TECHNICAL MEMORANDUM

246

Impact of ensemble size on ensemble prediction

R. Buizza and T. Palmer

Research Department

January 1998

European Centre for Medium-Range Weather Forecasts
Europäisches Zentrum für mittelfristige Wettervorhersage
Centre européen pour les prévisions météorologiques à moyen

Impact of Ensemble Size on Ensemble Prediction

R Buizza and T N Palmer

European Centre of Medium Range Weather Forecasts

Submitted to *Mon. Wea. Rev.*:28 May 1997

Revised version: 17 November 1997

ABSTRACT

The impact of ensemble size on the performance of the European Centre for Medium- Range Weather Forecasts (ECMWF) Ensemble Prediction System (EPS) is analyzed. The skill of ensembles generated using 2, 4, 8, 16 and 32 perturbed ensemble members are compared for a period of 45 days, from 1 October to 15 November 1996. For each ensemble configuration, the skill is compared with the *potential skill*, measured by randomly choosing one of the 32 ensemble members as verification (idealized ensemble). Results are based on the analyzes of the prediction of the 500 hPa geopotential height field. Various measures of performance are applied: skill of the ensemble mean, spread-skill relationship, skill of most accurate ensemble member, Brier score, Ranked Probability Score, Relative Operating Characteristic, and the Outlier Statistic.

The relation between ensemble spread and control error is studied using L^2, L^8, and L^∞ norms to measure distances between ensemble members and the control forecast or the verification. It is argued that the supremum norm is a more suitable measure of distance, given the strategy for constructing ensemble perturbations from rapidly-growing singular vectors. Results indicate that, for the supremum norm, any increase of ensemble size within the range considered in this paper is strongly beneficial. With the smaller ensemble sizes, ensemble spread does not provide a reliable bound on control error in many cases. By contrast, with 32 members, spread provides a bound on control error in nearly all cases. It could be anticipated that further improvement could be achieved with higher ensemble size still. On the other hand, spread skill relationship was not consistently improved with higher ensemble size using the L^2 norm.

The overall conclusion is that the extent to which an increase of ensemble size (particularly from 8 to 16, and 16 to 32 members) improves EPS performance, is strongly dependent on the measure used to assess performance. In addition to the spread/skill relationship, the measures which are most sensitive to ensemble size are shown to be the skill of the best ensemble member (particularly when evaluated on a point-wise basis), and the Outlier Statistic.

1. INTRODUCTION

Since December 1992, both the U. S. National Meteorological Center (NMC, now National Center for Environmental Prediction, NCEP) and ECMWF have supplemented their deterministic high-resolution forecast integration with a medium-range ensemble prediction (*Tracton and Kalnay*, 1993; *Palmer et al*, 1993). The development follows the theoretical and experimental work of *Epstein* (1969a), *Gleeson* (1970), *Fleming* (1971a-b) and *Leith* (1974).

Ensemble prediction provides an estimate of the forecast probability distribution of atmospheric states. The extent to which this estimate is reliable will depend on ensemble size. Because of limited computer resources, ensemble size must necessarily be very much smaller than the dimension of the phase space of the numerical weather prediction model used to perform the integrations. The purpose of this paper is to try to assess the impact of ensemble size on ensemble performance, for relatively small ensemble sizes. Specifically, five ensemble configurations based on 2, 4, 8, 16 and 32 perturbed members are compared over a 45-day period, from 1 October to 14 November 1996. A number of different performance measures are applied: ensemble-mean skill, spread/skill relationship, Brier Score, Outlier Statistic, Ranked Probability Score and Relative Operating Characteristic.

Individually, these measures do not necessarily prove that one configuration is superior to another (for example, the skill of the ensemble mean cannot be used to assess ensemble dispersion). However, together these measures provide a reasonably complete assessment of ensemble performance.

One important part of the paper concerns the assessment of possible relationships between spread and skill. We consider a hierarchy of distance norms for defining these quantities. At one end of the hierarchy lies the familiar root-mean-square distance. At the other end lies the supremum norm. We argue that the latter is a more appropriate norm given the ECMWF strategy of perturbing with fast-growing singular vector perturbations (*Buizza and Palmer*, 1995; *Molteni et al*, 1996).

Following *Buizza* (1997), we also assess the skill of a size-N ensemble in idealized mode, taking as verification an ensemble member randomly chosen among the 32 perturbed members.

The paper is organized as follows. The ensemble configurations are described in section 2. The methodology used to assess ensemble skill is discussed in Section 3. The performance of the different ensemble configurations are compared in Section 4. Conclusions are drawn in Section 5.

2. THE ENSEMBLE CONFIGURATIONS

Up until December 1996, the ECMWF EPS comprised 32 perturbed and one unperturbed (control) non-linear integrations of a version of the ECMWF model (*Simmons et al*, 1989; *Courtier et al*, 1991) with spectral triangular truncation T63 and 19 vertical levels (T63L19). The reader is referred to *Molteni et al* (1996) for a more complete description of the EPS.

The initial conditions of the 32 perturbed members are created by adding and subtracting perturbations to the control initial conditions. These perturbations are defined using the singular vectors (*Buizza and Palmer*, 1995) of the tangent propagator of the ECMWF model. The singular vectors identify the most unstable directions of the phase space of the system growing over a finite interval (the optimisation time interval), where the growth of any perturbation is computed as the ratio between the perturbation total energy at optimisation and initial time (i.e. a total energy norm is used).

The singular vectors are computed at T42L19 resolution with an optimisation time interval of 48 hours (*Buizza et al*, 1997), following a time evolving trajectory computed applying the complete ECMWF physical package, but using only a linear surface drag and vertical diffusion scheme (*Buizza*, 1994) when computing tangent forward and adjoint integrations.

Selection criteria are applied to select only 16 of the computed singular vectors. The selection criteria are such that the first 4 singular vectors are always chosen, and each subsequent singular vector (from the 5th onwards) is selected only if half of its total energy lies outside the regions where the singular vectors already selected are localized.

Once the 16 singular vectors have been selected, an orthogonal rotation in phase-space and a final re-scaling are performed to construct the ensemble perturbations. The purpose of the phase-space rotation is to generate perturbations which have the same globally-averaged energy as the singular vectors, but smaller local maxima and a more uniform spatial distribution. Moreover, the rotation ensures that all perturbations grow linearly at a similar rate. The rotation is defined to minimize the local ratio between the perturbation amplitude and the amplitude of the analysis error estimate given by the ECMWF 3D-var assimilation procedure (note that the estimate was given by the ECMWF Optimum Interpolation procedure up to 30 January 1996, time at which 3D-var became operational). At the time of this study, the re-scaling allowed perturbations to have local maxima up to $\alpha = \sqrt{2}$ larger than the local maxima of the analysis error estimate.

The 16 perturbations are added and subtracted to the control initial conditions (defined by the ECMWF 3D-var data assimilation procedure, *Courtier et al*, 1997) to define 32 perturbed initial conditions. Then, 32+1 (control) 10-day T63L19 non-linear integrations are performed.

The different ensemble configurations in this paper have been generated by sampling pairs of perturbed members, so that each ensemble contains perturbed members with initial conditions generated by adding and subtracting the same perturbation. Moreover, the ensembles have been generated so that, if $N1>N2$, configuration ens_N1 constitutes an extension of ens_N2, in the sense that it includes all the perturbed members of ens_N2, plus another $(N1-N2)$ members.

The skill of any configuration is computed using the operational ECMWF analysis as verification, while the *potential skill* of any configuration is computed considering a randomly chosen ensemble member of ens_32 as the verification. The type of ensemble arising when one member is chosen at random as the verification is often referred to as a "perfect-model ensemble". However, in addition to the removal of model error, this type of perfect-model ensemble has a fundamentally different skill characteristic to the actual ensemble. The ECMWF ensemble perturbations have a specified amplitude which can be taken to be approximately (but somewhat greater than) one standard deviation of the probability distribution of analysis error in the (phase space) direction chosen. As such, irrespective of model error, there is a definite non-zero probability that the verification lies outside the span of the ensemble (in the hypothesis of a Gaussian distribution, it would be about 32% if the perturbations were at precisely one standard deviation in amplitude). By contrast, in the perfect model ensemble the amplitude of the control forecast error is never larger than that of the forecast perturbations. Because of this important difference, we used the term *idealised ensemble* rather than "perfect-model ensemble".

3. VERIFICATION METHODOLOGY

The verification of a probability forecasting system is more complex than the verification of a single deterministic forecast (*Stanski et al*, 1989, *Wilks*, 1995). Thus, as for example in *Zhu et al* (1996), a complete analysis of an ensemble system should include probabilistic measures such as the Brier score (*Brier*, 1950), the ranked probability skill score (*Epstein*, 1969b), and other measures defined by signal detection theory (*Mason*, 1982). A brief explanation of the measures used in this paper is given below. Attention is focused on the 500 hPa geopotential height, and over two regions, the Northern Hemisphere extra-tropics (NH, region with latitude $\Phi \geq 30°N$), and Europe (latitude $30°N \leq \Phi \leq 75°N$, longitude $20°W \leq \Lambda \leq 45°E$).

3.1 Ensemble spread and forecast skill

Two of the most basic quantities used to verify an ensemble system are ensemble dispersion, or spread, and the skill of the control or ensemble-mean forecast. Arguably, the primary purpose of ensemble prediction is to use the former quantity as a predictor of the latter quantity. In some sense one expects situations when the ensemble dispersion is large to correspond to relatively poor forecast skill, and cases where ensemble dispersion is small, to correspond to relatively high forecast skill. It is commonplace to define skill in terms of root mean square error, which would correspond to choosing an L^2 distance function between a forecast and its verifying analysis. If such a distance function is used to define skill, then, for consistency, it should also be used in the definition of spread.

As noted by *Barker* (1991), there is no reason to expect a perfect correlation between spread and skill as determined, for example, by an L^2 measure. In particular, whilst small spread should indicate small control error, large spread may not necessarily imply large control error; it is possible for the spread to be relatively large, yet the control error to be relatively small. In this sense, in a perfect ensemble system, it may be more profitable to think of the spread estimate as providing some approximate bound on control error. The extent to which this is true should be dependent on ensemble size.

However, there is no theoretical reason why the L^2 measure is preferred over others. In particular it is possible to choose a distance measure to emphasise this notion of spread providing a bound on error. Consider the distance between two forecasts, or between a forecast and its verifying analysis, in terms of the supremum of the absolute value of the difference field, taken over all grid points in a given region. This supremum norm is an upper bound on all norms of the form L^n, n element N. In the discussion below, we assess the impact of ensemble size on spread skill relations using both the L^2 and supremum norms.

From a practical point of view, it is not particularly obvious which norm (L^2 or supremum) is the more useful. Certainly over a large enough area, the supremum norm has limited practical appeal, since the value of the norm is determined by the value of one forecast at one grid point. On the other hand, over sufficiently small areas, it is possible that the supremum norm is of more practical value.

