TECHNICAL MEMORANDUM

245

Impact of model resolution and ensemble size of the performance of an ensemble prediction system

R. Buizza, T. Petroliagis, T. Palmer,
J. Barkmeijer, M. Hamrud,
A. Hollingsworth, A. Simmons and N. Wedi

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 Model Resolution and Ensemble Size on the Performance of an Ensemble Prediction System

R Buizza,T Petroliagis,T Palmer, J Barkmeijer, M Hamrud, A Hollingsworth, A Simmons, N Wedi

European Centre for Medium-Range Weather Forecasts

Also to be published in *Q. J. R. Meteorol. Soc.*

ABSTRACT

Ensemble integrations over fourteen cases are described. These integrations test the relative impact of increase in ensemble size and in the resolution of the model used to integrate the ensemble. The ensembles are evaluated using a variety of statistical tests. Some of these indicate a relative advantage of an increase in ensemble size, whilst most tests suggest a relative advantage of an increase in model resolution. However, overall, the best performance was obtained by combining enhancement in model resolution (from T63L19 to T106L31) with an increase in ensemble size (from 32 to 50 members).

1. INTRODUCTION

The daily operational Ensemble Prediction System (EPS, *Palmer et al,* 1993; *Molteni et al,* 1996) implemented at the European Centre for Medium-Range Weather Forecasts (ECMWF) was, until 10 December 1996, based on 33 non-linear integrations of a low- resolution model version (spectral triangular truncation T63, 19 levels), one (the 'control' forecast) from the 12Z unperturbed analysis, and 32 from the 12Z analysis perturbed along unstable directions (*Buizza and Palmer,* 1995). This system is one of a number of medium- range ensemble systems running in operational or research mode (e.g. *Harrison et al,* 1995; *Houtekamer et al,* 1996; *Tracton and Kalnay,* 1993; *Toth et al,* 1996), and can be thought of as a development of seminal ideas by *Epstein* (1969a), *Gleeson* (1970), *Fleming* (1971a,b) and *Leith* (1974).

In all such cases, the ensembles are a finite sampling, $0(10^1$-$10^2)$, of the full forecast probability distribution function, whose evolution is governed by a Liouville equation (*Ehrendorfer,* 1994). The dimension of NWP models, $0(10^6$-$10^7)$, prohibits the direct solution of such a Liouville equation. The gross disparity between sample size and dimension is clearly problematic, leading to potentially serious sampling error. From a practical point of view, perhaps the most serious sampling error would be associated with an ensemble having small internal dispersion (and therefore forecasting high predictability) but nevertheless failing to encompass the verifying trajectory (and therefore having low skill).

The ECMWF strategy for constructing initial perturbations is based on sampling an unstable linear sub-space (comprising dominant singular vectors from the first 48 hours of the forecast). Whilst this strategy is designed to minimise the type of sampling error described above, it cannot be said that such error has been eliminated from the EPS. For example, taken over the hemisphere as a whole the unstable sub-space itself has dimension larger than 0(10^2- 103) (*Palmer et al,* 1997). Secondly, in the nonlinear range of the forecast, growing perturbations may not be directly associated with the unstable sub-space of the first 48 hours of the forecast.

In view of these remarks, it might seem natural to make use of the enhanced computational resources provided by ECMWF's Fujitsu VPP700 computer system to enhance the ensemble size. However, there is another factor, *a priori* of equal concern against which sampling error must be weighed. Part of the rationale for the EPS is as a tool to support the main T213L31 deterministic model. However, the behaviour of the T213 model and that of the T63 model are not entirely compatible. The most obvious differences are associated with prediction of weather-related phenomena, particularly in the vicinity of orography. For example, it is often found that low-level temperature forecasts from the T213 model lie outside the range of values provided by the EPS, purely because orographic heights for a given location are different. Moreover, the two models are not entirely consistent in terms of their gross statistics (e.g. global transient eddy kinetic energy, *Tibaldi et al,* 1990, or blocking activity; *A. Persson,* personal communication). Hence, in order to provide a practical product that is consistent with the main operational forecast at ECMWF, there is certainly a requirement to bring the resolution of the EPS model closer to the operational model. Of course, irrespective of the resolution of the main deterministic integration, if a goal of the EPS is to provide useful probability forecasts of weather events (e.g. heavy rainfall, strong winds etc), then the model used in the EPS must be capable of simulating such variables. It is clear that a T63L19 model has limited skill in this sense.

The purpose of this paper is to document results from EPS experimentation (14 cases from 1994-95) in which the benefits of increased model resolution were compared with the benefits of enhanced ensemble size. The impact of ensemble size on probabilistic precipitation prediction during three summer cases is also investigated. In addition to this work, there is continuing research on the specification of the initial perturbations, though this will not be discussed in this paper.

The outline of the paper is as follows. The EPS, the sampling strategy used to select the 14 case studies, and the different ensemble configurations tested are described in section 2. Since singular vectors with 31 vertical levels are used in some of the ensemble configurations, the impact of a vertical resolution increase on the singular vectors is analyzed in section 3. Sections 4, 5 and 6 are the main components of this paper in which results from the different ensemble configurations are compared. Conclusions are drawn in Section 7.

2. PRELIMINARIES

This Section is divided in three. Firstly, the EPS is briefly described (a more complete description is given in *Molteni et al,* 1996). Secondly, the selection criterion used to choose the 14 cases is discussed. Finally, a description of the EPS configurations tested in this paper is given.

2.1 The ECMWF EPS

The original operational ECMWF EPS comprised 32 perturbed non-linear integrations of an Eulerian version of the ECMWF model (*Simmons et al,* 1989; *Courtier et al,* 1991) with spectral triangular truncation T63 and 19 levels (T63L19), together with a control integration at the same resolution.

The initial conditions of the EPS perturbed members are created by adding and subtracting perturbations to the control initial conditions. The initial perturbations are defined using the 48-hour singular vectors (*Buizza and Palmer,* 1995) of an approximation 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 the 48-hour optimisation time interval. The inner product used to make these calculations is based on total energy, which appears roughly consistent in some measures with analysis error statistics (*Palmer et al*, 1997).

For the T63L19 version of the EPS, two sets of singular vectors were computed at T42L19 resolution 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 the tangent integrations. The first set was confined to grow in the Northern Hemisphere extra-tropics; the second set to grow in the Southern Hemisphere extra-tropics.

For each hemisphere, about 40 singular vectors were computed daily applying a Lanczos algorithm (*Strang*, 1986), with the exact number depending on the atmospheric flow. From these, 16 were selected. The selection criteria were such that, for each hemisphere, the first 4 singular vectors were always chosen, and each subsequent singular vector (from the 5th onwards) was selected only if half of its total energy lied outside the regions where the singular vectors already selected are localized.

After selection, the 16 singular vectors of each hemisphere were rotated in phase-space and finally re-scaled to construct the 16 ensemble perturbations. The phase-space rotation was applied to generate perturbations which have the same globally-averaged energy as the singular vectors, but smaller local maxima and a more uniform spatial distribution. The rotation was defined to minimize the local ratio between the perturbation amplitude and the amplitude of the analysis error estimate given by the ECMWF data assimilation system. The re-scaling allowed perturbations to have local maxima up to $\alpha = \sqrt{1.5}$ larger than the local maxima of the analysis error estimate. This procedure was applied in each hemisphere. Then the first Northern Hemisphere perturbation was added to the first Southern Hemisphere perturbation, and so on up to the 16th.

The 16 global perturbations were added and subtracted to the control initial conditions to define 32 perturbed initial conditions. Finally, 32+1 (control) 10-day T63L19 non-linear integrations were performed.

Both model changes and revisions of the methodology used to generated the initial perturbations alter the EPS. The major model changes since the EPS started (92.12.19) occurred on 93.08.04, when the new ECMWF surface and boundary layer scheme was introduced (*Viterbo and Beljaars*, 1995), and on 95.04.04, when the new ECMWF prognostic cloud scheme (*Tiedtke*, 1993, *Jacob*, 1994) and a new scheme for the representation of the sub-grid scale orography (*Lott and Miller*, 1995) were implemented. As a result of this study, both resolution and ensemble size were increased on 10 December 1996. Table 1 lists the major changes to the methodology used to generate the perturbed initial conditions.

2.2 Selection of cases

The 14 cases studied in this paper were selected by considering firstly the skill of the 500 hPa geopotential height predicted by the control integration of the T63L19 EPS at forecast day 5 over Europe, and secondly the operational EPS spread. Figure 1a shows the skill of the control (solid) and of the ensemble-mean (dashed), and the ensemble spread (dot), while Fig. 1b is a scatter diagram of ensemble spread/control skill. Attention has been paid to sample cases with above average, average, or poor control skill, and with either good or poor spread/skill relation. The (anomaly correlation based) control skill and ensemble spread for the 500 hPa geopotential height over Europe at forecast day 5, averaged among the 14 cases, are 0.75 and 0.88, while averaged over winter 1994-95 are 0.83 and 0.88, suggesting that the sampling is slightly biased towards cases with poor control skill. The three cases 94.11.09, 94.12.31 and 95.01.12 are the most problematic for the EPS, since smaller-than- average spread is not associated with higher-than-average control skill. They correspond, in fact, to the three crosses in the top/left corner of the scatter diagram in Fig. 1b. These cases may correspond to situations described in the Introduction where sampling error leads to a significant overestimate of forecast predictability.

2.3 The experimental EPS configurations

The experimental EPS configurations discussed in this paper are outlined in Table 2. The first row corresponds to the operational 32-member T63L19 EPS configuration. The second row describes a configuration where the model resolution was kept fixed, but the ensemble size increased to 128(+1), based on 64 (orthogonal) singular vector directions. The third row describes a configuration where the operational EPS perturbations were used, but integrated with a T106L31 model. The fourth row describes a configuration similar to the third, but with singular vectors computed using a 31-level tangent model. Finally, the fifth row describes a configuration with both increased resolution (T106L31) and enhanced ensemble size (50 members).

