheric Science. Ed.: I. T. Jolliffe and D. B. Stephenson. Wiley, 137-163.

Wang, X., & C. H. Bishop, 2003: A comparison of breeding and Ensemble Kalman Filter ensemble forecast schemes. JAS, 60, 1140-1158.

Wei, M., & 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, Inc., San Diego, pp. 467 (ISBN 0-12-751965-3).

Zhu, Y, G. lyengar, Z. Toth, M. S. Tracton, & T Marchok, 1996: Objective evaluation of the NCEP global ensemble forecasting system. Preprints of the 15th AMS Conference on Weather Analysis and Forecasting, 19-23 August 1996, Norfolk, Virginia, p. J79-J82.

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