Consider two fields $f(t) = f(x_g, t)$ and $h(t) = h(x_g, t)$, defined for each grid point value x_g inside a region Σ. Define the L^n norm

$$\|f(t)\|_{n,g} \equiv \left[\sum_{x_g \in \Sigma} w(x_g) f^n(x_g) / \sum_{x_g \in \Sigma} w(x_g) \right]^{1/n} \tag{1}$$

where $w(x_g) \equiv \cos(\phi_{x_g})$, ϕ_{x_g}, being the latitude of x_g. The distance between two fields f and h can be computed simply as the L^n norm of their difference,

$$d_n(f, h;t) \equiv \|f(t) - h(t)\|_{n,g} \tag{2}$$

As $n \to \infty$, the L^n norm tends to the supremum norm, with associated distance

$$d_\infty(f, h;t) \equiv max_{x_g \in \Sigma} |f(t) - h(t)| \tag{3}$$

The case $n=2$ corresponds to the root-mean-square (rms) difference, while $n \to \infty$ gives the maximum absolute value of the difference, inside the region Σ.

Let us now consider an ensemble of N+1 forecasts $f_j, j = 0, N$, where $j = 0$ denotes the control forecast. The spread of an ensemble of forecasts (relative to the control) is defined as

$$sp_n(t) \equiv \left[\frac{1}{N} \sum_{J=1}^{J=N} d_n^n(f_j, f_0;t) \right]^{1/n} \tag{4a}$$

$$sp_\infty(t) \equiv \max_{j=1,N} [d_\infty(f_j, f_0;t)] \tag{4b}$$

The skill of a forecast f_j is computed in terms of the L^n norm of the distance between the forecast and the analysis (or the ensemble member used as verification for the idealized ensemble).

In (4) the control forecast can be replaced by the ensemble mean

$$\bar{f} \equiv \frac{1}{N+1} \sum_{J=0}^{J=N} f_j \tag{5}$$

Spread/skill relationships are characterized by 2 indices, where the smaller the index, the more skilful the ensemble. The first index N_{ldia} gives the number of cases where the ensemble spread did not bound the control error. (The subscript *ldia* stands for lower- diagonal points, referring to a spread/skill scatter diagram - such as in Fig. 3 or 4). The second index N_{sl} is the number of cases with ensemble spread lower than average and control error larger than average. (The subscript *sl* stands for *s*mall ensemble spread and *l*arger control error, referring to the lower-right entry of the contingency table associated with a spread/skill scatter diagram.)

3.2 Best ensemble member and percentage of outliers

One important quality of a skilful ensemble prediction system, is that the verifying analysis should lie within the span of the ensemble. We verify this in three different ways.

The first two diagnostics measure the smallest distance between the verifying analysis and an ensemble member.

The first measure is based on the ensemble forecast with the smallest average L^2 distance from the analysis (i.e. rms error), taken over a prescribed area, such as the NH or Europe. This forecast will be referred to as the best ensemble member.

The second measure is based on finding the smallest error within the ensemble, independently at each point over a prescribed area such as the NH or Europe, and then taking the rms of these smallest error values over the prescribed area. In some sense, this latter measure may be more relevant to a forecaster who is only interested in the local point-wise skill of the EPS. We refer to this skill as the point-wise rms error.

The third measure is called Outlier Statistic. Consider the grid point values predicted by the $N+1$ ensemble members $f_j(x_g, t)$, ranked so that $-\infty \leq f_1 \leq . \leq f_{N+1} \leq \infty$. They define N+2 intervals. The probability of the analysis (or more generally of the verifying analysis) lying outside the EPS forecast range is defined as the sum of the probabilities of the analysis being in the two extreme categories, i.e. with $f_j < f_1$ or $f_j > f_{N+1}$. This will be called the Outlier Statistic.

3.3 Relative Operating Characteristic

Following *Stanski et al* (1989), consider the occurrence or non-occurrence of one event such as:'500hPa geopotential height anomaly larger than 50 m', and, for each grid point, check whether the event occurred and if it was predicted or not.

Considering all grid points x_g inside a region Σ, we can construct a two category contingency table, Table 1, where X can be referred to as the hits and the Z as the false alarms. Define the hit rate (i.e. the percentage of correct forecasts) as $X/(X+Y)$ and the false alarm rate (i.e. the percentage of forecasts of the event given that the event did not occur) as $Z/(Z+W)$. If these two rates are plotted against one each other on a graph, a single point results.

Signal detection theory (*Mason*, 1982) is a generalization of these ideas to probabilistic forecasts. Suppose that a forecast distribution generated using an ensemble system is stratified according to observation into ten 10% wide categories, as shown in Table 2, where $j=1,..\ 10$, and with the last category $j=10$ including also $P=100\%$. In other words, considering for example the event '500 hPa geopotential height anomaly larger than 50 m', this table has been constructed by computing, for each grid point, the predicted probability P for the event to occur, then by identifying the category j to which P belongs, and finally by adding 1 to either b_j or a_j, depending on whether the event did or did not occur.

For any given probability threshold $P=j*10\%$, the entries of this table can be summed to produce the four entries of a two by two contingency table:

$$W = \sum_{i=1}^{j} a_i \tag{6a}$$

$$Y = \sum_{i=1}^{j} b_i \tag{6b}$$

$$Z = \sum_{i=j+1}^{K} a_i \tag{6c}$$

$$X = \sum_{i=j+1}^{K} b_i \tag{6d}$$

Then, using the four entries *X, Y, Z, W*, the hit and false alarm rates can be calculated, and a point plotted on a graph. If this process is repeated for all probability thresholds from 0% to 100%, the result is a smooth curve called the relative operating characteristic (ROC).

A convenient measure associated with a ROC curve is the area under the curve, which decreases from 1 toward 0 as more false alarm rates occur. A value of 0.5 is considered as the lower bound for a useful forecast, since a system with such a *ROC-area* cannot discriminate between occurrence and non-occurrence of the event.

A second important measure is given by the separation of the conditional distributions of forecast probabilities given the occurrence and the non-occurrence of the event, which can be constructed using the a_j and b_j values. One measure of this separation is the distance between the means of the two distributions, normalized by the standard deviation of the distribution for non-occurrences, which we will name ROC-distance. Examples of ROC curves and conditional distributions are shown in Figs. 9 and 10.

Throughout this paper, *ROC-areas* and *ROC-distances* have been calculated considering different geographical regions, and 45-day average values have been computed simply by averaging *ROC-areas* and *ROC-distances*.

One of the main advantages of signal detection theory is that it can be used to compare deterministic and probabilistic forecasts (*Stanski et al*, 1989). Thus, a comparison can be made between the skill of an ensemble system and the skill of the control, the ensemble mean forecast or the ECMWF high-resolution T213L31 operational model.

3.4 Brier Score

One of the most common measures of accuracy for verifying two-category probability forecasts is the Brier score (*Brier*, 1950). Again, consider the event '500 hPa geopotential height anomaly larger than 50 m', and concentrate on the forecast probability distribution of Table 2, generated by an ensemble system considering all grid points $\mathbf{x}_g \in \Sigma$.

Following *Hsu and Murphy* (1986), the Brier score can be computed as the sum of three terms:

$$BS = \bar{o}(1-\bar{o}) + \sum_{j=1}^{N} \frac{N_j}{N}[(p_j - \bar{o}_j)^2 - (\bar{o}_j - \bar{o})^2] \tag{7}$$

where Pj=0.05+(j-1)*0.1 is the (centre) forecast probability of the j-*th* probability class, $\bar{o}_j = b_j/(a_j + b_j)$ is the relative frequency of the event, $N_j/N = a_j/A$ where $A = \Sigma_j b_j$ is the relative population of forecasts in the j-*th* class, $\bar{o} = b_j/(a_j + b_j)$ where $B = \Sigma_j b_j$ is the sample climatology.

A natural reference for the Brier score is the Brier score of the sample climatology

$$BS_{cli} = \bar{o}(1-\bar{o}) \tag{8}$$

The Brier Skill Score can be defined as

$$BSS = \frac{BS_{cli} - BS}{BS_{cli}} \tag{9}$$

45-day average Brier scores have been computed averaging *BS* of different cases. To prevent undesired weighting due to variations in the denominator, average Brier skill scores have been computed inserting average Brier scores in Eq. (9), and not by averaging skill scores.

3.5 Ranked probability score

The ranked probability score (*Epstein*, 1969b, *Stanski et al*, 1989) is intended for verifying contiguous multi-category probability forecasts. A perfect categorical forecast always receives a score of 1 and the worst possible categorical forecast receives a score of 0. It is sensitive to the error, in the sense that more credit is given to a forecast which concentrates its probability about the event that occurs.

Following *Stanski et al* (1989), for a given forecast time t and for each grid point x_g inside an area Σ, consider K mutually exclusive classes, and denote by $P=(P_1, .., P_K)$ the predicted ranked probability vector, and by $d=(d_1,.., d_K)$ the observation vector such that d_n=1 if class n occurs and zero otherwise. For example, consider 500 hPa geopotential height, concentrate on ten classes, eight of them 50 m wide from -200 m to 200 m, plus the two outer classes, and focus attention on the prediction of 500 hPa geopotential height anomalies with respect to climatology. Then, for each grid point, P_j is the predicted probability that the anomaly will be in the j-*th* class, and if the observed 500 hPa geopotential height is in the k-*th* class, d_k=1 and d_j=0 for any $j \neq k$.

The ranked probability score *RPS* is defined as:

$$RPS \equiv 1 - \frac{1}{K-1}\left[\sum_{i=1}^{K}\left(\sum_{n=1}^{i} P_n - \sum_{n=1}^{i} d_n\right)^2\right] \tag{10}$$

For each grid point, given the ranked probability scores for an ensemble, *RPS*, and for a standard forecast, RPS_{std}, the Ranked Probability Skill Score can be defined as:

$$RPSS \equiv \frac{RPS - RPS_{std}}{1 - RPS_{std}} \tag{11}$$

The *RPSS* ranges from *1* for a perfect forecast, to -∞, and negative *RPSS* indicate that the ensemble forecast is less accurate that the standard.

Throughout this paper, regional-average ranked probability scores have been computed averaging RPS values of all grid points $x_g \in \Sigma$. Moreover, 45-day average ranked probability scores have been computed averaging the RPS for the 45 different cases. By contrast, to prevent undesired weighting due to variations in the denominator, average ranked probability skill scores have been computed inserting average ranked probability scores in Eq. (11), and not by averaging skill scores.

4. IMPACT OF ENSEMBLE SIZE ON ENSEMBLE SKILL

In this section we analyze the impact of ensemble size on the different measures defined in the previous section. Attention is focused on the prediction of the 500 hPa geopotential height field, over the Northern Hemisphere (NH) and Europe (defined as the area with longitude between 20°W and 45°E, and latitude between 30°N and 75°N). Skill and potential skill for each ensemble configuration are compared.