All integrations from the five configurations have been run with the same model cycle (cycle 12r1.5, which was the operational cycle before the implementation of the new ECMWF prognostic cloud scheme and the new scheme for the representation of the sub-grid scale orography, see sub-section 2.1), and with the same perturbation initial amplitude ($\alpha = \sqrt{1.5}$, as it was in the operational EPS when the experimentation started). Note that T42L19 singular vectors were used to generate the perturbed initial conditions in configurations 32*T63, 128*T63 and 32*T106, while T42L31 singular vectors were used in configurations 32*T106SV31 and 50*T106SV31.

3. IMPACT OF RESOLUTION ON SINGULAR VECTORS

Buizza (1997a) has discussed the impact of horizontal resolution on singular vector characteristics. It was concluded that the increase in singular values when scales with total wave-number $n>42$ are introduced is smaller than the increase obtained when going from T21 to T42 (*Hartmann et al*, 1995). Specifically, an increase of horizontal resolution from T42 to T63 would give an extra 25% growth, for horizontal diffusion damping times on the smallest scale set to 6 and 12 hours, respectively. Considering their structure and geographical location, very small differences can be detected between T42 and T63 singular vectors. Moreover, T42 and T63 singular vectors have very similar total energy spectra, suggesting that the contribution of scales with total wave-number $n>42$ is rather small. The inclusion of scales with total wave-numbers $n>42$ could be more important when more physical processes are included in the linear forward and adjoint models. In fact, preliminary results indicate that the inclusion of moist processes induces a shift of the singular vectors' energy spectra towards large wave-numbers (*Errico and Ehrendorfer*, 1995; *Buizza et al*, 1996). Considering that, in terms of computer time, it costs approximately 5 times more to compute T63 instead of T42 singular vectors, it was decided not to increase the horizontal resolution until more physics is included in the forward and adjoint tangent models.

Concerning the impact of vertical resolution, the comparison of T42L19 and T42L31 singular vectors computed for the 14 case studies indicates that they are very similar, both in terms of growth rates, total energy spectra and vertical structure (not shown). Clearly, when non-linearly integrated with a T106L31 model version, T42L31 singular vectors grow in a more consistent way, as will be shown later when comparing ensembles run in configurations 32*T106 and 32*T106SV31. Since the extra cost of an increase of vertical resolution is small compared to the one needed for higher horizontal resolution, we decided to test the use of T42L31 singular vectors in the EPS.

4. IMPACT OF ENSEMBLE SIZE/RESOLUTION: STATISTICAL TESTS

As discussed by *Buizza* (1997b), a skilful ensemble prediction system is expected to fulfil at least the following three requirements:

i) the ensemble spread should be comparable with the error of the control forecast,

ii) small spread should indicate a skilful control forecast, and

iii) the verifying analysis should lie within the range covered by the ensemble forecasts.

Focusing on these requirements a number of probabilistic diagnostics are examined in this section, based on the 14 case studies. These include statistics of ensemble spread and control error, scatter diagrams of ensemble spread versus control error, ranked probability skill scores, relative operating characteristics, frequency of verifying analysis values lying outside the ensemble forecast range, cluster analysis and reliability analysis.

4.1 Ensemble spread and skill

Figure 2 shows, as a function of forecast time, the root-mean-square (rms) error of the controls, the rms spread (relative to the control) of the ensembles, and the rms error of the ensemble mean scores. The left hand panels refer to the 500 hPa geopotential height, and the right hand panels to the 1000 hPa geopotential height. Figures 2a,d show that, on average, the skill of the T106L31 controls are somewhat higher up to about day 5 than the corresponding T63L19 controls. The impact of enhanced resolution is more apparent at 1000 hpa. After day 5 there is no consistent difference between the controls.

The benefit of increasing resolution is more apparent, however, if one compares the ensemble spread. Figures 2b,e show that the two T106 ensemble configurations produce notably more spread than the T63 configurations. The effect here is somewhat more apparent at 500 hPa. Since ensemble spread is somewhat smaller than the control error, the most consistent results will be obtained with a configuration with small control error and large ensemble spread. This would argue for the importance of running the EPS with T106 resolution. A more complete comparison of the PDF of the ensemble spread and of the PDF of the control error confirms that a better agreement between the two distributions is achieved by the two high resolution configurations (not shown).

In terms of ensemble-mean skill, there is little difference between the configurations, though 50*T106SV31 has the lowest errors. This result is not especially surprising, given the relatively small impact of enhanced resolution on control skill. For example, it is well known that the impact of ensemble size on the ensemble mean skill is relatively weak for ensemble sizes exceeding about 10 (*Leith*, 1974).

The impact of resolution on skill can be seen more clearly by comparing the percentage of ensemble forecasts with anomaly correlation higher than 0.9 (Fig. 3a) or 0.8 (Fig. 3b). The benefit of higher resolution is less apparent if lower thresholds are considered, e.g. 0.6 (Fig. 3c).

4.2 Correspondence between small ensemble spread and high control skill

As discussed in sub-section 2.2, three cases (94.11.09, 94.12.31 and 95.01.12) are the most problematic, since smaller-than-average spread does not correspond with higher-than- average control skill (Fig. 1). For ease of comparison, contingency tables for small/large spread, small/large error, can be constructed from scatter diagrams as in Fig. 1, with the categories defined by the average values.

Table 3 gives the contingency tables for all configurations, relative to the 500 hPa geopotential height over Europe, at forecast day 5, and based on anomaly correlations. (In particular, the 32*T63 table refers to the scatter diagram of Fig.1). Table 3 shows that the best agreement between ensemble spread and control skill is achieved in configuration 50*T106SV31. Similar results are obtained if rms spread/control errors are used instead of anomaly correlations (see Table 3, values in brackets).

Although the statistical significance of these results can be questioned, they are reported because they give a more complete picture of the comparison. It should be stressed that the number of case studies has been limited by the availability of resources (the comparison reported in this work used about 3500 hours of a dedicated 128 processor CRAY- T3D machine).

4.3 Cluster analysis

For completeness of comparison, a cluster analysis has been performed on the 500 hPa geopotential fields over Europe. The cluster procedure applied is the same as the one used operationally at ECMWF (*Molteni et al*, 1996). Table 4 shows, for forecast day 5, the average number of clusters, the average anomaly correlation skill and the average population of the most populated and the best cluster. The clustering algorithm is set up so that at least 2 and no more than 6 clusters are always produced.

Configuration 50*T106SV31 is characterized by the largest number of clusters (reflecting the fact that this configuration has the largest spread). As a consequence, the cluster populations are relatively small. Considering the skill of the most populated cluster, configuration 32*T63 shows the highest skill. By contrast, the skill of the best cluster is higher for configurations 32*T106SV31 and 50*T106SV31. Considering configuration 32*T106SV31, it should be pointed out that, for three out of 14 cases, the best cluster comprised only 1 element, while this never occurred for any other configuration.

4.4 Ranked probability skill score and relative operating characteristic

The ranked probability score (*Epstein*, 1969b, *Stanski et al*, 1989) is intended for verifying multi-category probability forecasts. It is positively oriented, in the sense that the better forecasts get the higher scores, and it is defined in such a way that a perfect categorical forecast always receives the 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. The ranked probability skill score measures skill with respect to a standard, which in our case is persistence, and ranges from 1 (perfect forecast) to $-\infty$. The reader is referred to Appendix A for more details.

Ranked probability skill scores have been computed for probability predictions of 500 hPa geopotential height anomalies with respect to climatology. Forecasts and observed values have been classified in 10 categories, 8 of them 50 m wide covering a range from -200 m to 200 m, one characterized by values smaller than -200 m and one by values greater than +200 m. Table 5 reports the ranked probability skill scores for some forecast days for the Northern Hemisphere. Results indicate that, for any forecast time, the best skill scores are given by configuration 50*T106SV31. Similar results are obtained for smaller areas such as Europe. Moreover, results have been confirmed by considering 50 categories spaced by a 20 m interval (not shown).

Another verification procedure based on Signal Detection Theory (*Mason*, 1982, *Stanski et al*, 1989) has been applied. More specifically, relative operating characteristic curves (the reader is referred to Appendix A for more details) for probability predictions of 500 hPa geopotential height anomalies smaller than -50 m and - 25 m, and greater than 25 m and 50 m have been computed and compared. Considering a relative operating characteristic curve, two measures of importance can be defined. The first one, which we will call the *ROC-area*, is the area under the relative operating characteristic curve. It ranges from 1 for a perfect forecast (i.e. for a forecast with zero false alarm rates) to 0.5 for a useless forecast system which gives false alarms at the same rate as hits. The second one, which we will call *ROC-distance*, is a measure of the separation between the conditional distributions of forecast probabilities given the occurrence or non-occurrence of the event.

Table 6 reports ROC-areas and ROC-distances for the different ensemble configurations for some forecast days for the Northern Hemisphere, for probability predictions of anomalies smaller than -50 m or greater than 50 m. As for the ranked probability skill scores, these two measures have been proved to change very little with the ensemble configuration. Generally speaking, results show that higher ROC-areas and ROC-distances are given by configuration 50*T106SV31. Similar results have been obtained for the other two thresholds, and by considering other regions (not shown).

Generally speaking, Tables 5 and 6 indicate that both an increase on ensemble resolution and of ensemble size have a positive impact on the ensemble skill.

4.5 Probability of the analysis lying outside the ensemble forecast range

In comparing the probability of the analysis lying outside the ensemble forecast range, one needs to consider that, based on purely random sampling of the analysis error PDF in the limit of zero model error, the expected percentage of cases where the analysis is outside the ensemble is $100*[2/(N_{ens}+1)]$, where N_{ens} is the number of ensemble members.

Table 7 summarizes the percentages of analysis values lying outside the ensemble forecast range in addition to that expected by chance, for all configurations, for the 500 and 1000 hPa geopotential height at different forecast ranges over the Northern Hemisphere. Results show that a resolution increase alone (e.g. compare 32*T63 with 32*T106) has a small impact on this diagnostic. On the other hand, increasing ensemble size (e.g. compare 32*T63 with 128*T63) has a larger impact. Table 7 suggests that the percentage of wrongly predicted cases over this expected value $\{100*[p-2/(N_{ens}+1)]$, where p is the percentage given in Table 7} is smallest for configuration 50*T106SV31. Similar results are obtained for Europe.

It should be noted that there are two important reasons why one would not expect the percentage of outlier to equal the value $100*[2/(N_{ens}+1)]$. The first reason is model error. For example, the fact that the percentage of outlier is not substantially less for configuration 128*T63 than for 50*T106SV31 is an indication that T63L19 model error is greater than T106L31 model error. The second important reason is that the initial perturbations are not random samples of the analysis error PDF; rather, they are fixed-amplitude perturbations pointing along specific singular vector directions. The reduction in the percentage of outlier as the ensemble size increases depends crucially on the extent to which the new directions point into the sub-space spanned by analysis error. This in turn depends on the extent to which the energy metric describes accurately the analysis error covariance matrix (*Palmer et al*, 1997).

Note that it would be a trivial matter to reduce the percentage of outlier: one would merely increase the amplitude of the initial perturbations. However, in so doing, the resulting perturbations would be inconsistent with the analysis error PDF, in the sense that such initial perturbations would be very unlikely to occur.

4.6 Brier skill scores for probability predictions of temperature and precipitation

Brier skill scores have been computed for probability predictions of temperature anomalies (temperature 4 and 8 K warm/cold anomalies at 850 hPa) and for probability predictions of precipitation amounts larger than 1 and 10 mm/day (*Brier*, 1950; *Petroliagis et al*, 1997). The Brier skill score gives a measure of the skill of a probabilistic prediction compared to climatology. It is 1 for a perfect forecast, 0 for a probabilistic forecast which is no more accurate than a forecast based on climatology, and negative for even worse forecasts.

Table 8, which reports Brier skill score of probability prediction of temperature anomalies at forecast day 5, suggests that for the -8, -4, and +4 K categories, the best results are obtained with the 128*T63 configuration. This is a somewhat surprising result since it might be thought that the impact of the high resolution model on the skill of low-level temperature forecasts would be significant. On the other hand, for the +8 K category, it can be seen that the Brier skill score is higher for all of the T106 configurations than for the 128*T63 configuration.

Table 9, which lists Brier skill scores for probability prediction of precipitation amounts at forecast day 5 and 7, suggests that for the 1 mm/day category, the best results are obtained with the 128*T63 configuration. However, for the higher 10 mm/day category, the best results are obtained with the 50*T106SV31 configuration.

5. IMPACT OF ENSEMBLE SIZE/RESOLUTION: SYNOPTIC EVALUATIONS

Four case studies are analyzed in this Section, to further investigate the impact of increased resolution and ensemble size enlargement.

5.1 The case of 94.12.12: low-level temperature prediction at D+2.

This case is characterized, around forecast day 2, by a flow change from westerly to north-westerly over Scandinavia, and by a consequent cooling over the region. Attention is focused on the prediction of the position of the -8°C isotherm for 850 hPa temperature. Figure 4 shows that the -8°C isotherm predictions of the high-resolution ensembles (32*T106, 32*T106SV31 and 50*T106SV31) are more able to describe the T213L31 deterministic prediction. Moreover, some perturbed members of the high-resolution configurations are relatively close to the verification. The 50*T106SV31 configuration (Fig. 4e) includes perturbed members which follow the T213L31 prediction over Sweden, and correctly restricts the prediction of the area of cooling.

5.2 The case of 94.12.24: error reduction over Europe

This case is characterized, up to forecast day 5 and for all ensemble configurations, by very good predictions both by controls and perturbed members. By contrast, after forecast day 5, both the T63 and the T106 controls predict too fast a propagation of a trough over Europe, and by forecast day 8 both controls wrongly predict a westerly flow over the British Isles and a northerly flow across the Alpine chain and over Italy, while the analyzed field is characterized by a northerly flow over the British Isles and a westerly flow over Italy (see Fig. 5a for the T106 control prediction, Fig. 5c for the analysis, and Fig. 5d for the error of the T106 control). As a consequence, at forecast day 8, very few perturbed members of the 32*T63, 32*T106 or 128*T63 ensembles are characterized by an anomaly correlation skill higher than 0.6 (Fig. 6). By forecast day 9, none of the perturbed members of these three configurations has anomaly correlation skill higher then 0.6.

By contrast, two perturbed members of configuration 32*T106SV31 and four perturbed members of configuration 50*T106SV31 have anomaly correlation skill higher than 0.8 at forecast day 8, and have anomaly correlation skill higher than 0.6 at forecast day 9 (Fig. 6). (Note that these four forecasts are very good during the whole forecast range, and do not present any so-called return-of-skill.) The best forecast of configuration 50*T106SV31 is given by ensemble member number 28 (Fig. 5b), which has a very small error over Europe (Fig. 5e). Its initial conditions were generated by adding the perturbation shown in Fig. 5f to the control analysis. The time evolution of the divergence of the best ensemble member from the control seems to indicate that the perturbation located across the date line is responsible for the forecast improvement.

5.3 The case of 95.01.12: prediction from the best-cluster

This is the only case among the 14 selected when the T63 model is clearly better able to predict the correct atmospheric flow over Europe between forecast days 5 and 7. In fact, considering for example forecast day 7, although both the T63 and the T106 controls and the deterministic T213L31 model fail in forecasting a ridge over Russia (Fig. 7a), many 32*T63 and 128*T63 perturbed members have very high anomaly correlation skill, while just one 32*T106 member and four 50*T106SV31 members have anomaly correlation skill higher than 0.6 (Fig. 8). The poor performance is essentially due to a wrong prediction of a westerly rather than a southerly flow over the Scandinavian countries, due to a failure in predicting of a block development over that region.

Figure 7 shows the 500 hPa geopotential height field associated with the best cluster of four ensemble configurations. At forecast day 7 over Europe, the 32*T63 best cluster (Fig. 7c) is number 4 of a total of 4 clusters, comprises 4 out of 32 members, has an internal spread of 59 m, an rms error of 100 m and an anomaly correlation skill of 0.85. By contrast, the 32*T106 best cluster (number 2 of a total of 2 clusters), comprises 14 out of 32 members (Fig. 7b) has a lower internal spread (46 m), a higher rms error (195 m) and a lower anomaly correlation skill (0.36). The main deficiency of the 32*T106 ensemble prediction is the fact that it has quite a small spread (only two clusters are generated, with quite small internal spread), which should be associated with a quite small rms error of the control forecast. Configuration 32*T106SV31 has similar characteristics.

On the other hand, the 50*T106SV31 ensemble has a larger spread, as can be judged by the number of clusters produced (4). Moreover, its best cluster (number 4, with 6 out of 50 members, see Fig. 7d) has a larger internal spread (51 m), a smaller rms error (163 m) and a higher anomaly correlation skill (0.56) than the best 32*T106 cluster. Hence, in this case, increasing resolution alone was not helpful. However, the (unusual) deleterious effects of increased resolution were, in this case, ameliorated by larger ensemble size.

5.4 The case of 95.01.15: probability prediction for precipitation

This case is characterized by very intense precipitation over Bretagne, north-west of France, between the 18th and the 22nd of January 1995. Attention is focused on 24-hour accumulated precipitation, predicted for forecast day 3, 5 and 7 over Bretagne by ensembles started on 95.01.15.

Figure 9a shows the observed precipitation between January 17 and 18 (the 5 mm/day and the 15 mm/day isoline are drawn), and Figs. 9b-f give the probability of having more than 15 mm/day of rain from all ensemble configurations. The high resolution ensembles give a more correct location for the precipitation maximum, as indicated by the probability maxima. Moreover the probability values associated with this maximum are larger for the higher resolution model (95% for 32*T106, 32*T106SV31 and 50*T106SV31, and 65% for 32*T63 and 75% for 128*T63). Moreover, configuration 50*T106SV31 (Fig. 9f) is the only one with more than 15% probability of rain over the south of France. Similar considerations can be drawn by comparing the probability predictions for forecast day 5 and 7 (not shown).

6. FURTHER COMPARISON OF CONFIGURATIONS 32*T63 AND 50*T106SV31 DURING THREE SUMMER CASES

Six further ensemble experiments have been run for three summer cases, three in configuration 50*T106SV31 and three in configuration 32*T63. These experiments used a 15% larger initial amplitude (i.e. with $\alpha = \sqrt{2}$) and a more update model version (specifically, cycle 15r5). The three summer cases (96.09.21, 96.09.22 and 96.09.24) have been selected to investigate further the impact of ensemble size and model resolution and the prediction of a very intense and rapid cyclone development occurring over Northern Germany on the 29th of September (not shown).

Considering the forecasts started on the 21st of September and verified on the 29th (+8 day forecasts), both the T63L19 and the T106L31 control forecasts were not able to predict correctly the circulation over Europe, and this is reflected in their anomaly correlation skill (see solid curves with full dots in Figs. 10a,b). It is worth noting that also the high resolution T213L31 prediction was particularly poor (not shown). Considering the 32*T63 perturbed forecasts, none of them has anomaly correlation skill higher than 0.5 (Fig. 10a, dotted curves) while 8 of the 50*T106SV31 perturbed forecasts have anomaly correlation skill higher than 0.6, one of them with anomaly correlation skill higher than 0.75 during the whole 10-day forecast range.

The day after (i.e. the 22nd of September), the performance of the 32*T63 control and ensemble members improves dramatically (Fig. 10c). Comparing the two ensemble configurations, 50*T106SV31 is characterized by a better performance (Fig. 10d). This is especially true if one considers the forecast range between day 4 and 6.

Finally, the comparison of the ensemble forecasts started on the 24th further illustrates enhanced capability of configuration 50*T106SV31 to have some perturbed members with anomaly correlation skill higher than 0.6 during the whole 10-day forecast range (Figs. 10e,f). Moreover, results indicate that the skill of the perturbed forecasts is more independent from the skill of the control forecast for configuration 50*T106SV31 than for configuration 32*T63. In other words, the three 50*T106SV31 ensembles (Figs. 10c, d, f) have similar characteristics despite the fact that the skill of the control forecast varies substantially.

As a further comparison, percentages of analysis values lying outside the ensemble forecast range in addition to that expected by chance, averaged for the three summer cases, have been computed for the Northern Hemisphere (Table 10). Results strongly suggest that an increase of ensemble size and resolution would improve the performance of the ECMWF EPS (because of the limited number of cases, results for Europe are not statistically significant and thus are not shown).