4.1 Ensemble mean skill

Figure 1a shows the difference between the 45-day average rms error of the ensemble mean and the 45-day average skill of the control forecast for the extra-tropical NH. Results show that an ensemble size increase improves the average skill of the ensemble mean after forecast day 5, especially when increasing the ensemble size up to 8 perturbed members. Similar conclusions can be drawn considering other geographical regions (not shown).

Figure 1b shows the impact of ensemble size on the potential skill of the ensemble mean. It confirms that the impact is slightly stronger when going from 2 to 8 members, than when further enlarging the ensemble size. This effect was predicted by *Leith* (1974).

It can be seen that the skill of the ensemble mean forecast, between day 2 and day 4, is positive in the idealized ensemble, but slightly negative when verified against the actual analysis. This slightly negative impact has proven to be fairly typical when the T63 model was used for the operational EPS. Since December 1996, the operational EPS is run with a T_l159L31 resolution model, starting from initial conditions constructed using smaller initial perturbations. Operational results for the winter 1996/97 show that the improvement of the ensemble mean is more similar to the idealized ensemble results in Fig. 1b and less like the actual skill in Fig. 1a. This seems to indicate that in the old system, based on a T63L19 model which is known to be too little active compared to the atmosphere, too large initial perturbations were used to be able to get a reasonable amount of spread in the medium-range (i.e. after forecast day 3). Because of this choice, most of the perturbed members were characterized by a forecast error larger that the error of the control forecast, and this caused the error of the ensemble mean in the early forecast range to be, on average, larger than the error of the control forecast.

4.2 Relationship between spread and skill

Figure 2 shows spread/skill scatter diagrams for ens_2, ens_8, and ens_32 for day 5 over Europe using the L^2 norm (scatter diagrams for ens_4 and ens_16 are not shown for reason of space). Figure 3 shows a similar set of scatter diagrams using the L^∞ norm. Visually, the impact of ensemble size on the relationship between spread and skill, as suggested by Fig. 2, is not dramatic. By contrast, the impact of ensemble size in Fig. 3 is immediately apparent - as ensemble size is increased so the more the spread bounds the control error.

Figure 4 shows spread/skill scatter diagrams for ens_32 for day 5 over the NH for L^2, L^8. and L^∞. Again, this figure shows that a more convincing relationship can be found using the latter distance measure, with spread bounding error.

These results are confirmed by Table 3, which lists the indices N_{ldia} and N_{sl} of spread/skill relationship, both for the NH region and for Europe. For both these regions, it can be seen that with the L^∞ norm, N_{ldia} is a strongly decreasing function of ensemble size, both for the actual ensemble, and for the idealised ensemble (results in parentheses). By contrast, N_{ldia} does not depend strongly on ensemble size with the L^2 norm. Somewhat intermediate results are obtained using the L^8 norm, as might be expected.

It can be asked whether $n=8$ gives a satisfactory approximation to the supremum norm. Figure 5 shows the estimate of day 5 spread over the NH and Europe as a function of n. In fact, the spread does not approximate to its asymptotic value until $n=O(10^2)$.

Similar conclusions can be drawn considering other regions, and other forecast times.

4.3 Skill of the best ensemble member

The daily values of the rms error at forecast day 5 of the control and of the best member for each ensemble configuration are shown in Figs. 6a-b for the NH and Europe, with a 5-day running average applied to the error values. Figures 7a-d are similar to Figs. 6a-d but show the point-wise best member for each ensemble configuration averaged over the NH and Europe.

The results in Fig 6a-b show some interesting low-frequency variability. For example, over the whole NH, there are some periods where a 32-member ensemble has a notably lower best-member rms error. However, there are other periods when there is no obvious benefit of ens_32 over ens_16. For the European area, on the other hand, the periods are generally more extensive where the best member of ens_32 is notably more skilful than that of ens_16.

The fact that the impact of 32 members over 16 members is greater over Europe than over the whole NH, is probably a reflection of the relative geographical size of the two verification regions. This can be seen in the limit where the best member is evaluated independently at each point within either the NH or Europe (Fig. 7). Here it can be seen that the impact of increasing the ensemble size is more or less uniformly beneficial for any ensemble size, and for all cases considered. For example, the value of the best member for ens_32 is close to 1/3 of the value for ens_2.

4.4 Outlier Statistic

Table 4 shows the impact of ensemble size on the Outlier Statistic, specifically on the 45-day average percentage of outliers at different forecast times.

Results indicate that any increase of ensemble size has a noticeable effect in reducing the percentage of outliers. For verification against the real analyses, the decrease in the percentage of outliers with increasing ensemble size, arises in the early part of the forecast because the perturbations cover more of the hemisphere. As mentioned above, if the amplitude of perturbations was 1 standard deviation of the analysis error, the probability of the analysis lying outside the ensemble in the direction of a perturbation is about 32%. The fact that the day-3 outlier value of ens_2 is much greater than 32% is an indication of the poor NH coverage associated with a single perturbation.

Conversely, in the limit of no predictability, when all ensemble members are equally likely, then the probability of the analysis lying outside the ensemble is 200/(N+2)% (see Appendix). For N=32 this gives 6%. It can be seen that the actual values for ens_32 lie between these extremes. The difference between the actual value at day 10 (16%) and the "no predictability" reference value of 6% can be attributed partly to the fact that there is still some predictability at day 10, and partly to model error.

For the idealised ensemble, it is easy to show that the percentage of outliers should scale as [(32-N)/32]*[200/(N+2)]% for all forecast ranges (the first term denotes the probability that one of the ensemble members is not precisely equal to the verification - for the real ensemble, this probability is equal to unity). As shown in the table, these are the values that are computed from the idealised ensembles, modulo small sampling error.

4.5 Relative Operating Characteristic

Consider three thresholds *Thr=25, 50, 100 m.* Relative operating characteristic curves have been computed for the prediction of the event '500 hPa geopotential height anomaly larger than *Thr*' and for the event '500 hPa geopotential height anomaly smaller than -*Thr*', for the NH and Europe at different forecast times.

As an example, Fig. 8a shows the 45-day average relative operating characteristic for the event '500 hPa geopotential height anomaly larger than 50 m', for the NH at forecast day 5, for ens_32. Apart from the relative operating characteristic curve, three markers are plotted to show the false/hit rates for the control (square), the ensemble mean (cross) and the ECMWF high resolution T213L31 model (diamond). In this case, the ensemble system has a false/hit rate similar to the control and the T213L31 forecasts, while the ensemble mean has a slightly better rate. The *ROC-area* for Fig. 8a is 0.87. By contrast, the ensemble system has a false/hit rate similar to the T213L31 but better than the control and the ensemble mean at forecast day 7 (Fig. 8b). Note that, at forecast day 7, the *ROC-area* decreases from the value of forecast day 5 to 0.79, indicating a forecast deterioration. Similar considerations can be drawn considering the event '500 hPa geopotential height anomaly smaller than -50 m' (not shown).

Figure 8c shows the conditional distributions of forecast probabilities given the occurrence or non-occurrence of the event '500 hPa geopotential height anomaly larger than 50 m', i.e. associated with the ROC curve shown in Fig. 8a. The distance between the means of the two distributions, normalized by the standard deviation of the distribution of non-occurrence, is 1.31. Similarly, Fig. 8d shows the conditional distribution at forecast day 7. At this forecast day, the normalized distance is 0.80. The comparison of Figs. 8c and 8d confirms that the normalized distance (*ROC-distance*) is a useful indication of the separation of the two distributions.

Let us now compare the different ensemble configurations. Figures 9a shows the 45- day average *ROC-area* for the events '500 hPa geopotential height anomaly larger than 50 m'. Results indicate that the impact of an ensemble size increase up to 8 members is detectable during the whole forecast range, but that any further increase has a minor, although still positive, impact. Similar conclusions could be drawn by considering the event '500 hPa geopotential height anomaly smaller than -50 m' (not shown). By contrast, *ROC-distances* are insensitive to ensemble size (not shown). Considering the idealized ensembles, Fig. 9b confirms that the impact of ensemble size is less evident when at least 8 members are included in the ensemble. Results relative to other geographical regions or any other thresholds are similar (not shown).

4.6 Brier score

Figures 10a-b show the Brier scores and skill scores relative to Europe for the event '500 hPa geopotential height anomaly larger than 50 m'. Both the Brier score and the Brier skill scores improve when enlarging the ensemble size up to 8 members, while a smaller although still positive impact can be detected when further increasing the ensemble size. Figures 10a-b show that, for Europe, more than 1 day of predictability is gained by going from 2 to 8 members, and that a further increase of up to about half a day can be gained by enlarging the ensemble size to 32 members. Similar results are obtained for other thresholds and areas (not shown).

Figures 10c-d are analogous to Figs. 10a-b, but computed using the idealised ensembles. They confirm that the impact of ensemble size is less detectable once at least 8 members have been included in the ensemble.

4.7 Ranked probability score

Ranked probability scores and skill scores for the prediction of 500 hPa geopotential height anomaly with respect to climatology have been computed using 10 classes, with the 8 inner classes 50 m wide, and with 50 classes, with the 48 inner classes 20 m wide. Only ranked probability skill scores will be discussed. Moreover, the discussion will focus on scores relative to 10 classes, since the results are insensitive to the number of classes.

Figure 11a shows the ranked probability skill scores for the NH computed with persistence as standard [see Eq. (11)]. Results indicate that an increase in ensemble size improves the skill, especially when the size is increased from 2 to 4 members.

As a further comparison, the ranked probability skill scores of ens_4, ens_8, ens_16 and ens_32 have been computed using ens_2 as standard (Fig. 11b). Results confirm that the impact is always positive, but that the improvement is less evident once the ensemble has at least 8 members (dot curve in Fig. 11b). This is further highlighted by Fig. 11c, which shows, for each ensemble configuration ens_N, the ranked probability skill score computed with respect to ens_N/2 {e.g. the solid line in Fig. 11c shows *RPSS(ens_32)=[RPS(ens_32)- RPS(ens_16)]/[1-RPS(ens_16)]*}.

Figures 11d-f are analogous to Figs. 11a-c, but computed using the idealized ensembles. The comparison between Figs. 11a and 11d indicates that the potential *RPSS* values are about 0.2 higher that the real values, if persistence is used as standard, but only about 0.05 if *RPSS* are computed using ens_2 as standard. Considering the impact of ensemble size, Figs. 11f confirms the result of Fig. 11c, that the improvement in skill is smaller once at least 8 members have been included in the ensemble.

Similar conclusions can be drawn by considering other geographical regions (not shown).