7. CONCLUSIONS

The 32-member T63L19 ensemble prediction system is deficient because ensemble members do not span the range of uncertainty of analysis error, and also because the integrating model does not have resolution compatible with the T213L31 high resolution deterministic model. Ensembles for fourteen different cases, chosen to span a range of forecast skill, have been made for a variety of ensemble configurations. These include the original operational 32*T63 EPS configuration, an enhanced member 128*T63 configuration, an enhanced resolution 32*T106SV31 resolution (where the singular vectors are computed with 31 levels), and a combined enhanced member and resolution ensemble 50*T106SV31. For all ensemble configurations the perturbation initial amplitude has been set to the same value (specifically, $\alpha = \sqrt{1.5}$, see Section 2.1).

The results showed benefits of increases in both ensemble size and resolution. The comparison of ranked probability skill scores and relative operating characteristic curves indicates that both changes have a positive impact on ensemble skill. On the one hand, the outlier statistic (showing the percentage of cases where the verification lay outside the ensemble range) was most substantially reduced with increased ensemble size. Similarly for most categories studied, the Brier scores showed most skill for the largest ensembles. On the other hand, spread was increased with higher resolution, and the percentage of ensemble members with very high skill was highest in enhanced resolution forecasts. Three out of four of the presented case studies showed the importance of higher resolution on probability forecasts.

Results have shown that, even in configuration 50*T106SV31, ensemble spread is still underestimated. Part of the problem undoubtedly lies in the effect of model error. However, as new model cycles are introduced, one might expect to see this effect becoming smaller and smaller. On the other hand, the effect of random model error must be accounted for more explicitly, e.g. by introducing stochastic diabatic forcing in the model equations (this strategy is currently under investigation at ECMWF). A second reason for the discrepancy may lie in the choice of amplitude of initial perturbations. In fact, in operations, since March 1996, the amplitude of the initial perturbations for the EPS has been larger by a factor of 33% than that used in these experiments. However, the choice of perturbation amplitude is not well defined, as it requires knowledge of the projection of analysis error onto the unstable sub- space. In parallel with the experimentation described here, research is ongoing to refine further the structure and amplitude of initial perturbations. This includes the direct use of observations to constrain the singular vector structures (through the use of the 3D-VAR Hessian) and the addition of evolved singular vectors (which are virtually orthogonal to the initial singular vectors) to the initial perturbations, as a means of describing the stable sub- space of analysis error.

Overall, results pointed to the fact that the next generation operational EPS should include both enhanced resolution and enhanced membership. The 50*T106SV31 configuration gave the best performance throughout the tests.

On 10 December 1996, both ensemble size and resolution increased in the operational EPS running on ECMWF's new Fujitsu VPP700 computer system. Ensemble size was increased to 50 perturbed forecasts, and vertical resolution was increased to 31 levels for both the singular vector calculations and the main forecasts, as in the study reported here. Developments in numerical technique had enabled introduction of a two-time-level semi- Lagrangian scheme with a 45-minute time step. The spectral horizontal resolution was increased to T159, with the model integrated using the so-called "linear-grid" option whereby the computational grid is the same as the standard "quadratic" grid used for the T106 integrations reported here. Comparison of the operational performance of the EPS in 1997 with that 1996 confirms the benefits of increased resolution and ensemble size identified in this paper.

ACKNOWLEDGEMENT

We thank Zoltan Toth and Eugenia Kalnay for their very useful comments.

APPENDIX A

TWO VERIFICATION METHODS FOR PROBABILITY FORECASTS

This appendix briefly describes two verification methods which are particularly useful for probability forecasts.

Ranked Probability Skill Score

Following *Stanski et al* (1989), for a given forecast time *t* and for each grid point *x*, let us consider *K* mutually exclusive classes, and let us denote by $P=(P_1, .., P_K)$ the ranked probability vector, and by $d=(d_1, .., d_K)$ the observation vector such that dn=1 if class n occurs and zero otherwise. The ranked probability score *RPS (x, t, P, d)* is defined as:

$$RPS(x, t, P, d) = 1 - \frac{1}{K-1}\left[\sum_{i=1}^{K}\left(\sum_{n=1}^{i} p_n - \sum_{n=1}^{i} d_n\right)^2\right] \tag{A.1}$$

Given a ranked probability score *RPS* for an ensemble and for a standard forecast, which in our case is *persistency*, the Ranked Probability Skill Score can be defined as:

$$RPSS(ens) = \frac{RPS(ens) - RPS(persistence)}{1 - RPS(persistence)} \tag{A.2}$$

Throughout this paper, regional-mean ranked probability scores have been computed averaging *RPS* of all grid points *x* belonging to a specific geographical region. Moreover, case-mean ranked probability scores have been computed averaging *RPS* of different cases. To prevent undesired weighting due to variations in the denominator, average ranked probability skill scores have been computed inserting average ranked probability scores in equation (A.2), and not by averaging skill scores.

Signal Detection Theory

Again following *Stanski et al* (1989), let us consider a two category contingency table:

	Forecast=YES	Forecast=NO	Total observed
Observed=YES	X	Y	X+Y
Observed=NO	Z	W	Z+W
Total forecast	X+Z	Y+W	

where X can be referred to as the hits and the Z as the false alarms. Let us define the hit rate (i.e. the percentage of correct forecast) 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 is a generalization of these ideas to probabilistic forecasts. Suppose to have a forecast distribution stratified according to observation into ten 10% wide categories:

Probability range	Observerd=NO	Observed=YES
$(j-1)*10\% \le prob < j*10\%$	a_j	b_j

with *j=1,.. 10*, with the last category *j=10* including also *prob*=100%

For a given probability threshold *prob=j**10%, the entries of this table can be summed to produce the four entries of a two by two contingency table, the hit and false alarm rates calculated, and a point plotted on a graph. Specifically, the four entries of a two by two contingency table for probability threshold *prob=j**10% are:

$$W = \sum_{i=1}^{J} a_i \tag{A.3a}$$

$$Y = \sum_{i=1}^{J} b_i \tag{A.3b}$$

$$Z = \sum_{i=j+1}^{\Lambda} a_i \tag{A.3c}$$

$$X = \sum_{i=i+1}^{\Lambda} b_i \tag{A.3d}$$

If the process is repeated for all probability thresholds from *0%* to *100%*, the result is a smooth curve called the relative operating characteristic (ROC).

One 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 measure of importance 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 (for example, the conditional distribution of forecast probabilities given the occurrence of the event is given by a_j). 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.

Throughout this paper, ROC-areas and ROC-distances have been computed considering different geographical regions. Case-mean values have been computed averaging ROC-areas and ROC-distances of different cases.

The reader is referred to *Stanski et al* (1989) for a more detailed mathematical definition.

REFERENCES

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, 1997a. Linear time evolution of T21, T42 and T63 singular vectors. *J. Atmos. Sci.*, in press.

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

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

Buizza, R, Palmer, T N, Barkmeijer, J, Gelaro, R, and Mahfouf, J F, 1996. Singular vectors, norms and large-scale condensation. American Meteorological Society pre-prints of the *11th conference on Numerical Weather Prediction*, 19-23 August 1996, Norfolk, Virginia, US.

Courtier, P, Freyder, C, Geleyn, J F, Rabier, F, and Rochas, M, 1991. The Arpege project at Météo France. Proceedings of the ECMWF Seminar on *Numerical methods in atmospheric models*, 9-13 September 1991, Vol. 2, ECMWF, Shinfield Park, Reading RG2 9AX, UK, pp 324.

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.

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

Errico, E R, and Ehrendorfer, M, 1995. Moist singular vectors in a primitive-equation regional model. American Meteorological pre-prints of the *Tenth conference on atmospheric and oceanic waves and stability*, 5-9 June 1995, Big Sky, Montana, US, 272 pp.

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.

Harrison, M S J, Richardson, D S, Robertson, K, and Woodcock, A, 1995. *Medium-range ensembles using both the ECMWF T63 and Unified models - An initial report*, Technical Report no. 153, Forecasting Research Division, Meteorological Office, London Road, Bracknell, Berkshire RG12 2SZ, UK.

Hartmann, D L, Buizza, R, and Palmer, T N, 1995. Singular vectors: the effect of spatial scale on linear growth of disturbances. *J. Atmos. Sci.*, **55**, 22, 3885-3894.

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

Jacob, C, 1994. The impact of the new cloud scheme on ECMWF's Integrated Forecasting System (IFS). Proceedings of the ECMWF Workshop on *Modelling, validation and assimilation of clouds*, 31 October-4 November 1994, ECMWF, Shinfield Park, Reading RG2 9AX, UK, pp 464.

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

Lott, F, and Miller, M, 1995. A new sub-grid scale orographic drag parametrization: its formulation and testing. ECMWF Technical Memorandum n. 218, ECMWF, Shinfield Park, Reading RG2 9AX, UK.

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

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

Palmer, T N, Molteni, F, Mureau, R, and Buizza, R, 1993. Ensemble prediction. ECMWF Seminar proceedings *'Validation of models over Europe: Vol. 1'*, ECMWF, Shinfield Park, Reading, RG2-9AX, UK.

Palmer, T.N., Gelaro, R, Barkmeijer, J, and Buizza, R, 1997. Singular vectors, metrics and adaptive observations. *J. Atmos. Sci.*, in press.

Petroliagis, T, Buizza, R, Lanzinger, A, and Palmer, T N, 1997. Potential use of the ECMWF Ensemble Prediction System in cases of extreme weather events. *Meteorol. Appl.*, **4**, 69-84.

Simmons, A J, Burridge, D M, Jarraud, M, Girard M, and Wergen, W, 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 M3H 5T4, pp 114.

Strang, G, 1986. *Introduction to applied mathematics*. Wellesley-Cambridge press, pp 758.

Tibaldi, S, Palmer, T N, Brancovic, C, and Cubasch, U, 1990. Extended-range predictions with ECMWF models: influence of horizontal resolution on systematic error and forecast skill. *Q. J. R. Meteorol. Soc.*, **116**, 835-866.

Tiedtke, M, 1993. Representation of clouds in large-scale models. *Mon. Wea. Rev.*, **121**, 11, 3040-3060.