5. DISCUSSION AND CONCLUSIONS

This work has focused on the impact of ensemble size on the performance of the ECMWF Ensemble Prediction System as it was before December 1996 (at the time of this study). This system used T42L19 singular vectors optimized over 48 hours to generate the perturbed initial conditions. These perturbations were integrated using a T63L19 non-linear model.

As well as studying the ensemble mean, spread-skill relationships, and the skill of the best ensemble member, measures specifically designed to verify probabilistic forecasts (e.g. ranked probability skill score, relative operating characteristic curves, Brier skill scores) have been used.

Attention has been focused on a 45-day period, from 1 October to 15 November 1996, on the 500 hPa geopotential height field, and on two areas, the extra-tropical Northern Hemisphere (NH) and Europe. Ensembles characterized by 2, 4, 8, 16 and 32 perturbed members have been compared. The skill of each ensemble configuration, measured using the operational analysis as verification, has been compared to its potential skill, measured using an ensemble member, randomly chosen among the 32 perturbed ensemble members, as verification.

The relation between ensemble spread and control error has been studied using scatter diagrams of ensemble spread versus control error, with distances measured using L^2, L^8 and L^∞ norms (Figs. 2-5). Two skill indices, the number of cases for which ensemble spread did not constitute a bound on the control error (N_{ldia}), and the number of cases with below average spread and above average control error (N_{sl}), have been used to investigate the spread/skill relation. Considering the analysis as verification, results indicate that, for the supremum L^∞ norm, any increase in ensemble size is strongly beneficial, since it decreases both indices N_{ldia} and N_{sl} (Table 3). Results obtained for the idealized ensembles show an even larger positive impact of ensemble size. This larger positive impact is not realized when verifying against an analysis due to the existence of (unaccounted for) model errors. It is worth to mention that work is in progress at ECMWF to modify the EPS so that model errors are taken into account.

Table 5 gives a schematic summary of the results reported in Section 4. A plus, or a minus, under each item is an indication of a positive, or negative, impact of ensemble size. The overall outcome is that although the impact is positive on all measures, the extent to which an increase in ensemble size improves the EPS performance depend strongly on the measure itself

Beside the two spread/skill indices, the most positive sensitive measures to ensemble size have proved to be the skill of the best ensemble member (particularly when evaluated on a point-wise independent basis), the Outlier Statistic (i.e. the percentage of analysis values lying outside the ensemble forecast range), and the ranked probability skill score. Considering spread/skill relation, results suggest that ensemble spread bounds the error of the control forecast if the supremum norm L^{∞} is used to measure distances.

Considering the skill of the best ensemble member, any doubling of the ensemble size reduces the average rms error of the best ensemble member at forecast day 5, e.g. on average by about 8% over Europe (Fig. 6). However, the impact of ensemble size has been proven to be even more dramatic on the rms error of the point-wise best ensemble member. In fact, e.g. at forecast day 5 over Europe, any doubling of ensemble size reduces, on average, the rms error by about 20% (Fig. 7).

Another measure which showed a positive impact of increasing the ensemble size is the Outlier Statistic. For example, at forecast day 5 for the NH, the percentage of outlier decreases from 66% to 41% when increasing the ensemble size from 2 to 4, and it further reduces to 26% when enlarging the ensemble size up to 32 perturbed members (Table 4).

Sensitivity to ensemble size was also detected using the Ranked Probability Skill Score. Considering, for example, the forecast time when the ranked probability skill score relative to the prediction of 500 hPa geopotential height anomalies with respect to climatology crosses the 0.5 line as a limit of useful prediction, results indicate that predictability increases by about 1 day when going from 2 to 8 members, and by about 12 hours when further going to 32 members (Fig. 11).

Positive, though smaller, impact was also detected using signal detection theory, and Brier score and skill scores. Considering, for example, the ROC curves relative to the probability prediction of the event '500 hPa geopotential height anomaly larger (smaller) then 50 (-50) m' for the NH, results indicate that the are under the ROC curve increases when enlarging the ensemble size from 2 to 8 members, while it shows only a small although still positive increase when further adding more members (Fig. 9).

Similar conclusions can be drawn by considering Brier skill scores. Considering, for example, the event '500 hPa geopotential height anomaly larger (smaller) then 50 (-50) m', the forecast time when the Brier skill score reaches the limit of skilful prediction (i.e. when it reaches zero) increases by more than 1 day when going from 2 to 8 members, but it does not change by further increasing the ensemble size (Fig. 10).

At last, considering the ensemble mean, results have shown that its skill is sensitive to ensemble size only when up to 8 members are included in the ensemble. This confirms the conclusions of *Leith* (1974), that once 8 members have been included enough 'intelligent filtering' has been applied, and any further ensemble size increase gives only a refinement, in the ensemble mean field.

We conclude with an important caveat. The results shown in this paper cannot be generalised to conclude that the impact of ensemble size is necessarily small for increases in ensemble size above 32. The dynamical system underlying the truncated equations of motion is not believed to be low-dimensional, and hence a complete estimate of the forecast probability-distribution-function is very unlikely to be well sampled by 32 members. Hence we should view the conclusions of this paper to be specific to the types of diagnostic and types of event studied. To take a specific example; suppose we asked for the Brier Score or the Relative Operating Characteristic for a bounded event over a specific domain and for a limited period (for example the probability that the surface temperature lies within 2 degree of freezing point

over the United Kingdom in December). This contrasts with the types of semi-infinite events (anomaly greater than given threshold) defined over a large region (e.g. Europe) over a whole season.

Similarly one might ask about events defined, not in terms of a single grid point, but over several. For example, an event could be defined by the geopotential height being higher than some threshold at one grid-point, and, simultaneously, lower than some threshold at a different grid-point. The Brier score would be based on joint probability statistics for the two grid-points. This type of score would be relevant in assessing the Brier-score skill of, for example, blocking, using the Tibaldi-Molteni blocking index (*Tibaldi and Molteni*, 1990).

For either of these types of events, it is quite possible that reliable probabilities are unobtainable with only 32 members, and that much larger ensemble sizes are needed. We shall be assessing the impact of ensemble size on these more specific types of events in a future study.

ACKNOWLEDGEMENTS

Acknowledgements should go to Andreas Lanzinger for his helpful suggestions given during the development of signal detection theory diagnostic tools. Adrian Simmons and Olivier Talagrand are also acknowledged for their valuable comments on an earlier version of this paper.

APPENDIX

Consider an idealized 2-dimensional probability-distribution-function at forecast day d+2 for an ensemble compounded of one unperturbed (control) and four perturbed forecasts (x_j, j=0.4, and some hypothetical point t where truth lies. Suppose that the EPS perturbations have $\pm f\sigma$ amplitude, that they are orthogonal and symmetric, and suppose that the coordinate axes are centred at the control.

Let t be distant r from the control forecast. The distance from t to one of the perturbed forecasts (say x_1) is

$$d = \sqrt{(f\sigma)^2 + r^2 - 2r(fr)\cos(\theta)} \tag{A.1}$$

which is less than r if $r\cos(\theta) > f\sigma/2$, where θ is the angle between t and x_1.

Repeating for all points x_j, j=1,4 this inequality defines (in the 2-dimensional plane) a set of lines (at $\pm f\sigma/2$) parallel to the axes, bounding the regions containing the truth where n=0, 1, 2 perturbed EPS forecasts are more skilful than the control.

The expected number N_e of perturbed forecasts that are more skilful than the control can be calculated by multiplying, for each region, n (i.e. the number of perturbed forecasts more skilful than the control) by the appropriate probability values for a 2-dimensional normal distribution.

For example, for f=1, from normal distribution tables the probability that truth lies outside the $\pm\sigma/2$ boundary is about 0.6. Hence, the expected number N_e is

$$[(0.4)^2 * 0.0] + [2 * 0.4 * 0.6 * 1] + [(0.6)^2 * 2] = 1.2 \tag{A.2}$$

The fraction of forecasts better than the control is therefore about 1.2/4=0.3, or 30%.

More generally, for an n dimensional space and an ensemble with $2n$ members, let p denote the probability that truth lies outside the $\pm f\sigma/2$ boundary. Then the expected number N_e is given by the binomial distribution expansion

$$N_e = \sum_{r=0}^{n} \left[\begin{matrix} n \\ r \end{matrix} \right] p^r (1-p)^{n-r} r \tag{A.3}$$

For large n (16 for the EPS) we can use the normal distribution approximation to the binomial distribution, so that

$$N_e \approx np \tag{A.4}$$

and the fraction of forecasts better than the control is $p/2$.

With f=1, p=0.6 as above, then again about 30% of forecasts should do better than the control.

REFERENCES

Barker, T W, 1991. The relationship between spread and forecast error in extended-range forecasts. *J. Climate*, **4**, 733-742.

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

Buizza, R, 1994. Sensitivity of optimal unstable structures. *Q. J. R. Meteorol. Soc.*, **120**, 429-451.

Buizza, R, 1997. Potential forecast skill of ensemble prediction, and spread and skill distributions of the ECMWF Ensemble Prediction System. *Mon. Wea. Rev.*, **125**, 1, 99-119.

Buizza, R, and T N Palmer, 1995. The singular vector structure of the atmospheric general circulation. *J. Atmos. Sci.*, **52**, 9, 1434-1456.

Buizza, R, Gelaro, R, Molteni, F, and T N Palmer, 1997. Predictability studies with high resolution singular vectors. *Q. J. R. Meteorol. Soc.*, **123.**, 1007-1033

Courtier, P, C Freyder, J F Geleyn, F Rabier and M Rochas, 1991. The Arpege project at Météo France. Pp. 192-231 in Proceedings of the ECMWF seminar on *Numerical methods in atmospheric models*, ECMWF, Shinfield Park, Reading RG2-9AX, 9-13 September 1991, Vol. 2.

Courtier, P, Andersson, E, Heckley, W, Pailleux, J, Vasiljevic, D, Hollingsworth, A, Rabier, F, and Fisher, M, 1997. The ECMWF implementation of three dimensional variational assimilation (3D-var). Part I: Formulation. *Q. J. R. Meteorol. Soc.*, in press.

Epstein, E S, 1969a. Stochastic dynamic predictions. *Tellus*, **21**, 739-759.

Epstein, E S, 1969b. A scoring system for probability forecasts of ranked categories. *J. of Appl. Meteorol.*, **8**, 985-987.

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.

Gleeson, T A, 1970. Statistical-dynamical predictions. *J. Appl. Meteorol.*, **9**, 333-344.

Hsu, W, and Murphy, A H, 1986. The attributes diagram. A geometrical framework for assessing the quality of probabilistic forecasts. *International Journal of Forecasting*, **2**, 285- 293.

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

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

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

Palmer, T N, F Molteni, R Mureau, R Buizza, P Chapelet and J Tribbia, 1993. Ensemble prediction. ECMWF Seminar Proceedings on *Validation of models over Europe: Vol. 1*, ECMWF, Shinfield Park, Reading RG2 9AX, UK, 285 pp.