Toth, Z, Kalnay, E, Tracton, S, Wobus, R, and Irwin, J, 1996. A synoptic evaluation of the NCEP Ensemble. Proceedings of the *Fifth workshop on Meteorological Operational Systems*, 13-17 November 1995, European Centre for Medium-Range Weather Forecasts, Shinfield Park, Reading, RG2 9AX, UK.

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

Viterbo, P, and Beljaars, A C M, 1995. An improved land surface parametrization scheme in the ECMWF model and its validation. *J. Clim.*, **8**, 2716-2748.

92.12.19	EPS starting date. The EPS is run weekly on Saturdays, Sundays and Mondays, with T21L19 global singular vectors, maximized over a 36 hour time interval, with initial amplitude scaled with $\alpha = \sqrt{2}$.
93.02.20	The singular vectors are maximized only over the Northern Hemisphere extra-tropics ($\phi \geq 30°N$)
94.05.01	EPS starts being run daily
94.08.23	The singular vectors' optimisation time interval is increased from 36 to 48 hours, and the perturbation initial amplitude is increased, $\alpha = 2$
95.03.14	The singular vectors' horizontal resolution is increased to T42, and the perturbation initial amplitude is reduced, $\alpha = \sqrt{1.5}$
96.03.04	Global EPS: perturbations growing in the Southern Hemisphere extra-tropics are added to the perturbed initial conditions.
96.03.25	The perturbation initial amplitude is increased, $\alpha = \sqrt{2}$
96.12.10	The singular vectors' vertical resolution is increased to L31, the model resolution is increased to $T_L159L31$, the number of perturbed members is increased to 50.

Table 1: List of major modifications implemented in the construction of the EPS initial conditions

EPS configuration	size	forecast resolution	Singular vectors' resolution
32*T63	32	T63L19	T42L19
128*T63	128	T63L19	T42L19
32*T106	32	T106L31	T42L19
32T106L31	32	T106L31	T42L31
50*T106SV31	50	T106L31	T42L31

Table 2: Characteristics of the EPS configurations tested

Configuration			
32*T63		Small error	Large error
	Small spread	3 (4)	**4 (4)**
	Large spread	3 (3)	4 (3)
	cc = 0.09 (0.11)		
32*T106		Small error	Large error
	Small spread	3 (5)	**4 (4)**
	Large spread	2 (3)	5 (2)
	cc = 0.12 (-0.10)		
32*T106SV31		Small error	Large error
	Small spread	3 (5)	**2 (4)**
	Large spread	2 (3)	7 (2)
	cc = 0.12 (-0.02)		
50*T106SV31		Small error	Large error
	Small spread	3 (5)	**1 (3)**
	Large spread	2 (3)	8 (3)
	cc = 0.22 (-0.05)		
128*T63		Small error	Large error
	Small spread	3 (4)	**5 (3)**
	Large spread	3 (3)	5 (4)
	cc = 0.25 (-0.27)		

Table 3: Contingency tables for small/large spread, small/large error, computed for the 500 hPa geopotential height, over Europe at forecast day 5, using anomaly correlation or rms distances (values in brackets). [In terms of anomaly correlations, small/large denotes an anomaly correlation value smaller/larger than average, while in terms of rms distances, small/large denotes a distance (either rms spread or rms error) smaller/larger than average.] For each configuration, correlation coefficients are also reported

Configuration	$N_{clusters}$	cluster 1		cluster 2	
		$\%_{mem}$	ACC	$\%_{mem}$	ACC
32*T63	4.2	40%	0.69	27%	0.78
32*T106	4.6	40%	0.64	32%	0.79
32*T106SV31	5.4	36%	0.66	18%	0.83
50*T106SV31	5.8	31%	0.64	21%	0.81
128*T63	5.2	32%	0.65	21%	0.77

Table 4: Average results of cluster analysis of ensembles run in all configurations, relative to 500 hPa geopotential height over Europe, at forecast day 5. Column 1: number of cluster; column 2: percentage of ensemble members of the most populated cluster; column 3: anomaly correlation skill of the most populated cluster; columns 4-5: as columns 2-3 but for the best cluster

Configuration	Ranked Probability Skill Score - forecast day			
	3	5	7	10
32*T63	0.670	0.548	0.484	0.454
32*T106	0.688	0.558	0.495	0.462
32*T106SV31	0.685	0.561	0.503	0.465
50*T106SV31	**0.695**	**0.571**	**0.509**	**0.471**
128*T63	0.683	0.554	0.495	0.463

Table 5: Ranked probability skill scores for categorical prediction of 500 hPa geopotential height anomaly (with respect to climatological values), over the Northern Hemisphere at different forecast ranges. For each forecast time, bold identifies the best value

Configuration	ROC-area - forecast day				ROC-distance - forecast day			
	3	5	7	10	3	5	7	10
32*T63	0.935	0.878	0.828	0.727	2.34	1.57	1.03	0.48
32*T106	0.942	0.886	0.833	0.738	**2.37**	1.55	0.98	0.46
32*T106SV31	0.942	0.886	0.838	0.743	2.35	1.452	0.99	0.46
50*T106SV31	**0.946**	**0.892**	**0.842**	**0.747**	2.36	1.55	1.00	0.47
128*T63	0.941	0.882	0.836	0.734	**2.37**	**1.59**	**1.05**	**0.49**

Table 6: Signal detection theory: (a) ROC-area and ROC-distance for probability prediction of 500 hPa geopotential height anomaly (with respect to climatological values) smaller than -50 m, over the Northern Hemisphere at different forecast ranges; (b) as (a) but for anomaly larger than 50 m. For each forecast time, bold identifies the best value.

Configuration	Z500 - forecast day				Z1000 - forecast day			
	3	5	7	10	3	5	7	10
32*T63	29	24	19	13	28	24	18	11
32*T106	26	22	14	8	26	21	14	6
32*T106SV31	26	20	12	6	26	20	12	6
50*T106SV31	20	16	11	5	20	15	11	4
128*T63	19	18	14	8	20	17	14	7

Table 7: Percentages of analysis values lying outside the ensemble forecast range minus the expected value [i.e. p-2/(Nens+1), where p is the percentage of analysis values lying outside the ensemble forecast range, see Section 4.5], relative to the 500 and 1000 hPa geopotential height over the Northern Hemisphere at different forecast ranges, for all configurations

Configuration	-8°	-4°	+4°	+8°
32*T63	.026	.259	.440	.385
32*T106	.016	.252	.429	.411
32*T106SV31	.023	.252	.428	**.417**
50*T106SV31	.041	.273	.439	**.417**
128*T63	**.060**	**.285**	**.449**	.405

Table 8: Brier skill score (parenthesis) for probability prediction of temperature anomalies of -8, -4, +4 and +8 degrees, relative to temperature at 850 hPa, over the Northern Hemisphere, at forecast day 5

Configuration	forecast day 5		forecast day 7	
	1mm/day	10mm/day	1mm/day	10mm/day
32*T63	.286	.066	.201	.009
32*T106	.286	.095	.219	.078
32*T106SV31	.285	.097	.219	.078
50*T106SV31	.298	**.104**	.230	**.091**
128*T63	**.299**	.087	**.238**	.049

Table 9: Brier skill score for probability prediction of precipitation amounts of 1 and 10 mm/day, over the Northern Hemisphere, at forecast day 5 and 7

Configuration	NH - forecast day			
	3	5	7	10
32*T63	63	53	36	17
50*T106SV31	20	16	11	3

Table 10: Percentages of analysis values lying outside the ensemble forecast range minus the expected value [i.e. p-2/(Nens+1), where p is the percentages of analysis values lying outside the ensemble forecast range, see Section 4.5], relative to the 500 hPa geopotential height over the Northern Hemisphere at different forecast ranges, for configurations 32*T63 and 50*T106SV31, averaged during the three summer cases (96.09.21, 96.09.22 and 96.09.24)

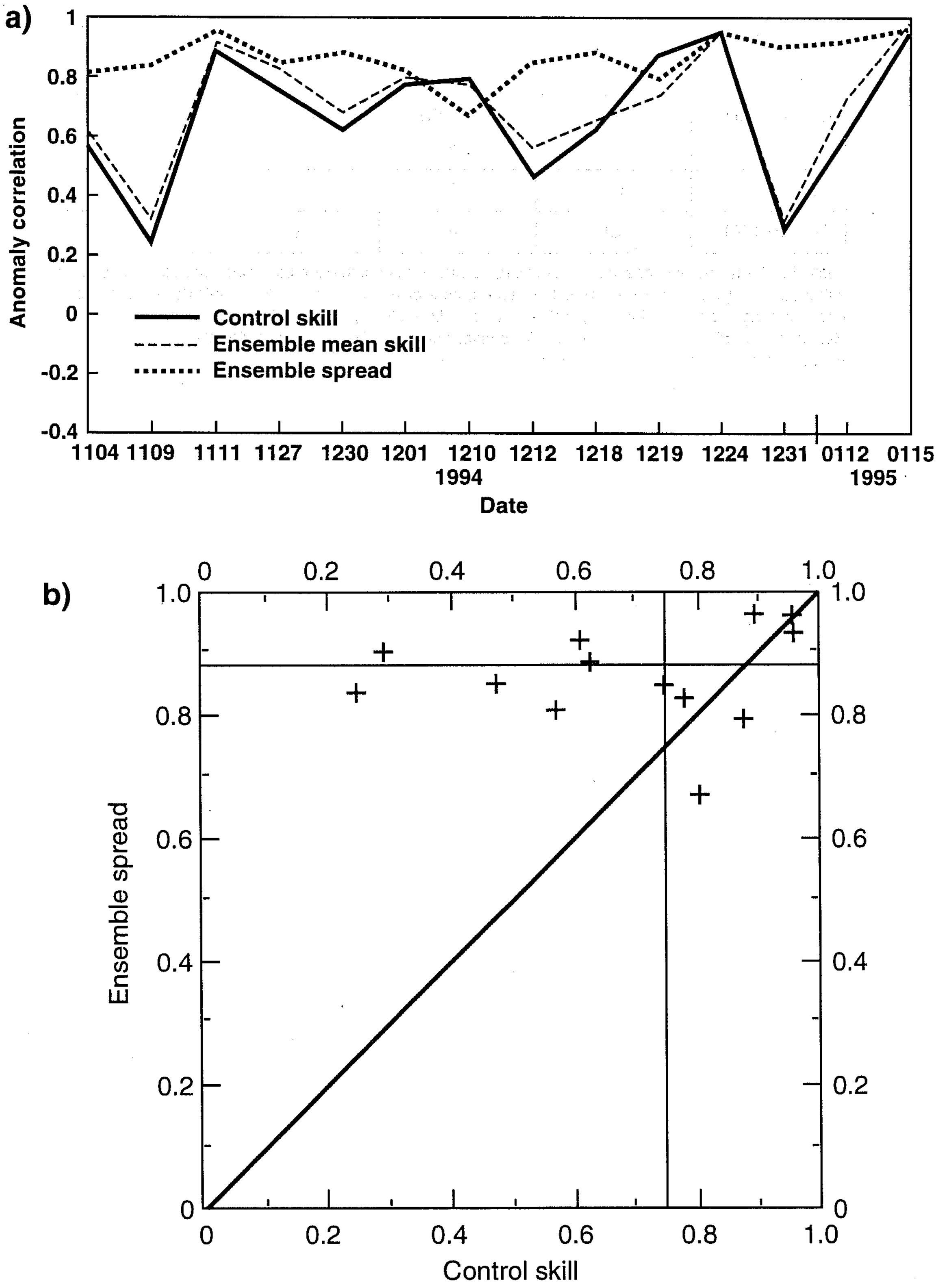

Fig. 1(a) Anomaly correlation based skill of the control (solid) and of the ensemble- mean (dash), and ensemble spread around the control (dot), of the 500 hPa geopotential height over Europe, at forecast day 5, and (b) scatter diagram of the (anomaly correlation based) control skill (abscissa) versus the ensemble spread around the control (ordinate), for the 14 case studies. [Note that two crosses almost overlap at point (0.62;0.87).]

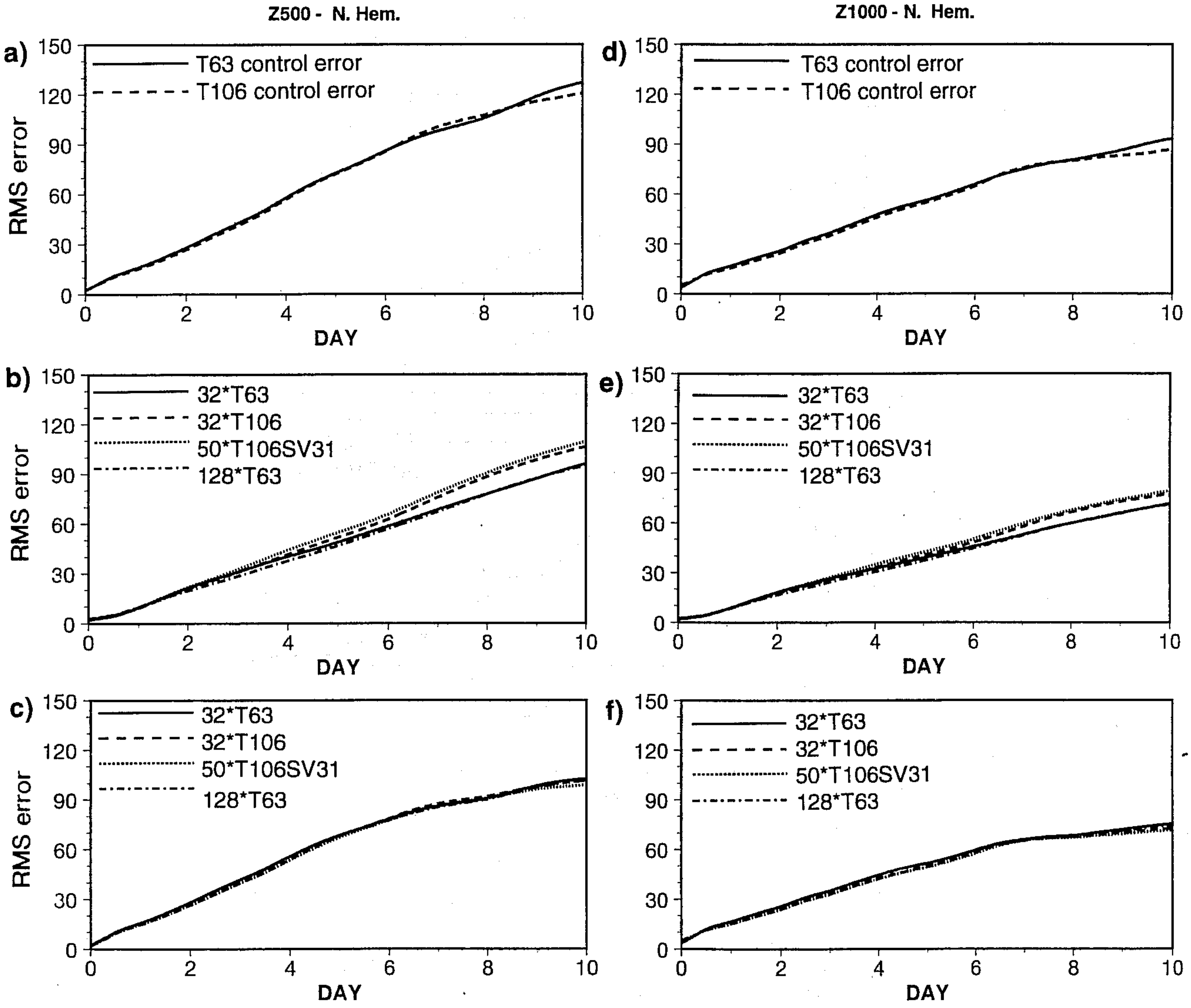

Fig. 2(a) Average rms error of T63L19 (solid) and T106L31 (dash) control forecasts, (b): average rms spread of 32*T63 (solid), 32*T106 (dash), 50*T106SV31 (dot) and 128*T63 (chain-dash) ensembles, (c): as (b) but for the rms rror of the ensemble mean, relative to the 500 hPa geopotential height over the Northern Hemisphere. (d-e-f): as (a-b-c), respectively, but for the 1000 hPa geopotential height. Abscissa: forecast day. Ordinate: rms values.

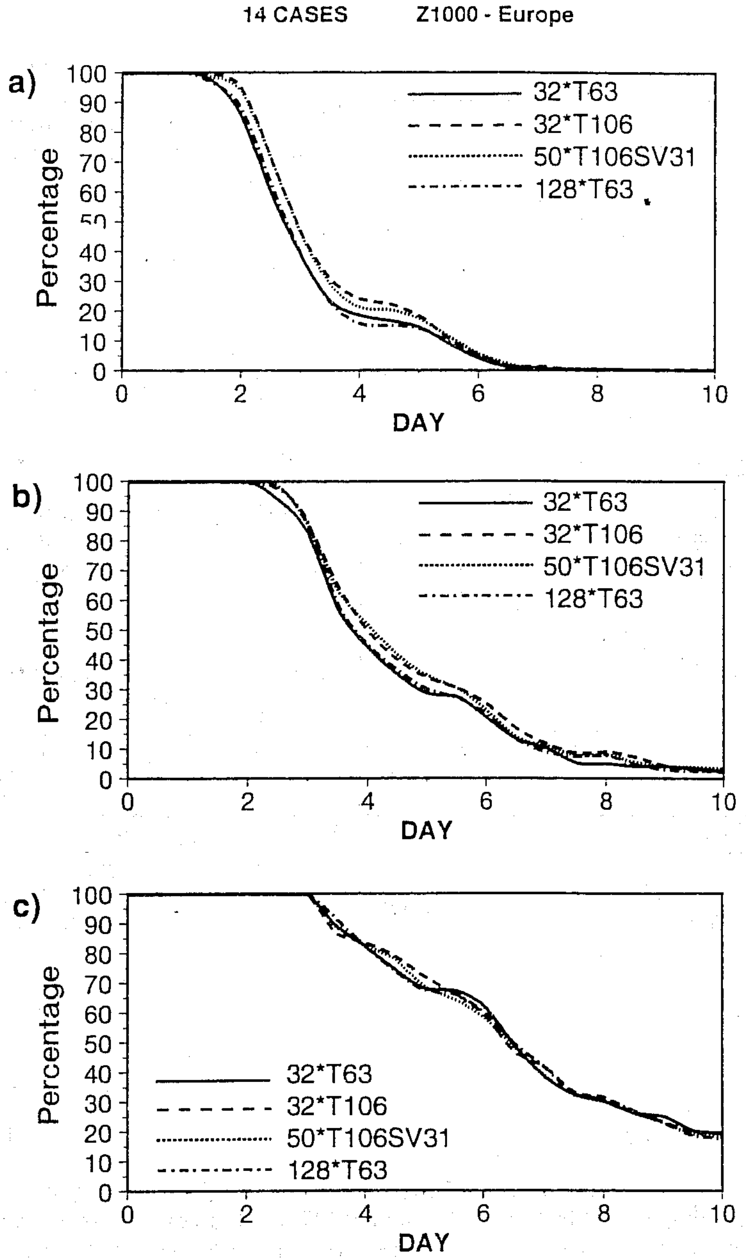

Fig. 3 Percentage of perturbed members with anomaly correlation skill higher than (a) 0.9, (b) 0.8, and (c) 0.6, in 32*T63 (solid), 32*T106 (dash), 50*T106SV31 (dot) and 128*T63 (chain-dash) ensembles. Values refer to the 1000 hPa geopotential height over Europe. Abscissa: forecast day. Ordinate: percentage.