Simmons, A J, D M Burridge, M Jarraud, C Girard and W Wergen, 1989. The ECMWF medium-range prediction models development of the numerical formulations and the impact of increased resolution. *Meteorol. Atmos. Phys.*, **40**, 28-60.

Stanski, H R, Wilson, L J, and Burrows, W R, 1989. Survey of common verification methods in meteorology. Research Report n. 89-5, Atmospheric Environment Service, Forecast Research Division, 4905 Dufferin Street, Downsview, Ontario, Canada M3H5T4, pp 114.

Tibaldi, S, and Molteni, F, 1990. On the operational predictability of blocking. *Tellus*, **42A**, 343-365.

Tracton, M S and E Kalnay, 1993. Operational ensemble prediction at the National Meteorological Center: practical aspects. *Weather and Forecasting*, **8**, 379-398.

Zhu, Y, Iyengar, G, Toth, Z, Tracton, S, and Marchok, T, 1996. Objective evaluation of the NCEP global ensemble forecasting system. Proceedings of the *1th Conference on Numerical Weather Prediction*, August 19-23 1996, Norfolk, Virginia, US, pp 632.

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

Tables

	fc=YES	fc=NO	tot obs
obs=YES	X	Y	X+Y
obs=NO	Z	W	Z+W
tot fc	X+Z	Y+W	

Table 1: An example of contingency table

Probability range	obs=NO	obs=YES
$(j\text{-}1)*10\% \leq P < j*10\%$	a_j	b_j

Table 2: Probability forecast distribution

	NH - Z500 - day 5				
	ens_2	ens_4	ens_8	ens_16	ens_32
N_{ldia}-L^2 norm	39 (25)	36 (21)	36 (**18**)	**35** (20)	37 (20)
N_{ldia}-L^8 norm	39 (22)	28 (20)	24 (16)	22 (11)	**19** (**10**)
N_{ldia}-L^∞ norm	21 (17)	17 (8)	9 (4)	5 (1)	**1** (**0**)
N_{sl}-L^2 norm	10 (4)	8 (4)	8 (6)	8 (**4**)	**7** (**4**)
N_{sl}-L^8 norm	7 (5)	**4** (**5**)	**4** (6)	**4** (**5**)	**4** (**5**)
N_{sl}-L^∞ norm	**4** (7)	**4** (3)	7 (6)	5 (6)	6 (**4**)

Table 3a: pread/skill N_{ldia} index (number of cases where the ensemble spread did not bound the control error, i.e. number of points below the diagonal) and Nsl index (number of cases with ensemble spread lower than average and control error larger than average, i.e. number of points in the lower-right quadrant) for the NH. Bold values identify the most skilful results.

	Europe - Z500 - day 5				
	ens_2	ens_4	ens_8	ens_16	ens_32
N_{ldia}-L^2 norm	31 (19)	**26** (20)	27 (19)	**26** (15)	28 (**13**)
N_{ldia}-L^8 norm	28 (20)	23 (16)	20 (13)	15 (11)	**12** (**7**)
N_{ldia}-L^∞ norm	23 (12)	17 (6)	8 (4)	4 (3)	**2** (**0**)
N_{sl}-L^2 norm	9 (10)	9 (8)	10 (7)	10 (**6**)	**8** (**6**)
N_{sl}-L^8 norm	8 (9)	**6** (7)	7 (**5**)	**6** (6)	7 (6)
N_{sl}-L^∞ norm	9 (7)	9 (6)	**5** (**5**)	6 (7)	6 (6)

Table 3b: pread/skill N_{ldia} index (number of cases where the ensemble spread did not bound the control error, i.e. number of points below the diagonal) and N_{sl} index (number of cases with ensemble spread lower than average and control error larger than average, i.e. number of points in the lower-right quadrant) for Europe at forecast day 5. Bold values identify the most skilful results.

	fc-day 3	fc-day 10	fc-day 10	Reference %
ens_2	65 (47)	66 (46)	59 (47)	50 (47)
ens_4	52 (29)	53 (28)	45 (28)	33 (29)
ens_8	41 (14)	41 (13)	32 (14)	20 (15)
ens_16	34 (5)	33 (5)	23 (5)	11 (5)
ens_32	**28 (0)**	**26 (0)**	**16 (0)**	6 (0)

Table 4: Outlier Statistic for the NH at forecast day 3, 5 and 10, verified against the analysis, and using a randomly chosen ensemble member as verification (values in brackets). A reference percentage of outliers is also reported (last column, see text). Bold values identify the most skilful results.

Diagnostic	2 → 8 members	8 → 32 members
rms error of ens mean (Fig. 1)	decreases (+)	neutral
spread/skill relation (Figs. 2-4, Table 3)	improves (+)	improves(+)
rms error of best member (Fig. 6)	decreases (++)	decreases (++)
rms error of point-wise best member (Fig. 7)	decreases (++)	decreases (++)
Outlier Statistic (Table 4)	decreases (++)	decreases (++)
ROC (Figs. 9)	increases (+)	neutral
BS and BSS (Fig. 10)	increases (+)	increases (+)
RPSS (Fig. 11)	increases (++)	increases (+)

Table 5: Summary of the impact of ensemble size on the ensemble skill as measured using the different diagnostics. A +/- indicates an improvement/deterioration, a ++/-- a large improvement/ deterioration

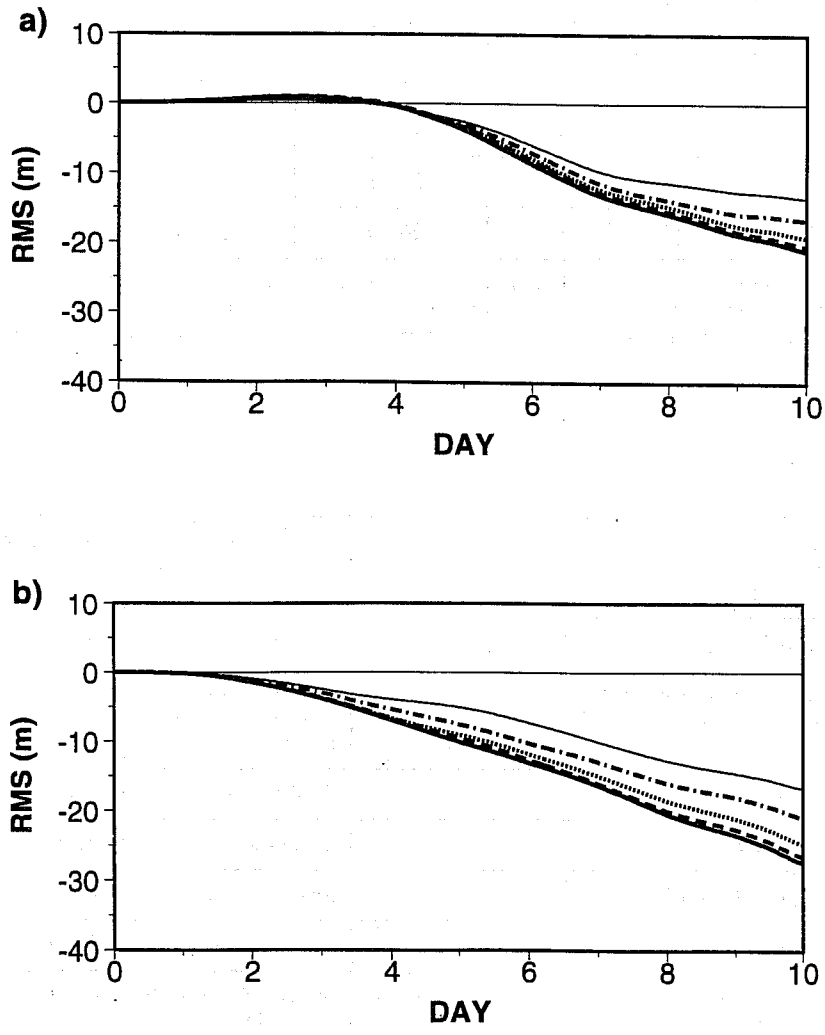

Fig. 1 (a) Difference between the 45-day average rms error of the ensemble mean and the control, for the 500 hPa geopotential height over the NH, for ens_2 (thin solid), ens_4 (chain dash), ens_8 (dot), ens_16 (dash) and ens_32 (solid). (b): as (a) but for the idealized ensembles.

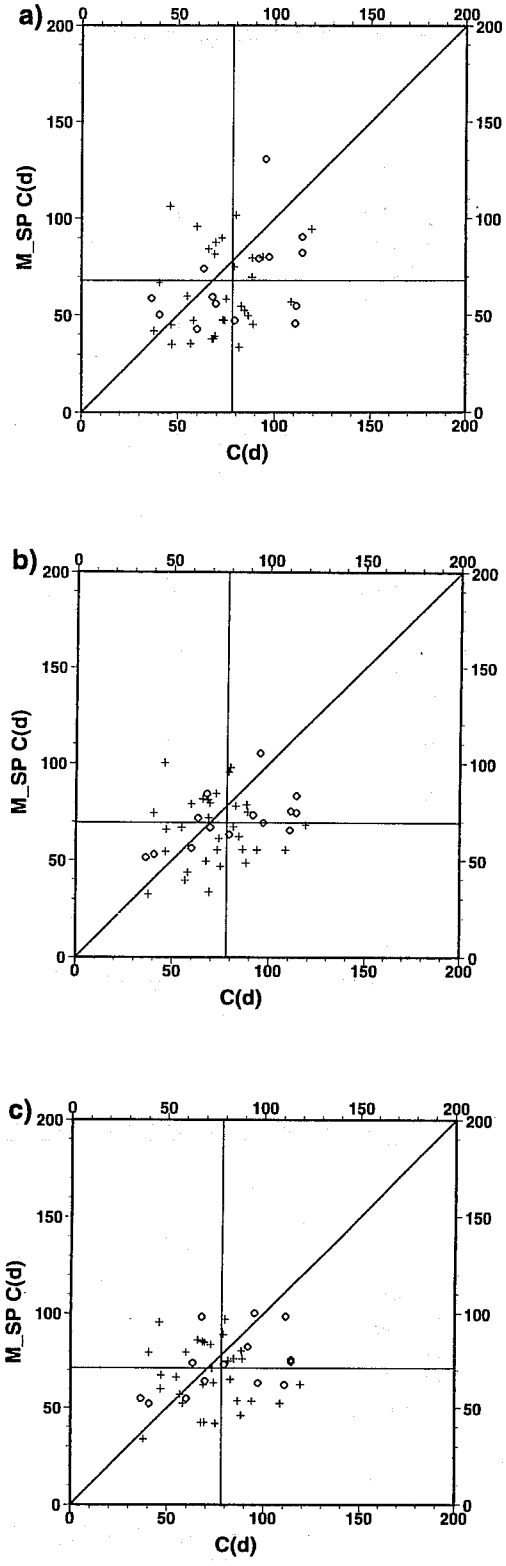

Fig. 2 Scatter diagram of ensemble spread around the control [ordinate, M_SP C(d)] versus control error [abscissa, C(d)], computed using the L^2 norm, over Europe at forecast day 5, for (a) ens_2, (b) ens_8, and (c) ens_32. Crosses identify October cases, diamonds November cases, and constant lines show the average ensemble spread and the average rms error.