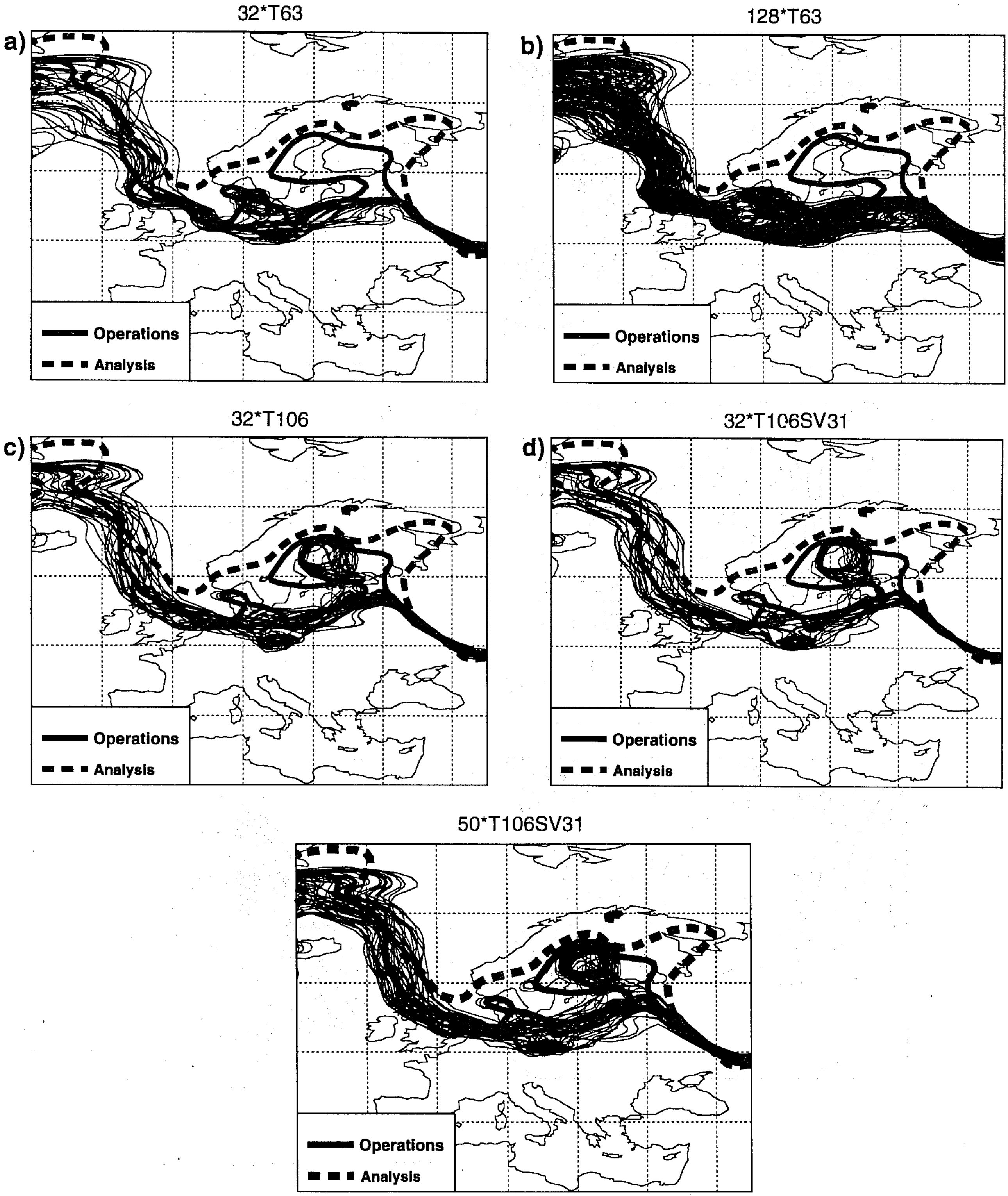

Fig. 4 94.12.12 case, 850 hPa Temperature: position of -8°C isotherm of the analysis (dash), T213L31 deterministic prediction (bold solid), and ensemble members (thin solid), of configurations (a) 32*T63, (b) 128*T63, (c) 32*T106, (d) 32*T106SV31 and (e) 50*T106SV31.

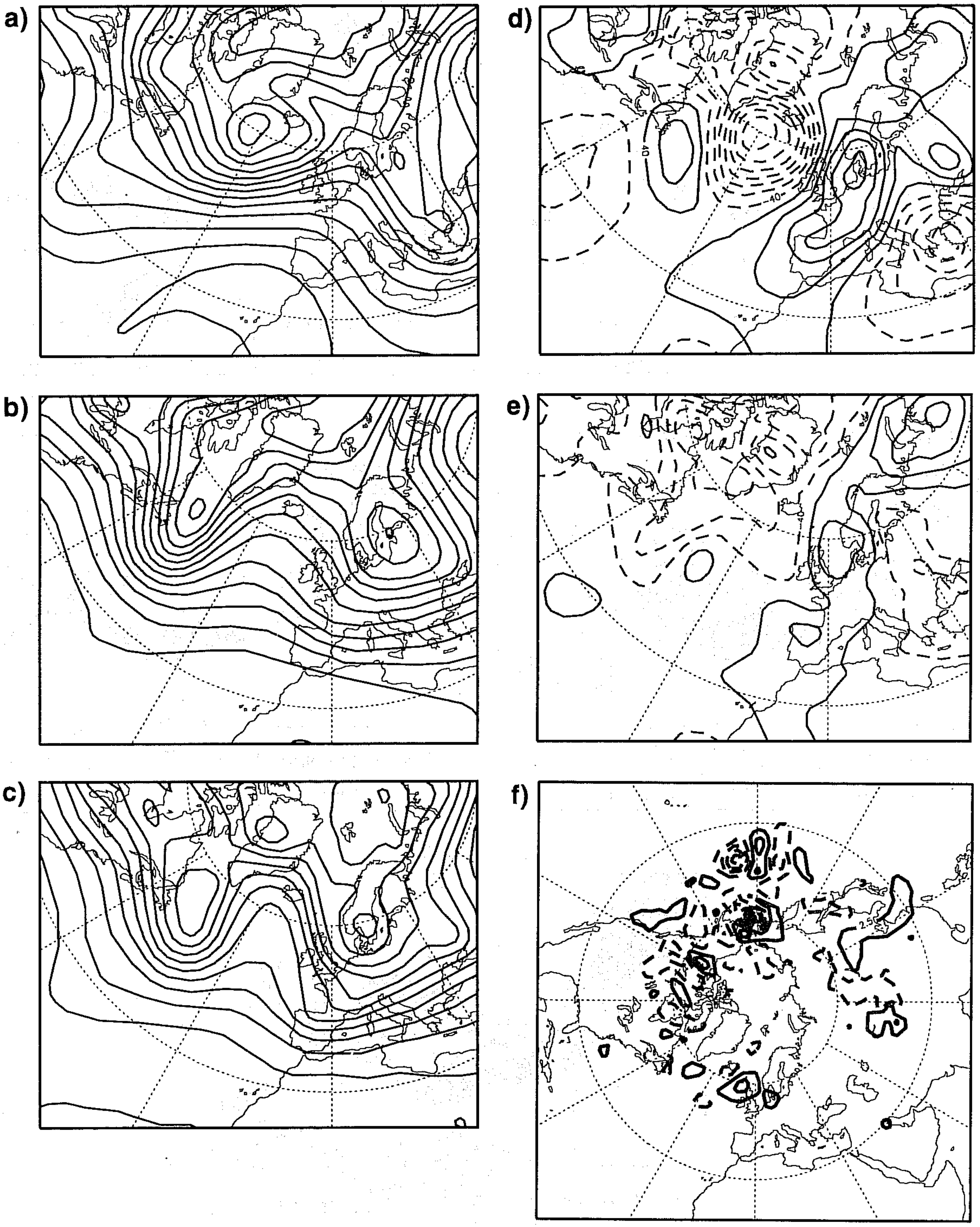

Fig. 5 94.12.24 case, configuration 50*T106SV31, 500 hPa geopotential height: (a) T106 control and (b) best member prediction at forecast day 8, (c) analysis for 95.01.01, (d) error of the control, (e) error of the best forecast, and (f) initial perturbation used to generate the perturbed initial condition. Contour isoline (a-e) 80 m, and (f) 5 m.