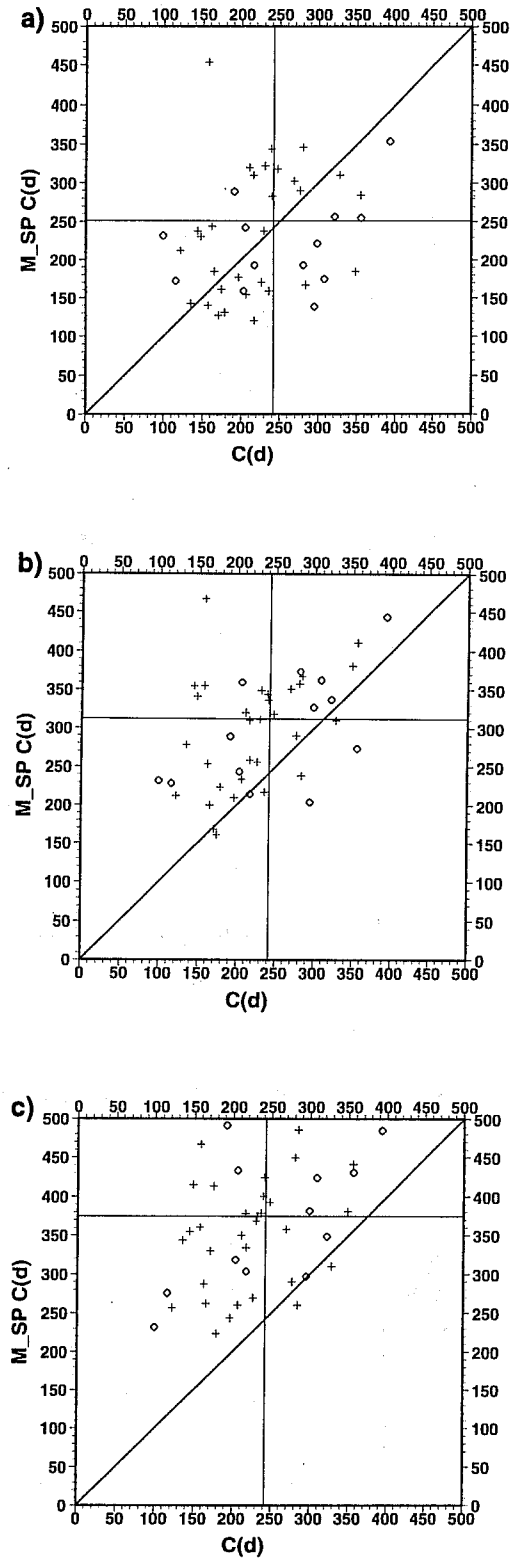

Fig. 3 As Fig. 3 but for distances computed using the L^{∞} norm.

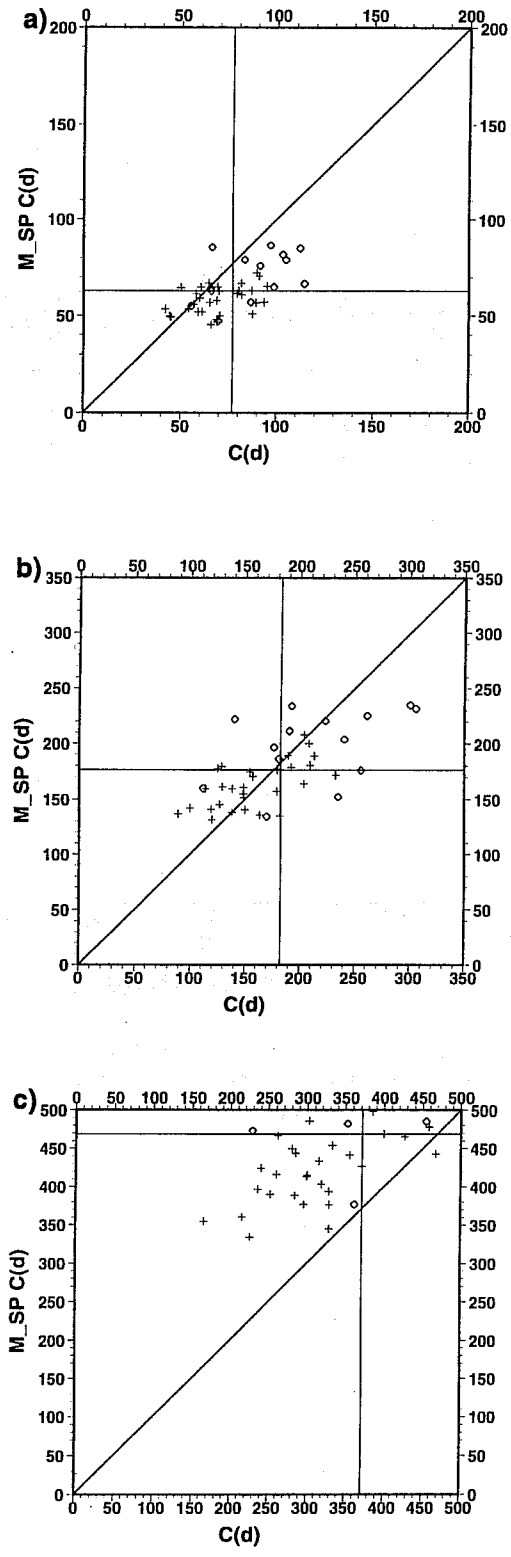

Fig. 4 Scatter diagram of ensemble spread around the control [ordinate, M_SP C(d)] versus control error [abscissa, C(d)], for the NH at forecast day 5, computed using the (a) L^2, (b) L^8, and (c) L^∞ norm.

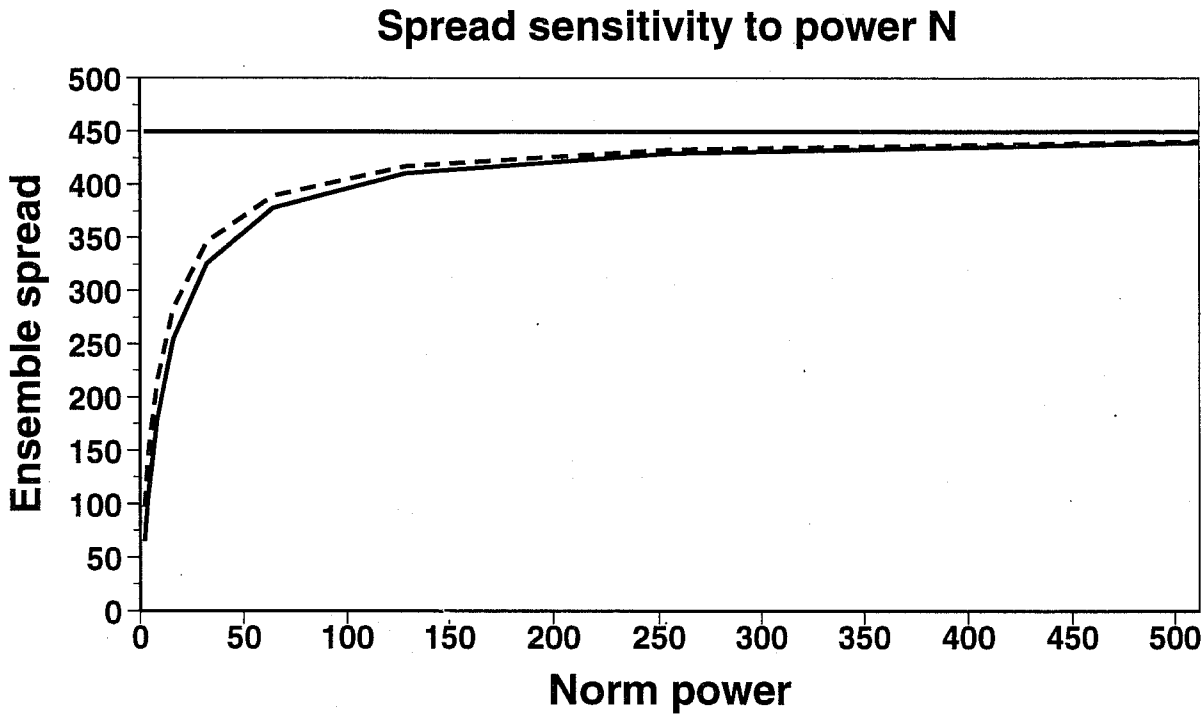

Fig. 5 Ensemble spread at forecast day 5 over the NH (solid) and Europe (dash), as a function of the power n of the norm L^n used to computed distances, for a single (randomly chosen) case study. The solid constant line represents the value of the supremum L^{∞} norm for Europe and the NH, which for this particular case happened to coincide.

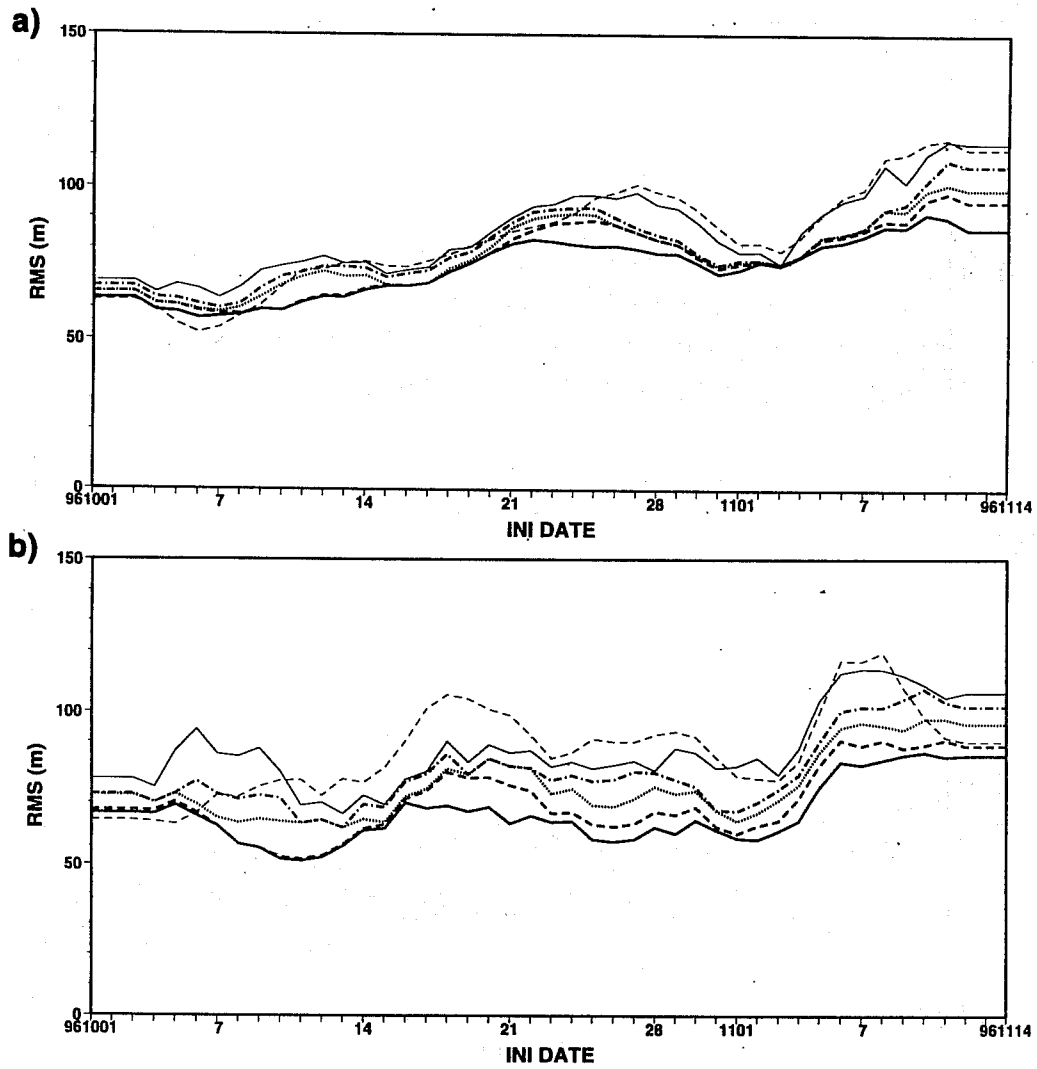

Fig. 6 5-day running mean of the rms error of the control (thin dash), and of the best member of ens_32 (heavy solid), ens_16 (dash), ens_8 (dot), ens_4 (chain dash), and ens_2 (thin solid), at forecast day 5 for (a) the NH and (b) Europe. The best member is the member which has the smallest rms error at forecast day 5. Initial dates (INI DATE) are reported on the abscissa axis.

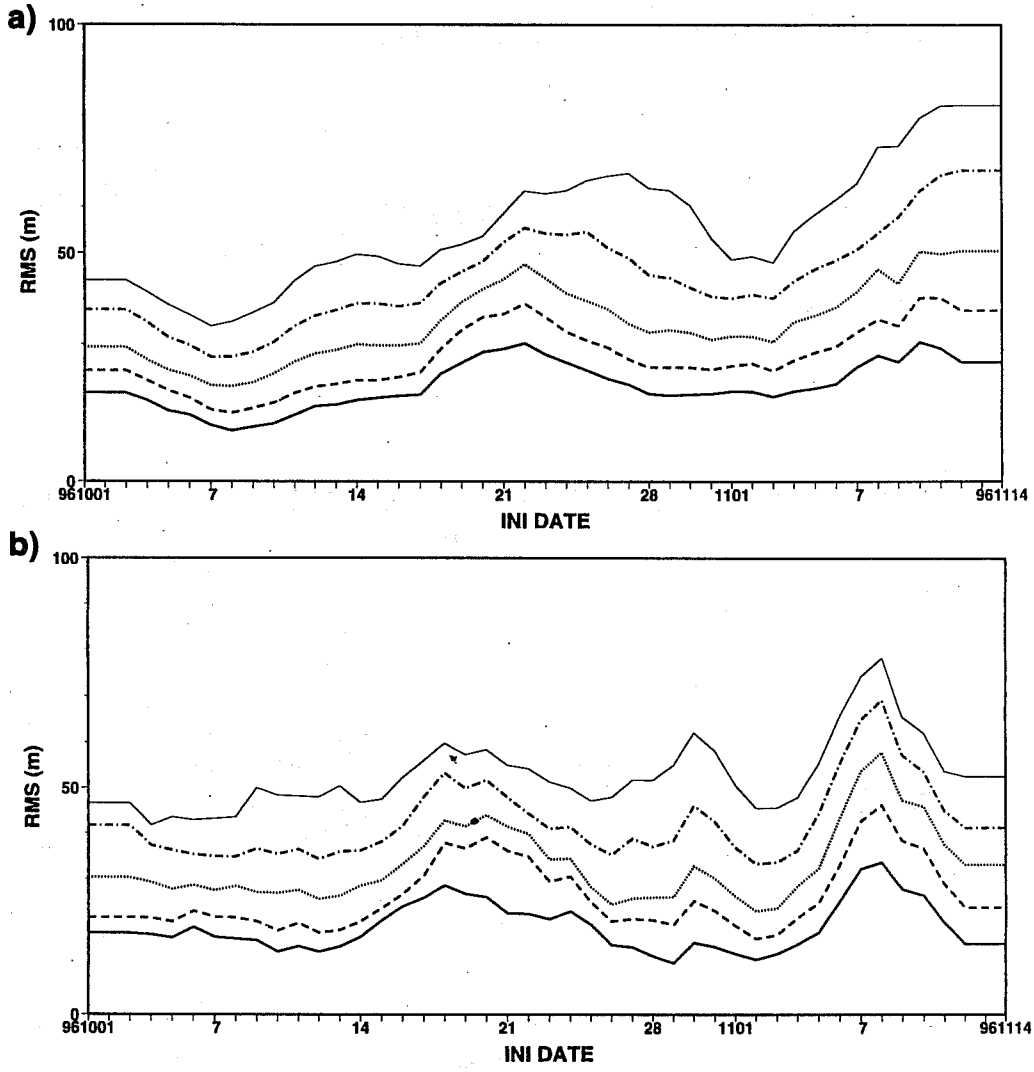

Fig. 7 5-day running mean of the average rms value of point-wise best ensemble member for ens_32 (heavy solid), ens_16 (dash), ens_8 (dot), ens_4 (chain dash), and ens_2 (thin solid), at forecast day 5 for (a) the NH and (b) Europe. Initial dates (INI DATE) are reported on the abscissa axis.

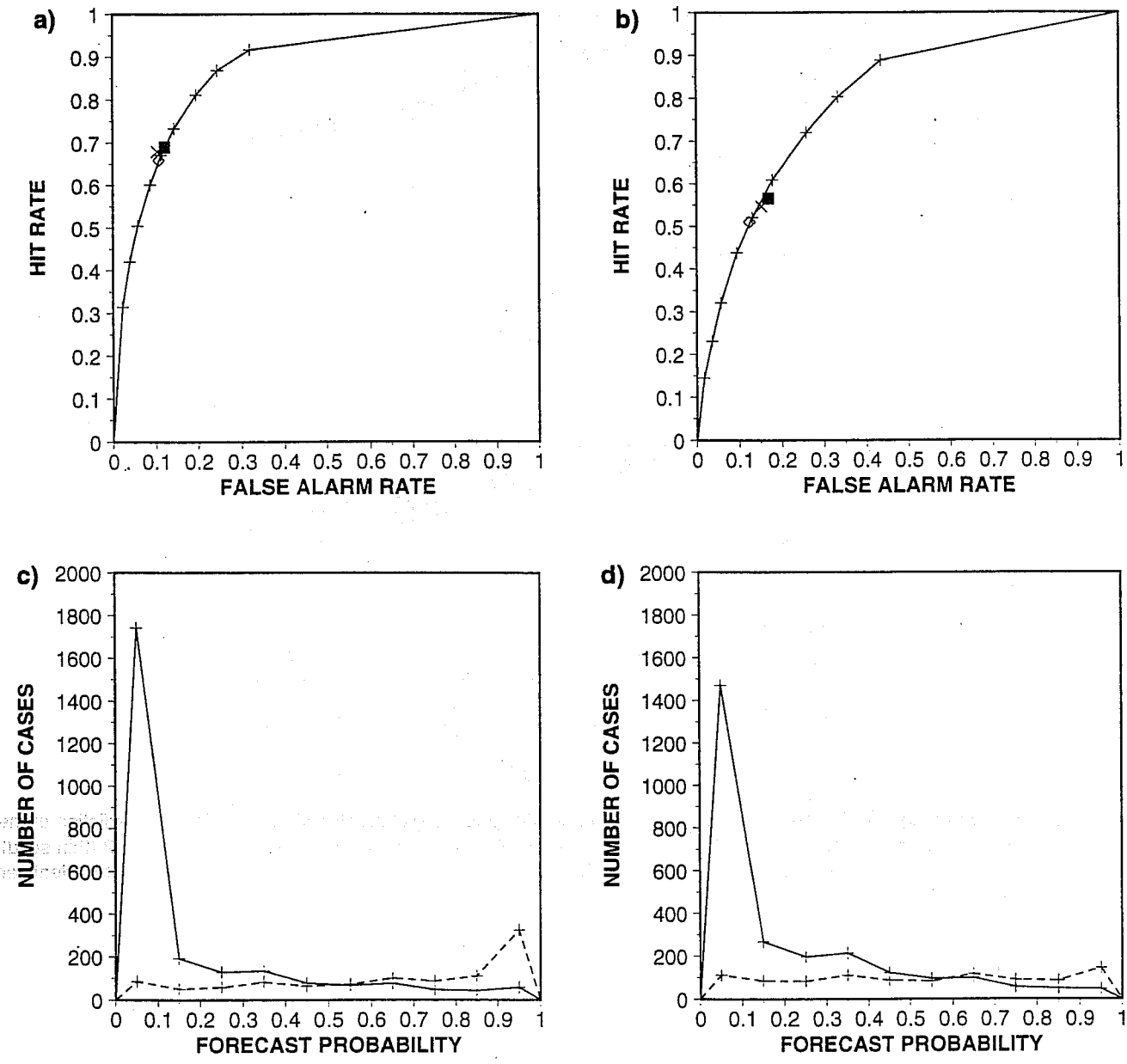

Fig. 8 (a) 45-day average relative operating characteristic for the prediction by configuration ens_32 of the event '500 hPa geopotential height anomaly larger than 50m', at forecast day 5 for the NH. Markers also show the 45-day average false/hit rates of the control (square), the ensemble mean (cross) and the ECMWF high-resolution T213L31 model (diamond). (b) as (a) but at forecast day 7. (c): 45-day average conditional distribution of forecast probabilities given the non-occurrence (solid) and the occurrence (dash) of the event '500 hPa geopotential height anomaly larger than 50m', at forecast day 5 for the NH, relative to the relative operating characteristic curve of panel (a). (d): as (c) but relative to the relative operating characteristic curves of panel (b).

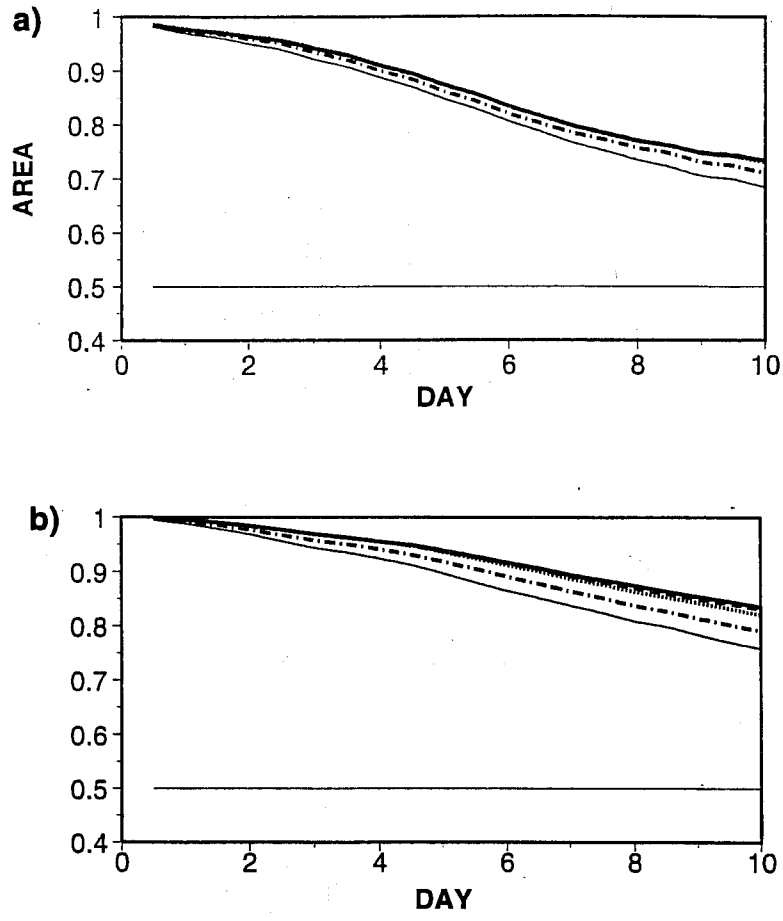

Fig 9. 45-day average (a) ROC-area computed from the relative operating characteristic curves for the prediction of the event '500 hPa geopotential height anomaly larger than 50m', at forecast day 5 for the NH, for ens_2 (thin solid), ens_4 (chain dash), ens_8 (dot), ens_16 (dash) and ens_32 (heavy solid). (b): as (a) but for the idealized ensembles.

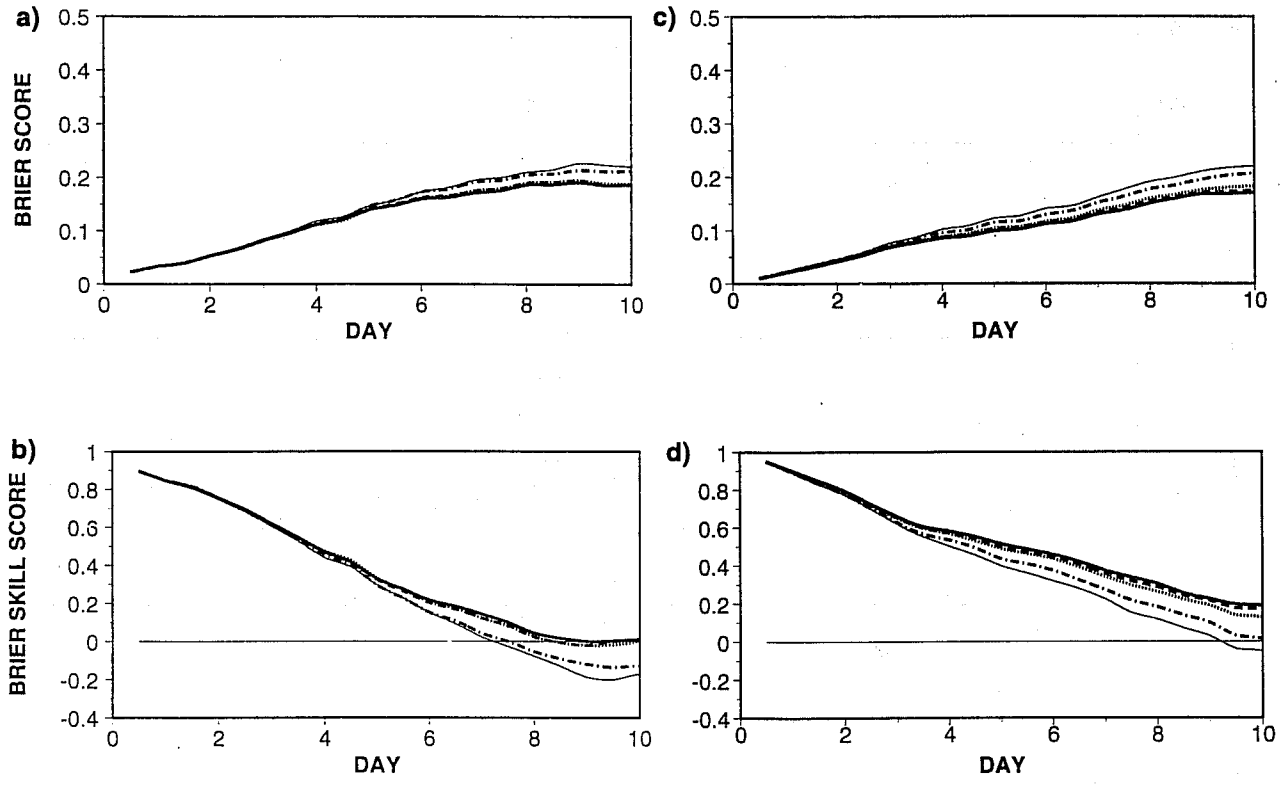

Fig. 10 (a) 45-day average Brier score for the event '500 hPa geopotential height anomaly larger than 50m' over Europe, for ens_2 (thin solid), ens_4 (chain dash), ens_8 (dot), ens_16 (dash) and ens_32 (heavy solid). (b): as (a) but for the Brier skill score. (c-d): as (a-b) but for the idealized ensembles.

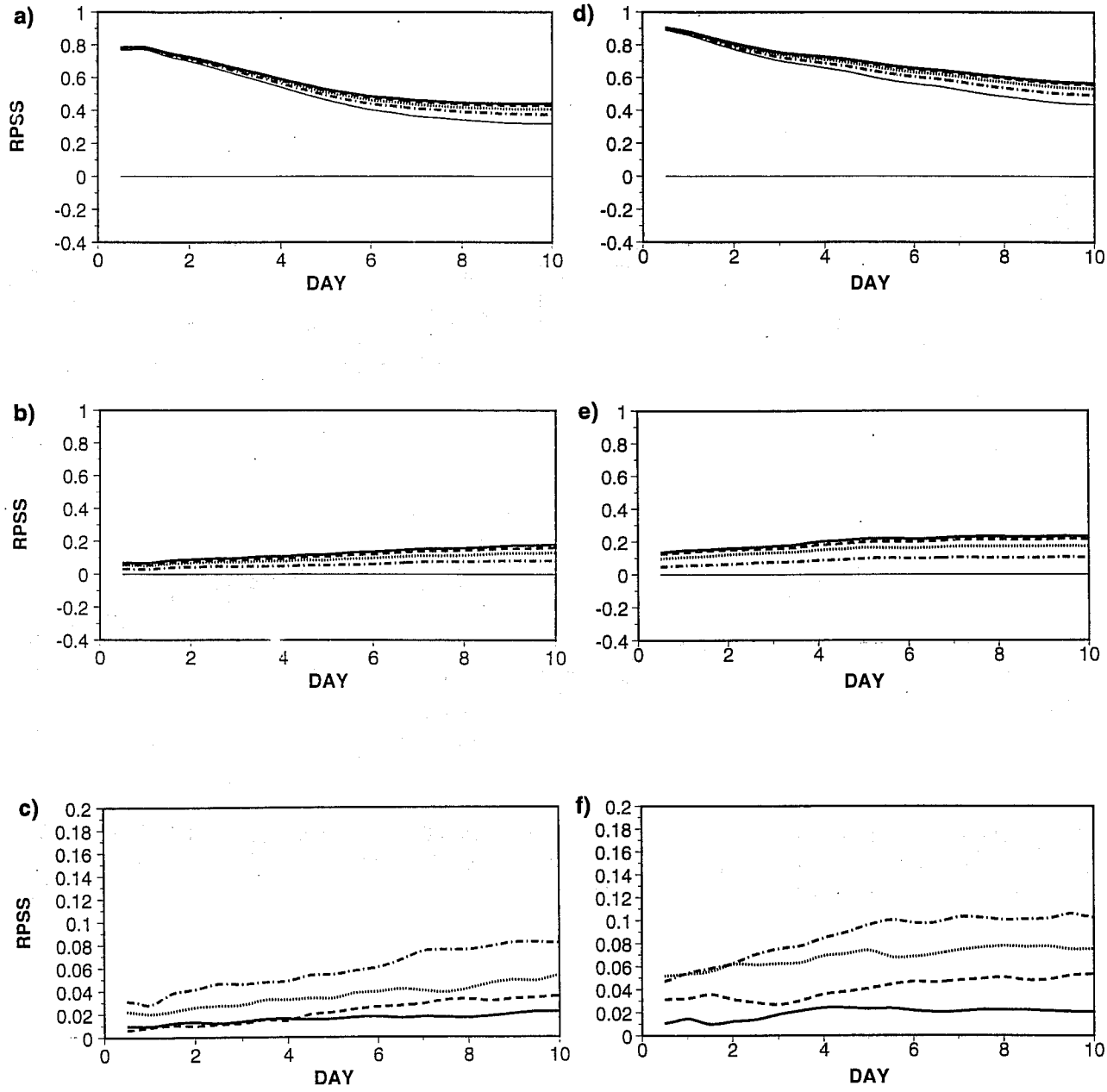

Fig. 11 (a) ranked probability skill score for the prediction of the 500 hPa geopotential height anomaly (with respect to climatology) over the NH, computed for 10 classes, with the 8 inner classes 50 m wide, with persistence as standard [see Eq. (10)], for ens_2 (thin solid), ens_4 (chain dash), ens_8 (dot), ens_16 (dash) and ens_32 (heavy solid). (b): as (a) but with ens_2 as standard. (c): as (a) but for the relative improvement {i.e., when ens_N is considered, for RPSS(ens_N)=[RPS(ens_N)-RPS(ens_N/2)]/[1-RPS(ens_N/2)], see text for more details}. (d-e-f): as (a-b-c), respectively, but for the idealized ensembles.