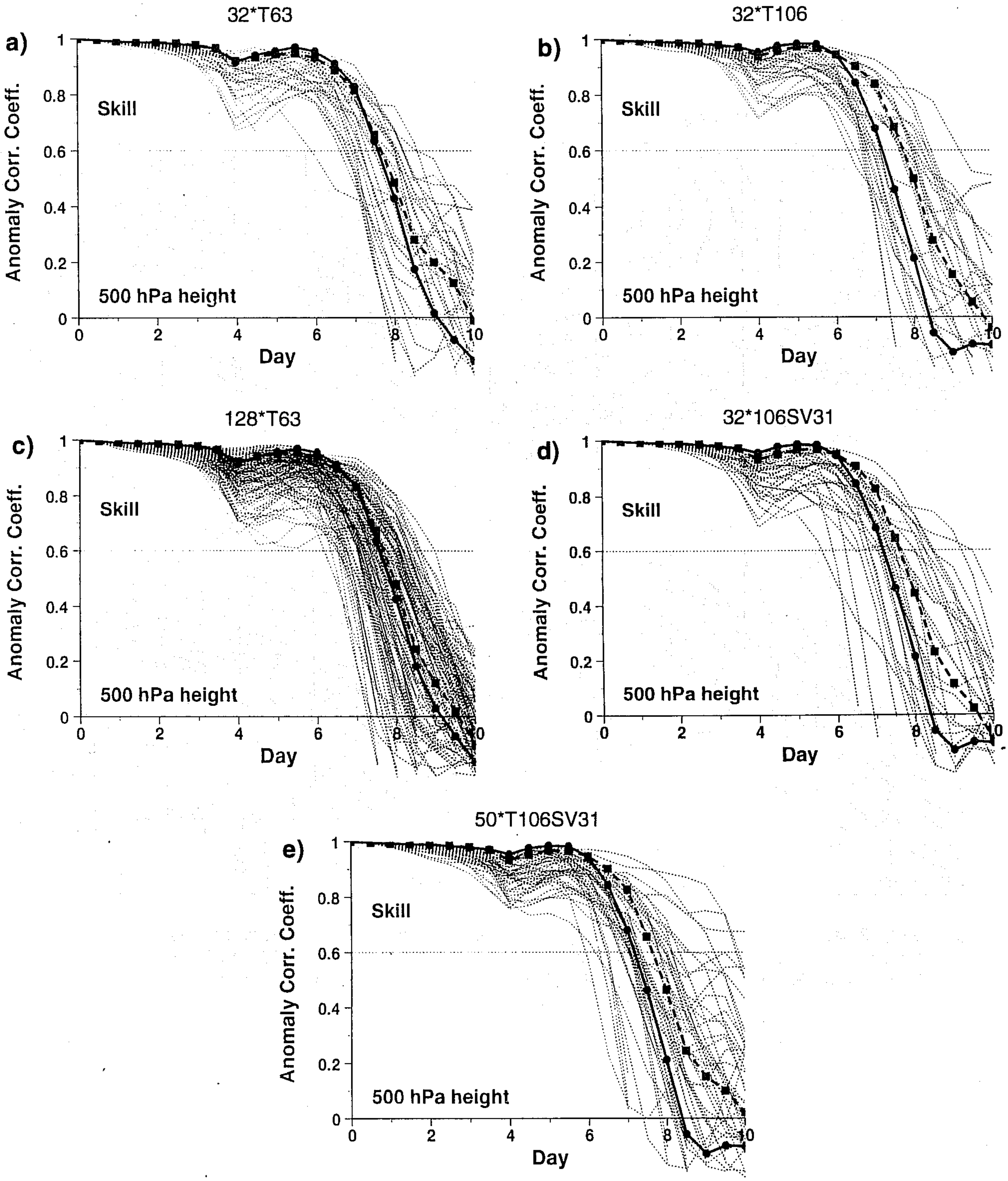

Fig. 6 94.12.24 case, 500 hPa geopotential height: anomaly correlation skill over Europe of the predictions given by the control (solid), the ensemble-mean (dash) and the perturbed members (dot) of configuration (a) 32*T63, (b) 32*T106, (c) 128*T63, (d) 32*T106SV31 and (e) 50*T106SV31.

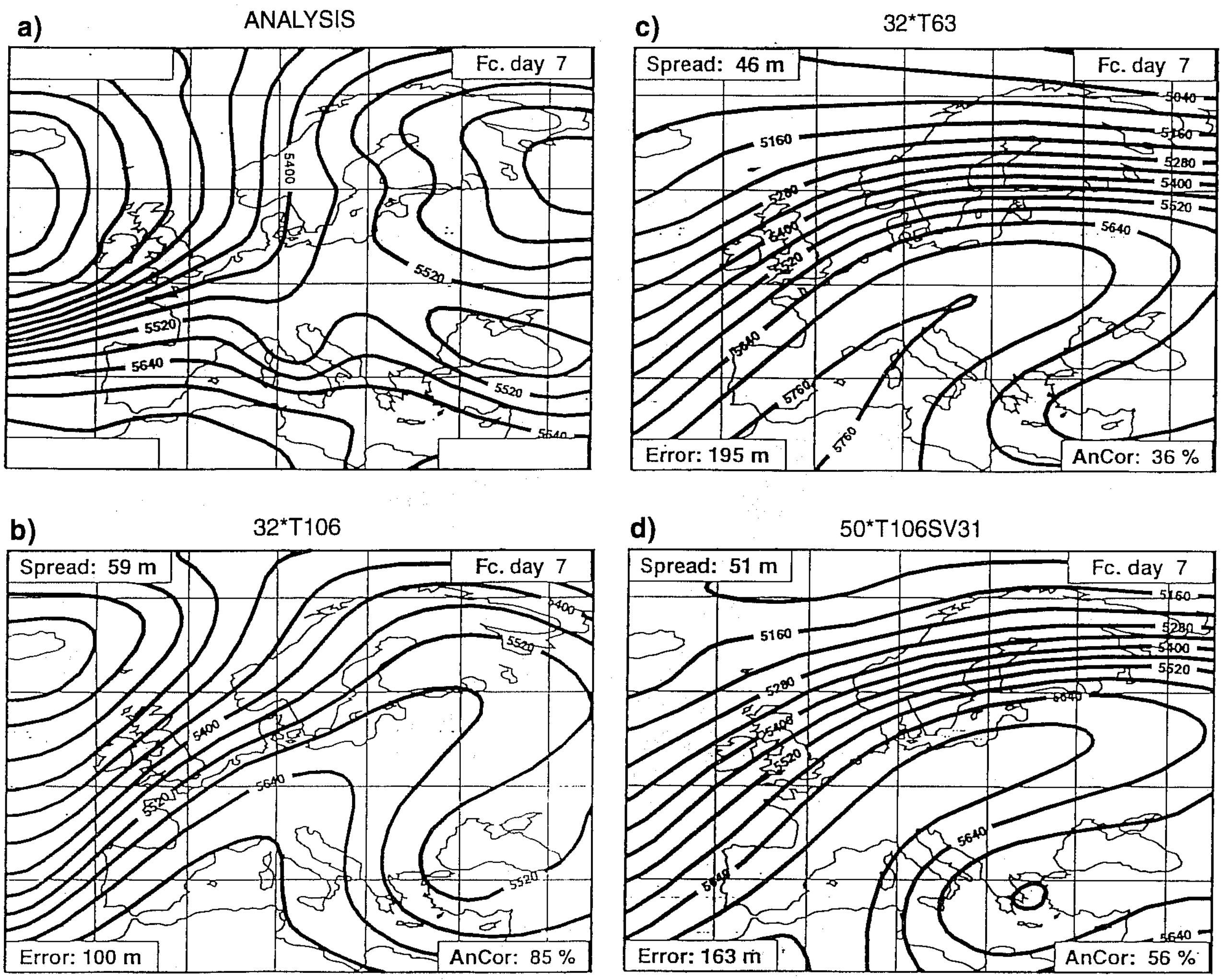

Fig. 7 5.01.12 case, 500 hPa geopotential height: (a) analysis corresponding to forecast day 7, and prediction for forecast day 7 of the best cluster from configuration (b) 32*T106, (c) 32*T63 and (d) 50*T106SV31. Contour isoline 40 m. For panels b-d, the top-left value is the cluster spread, the bottom-right value is the cluster rms error, and the bottom-left value is the cluster anomaly correlation skill.

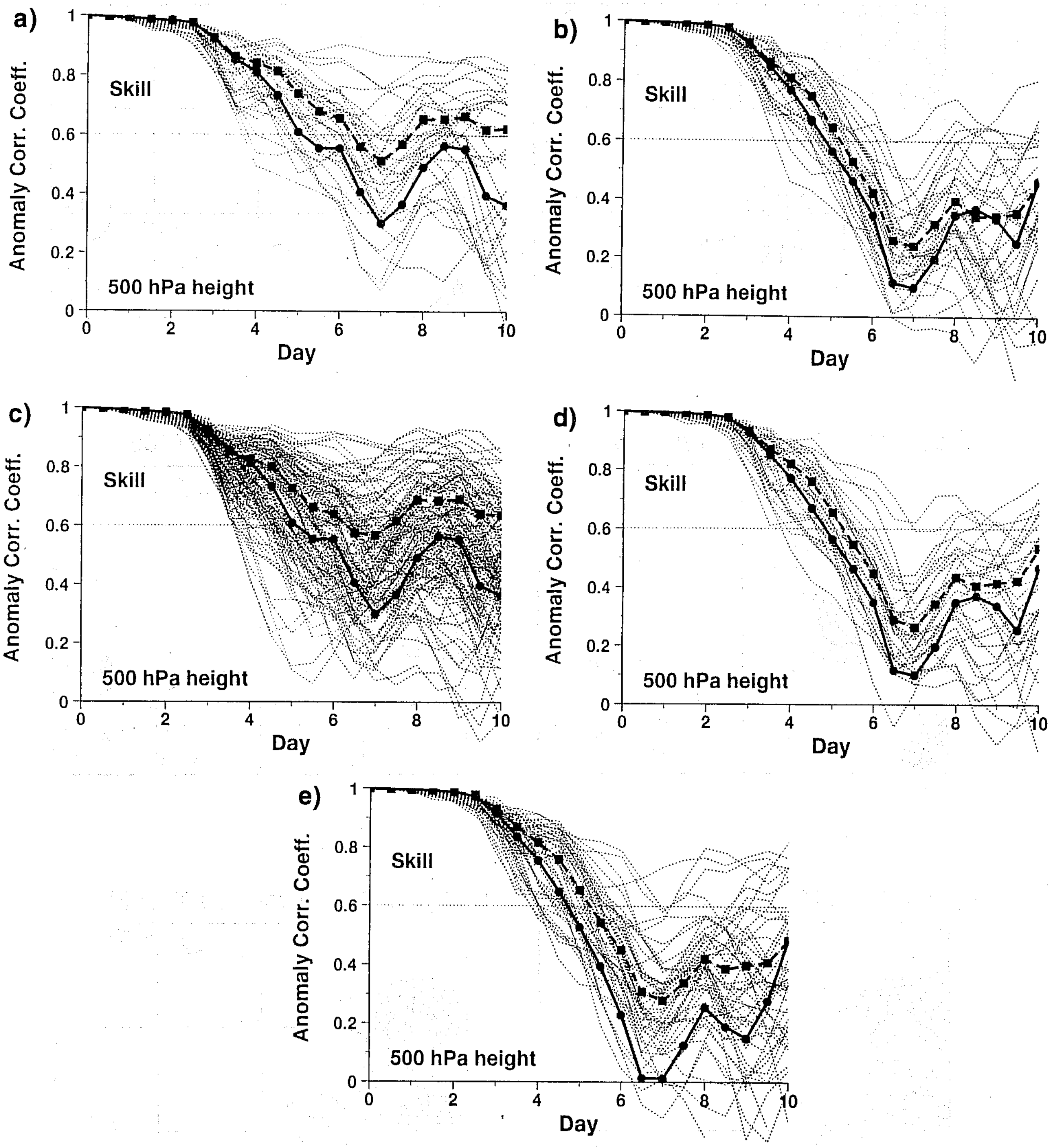

Fig. 8 95.01.12 case, 500 hPa geopotential height: anomaly correlation skill over Europe of the predictions given by the control (solid), the ensemble-mean (dash) and the perturbed members (dot) of configuration (a) 32*T63, (b) 32*T106, (c) 128*T63, (d) 32*T106SV31 and (e) 50*T106SV31.

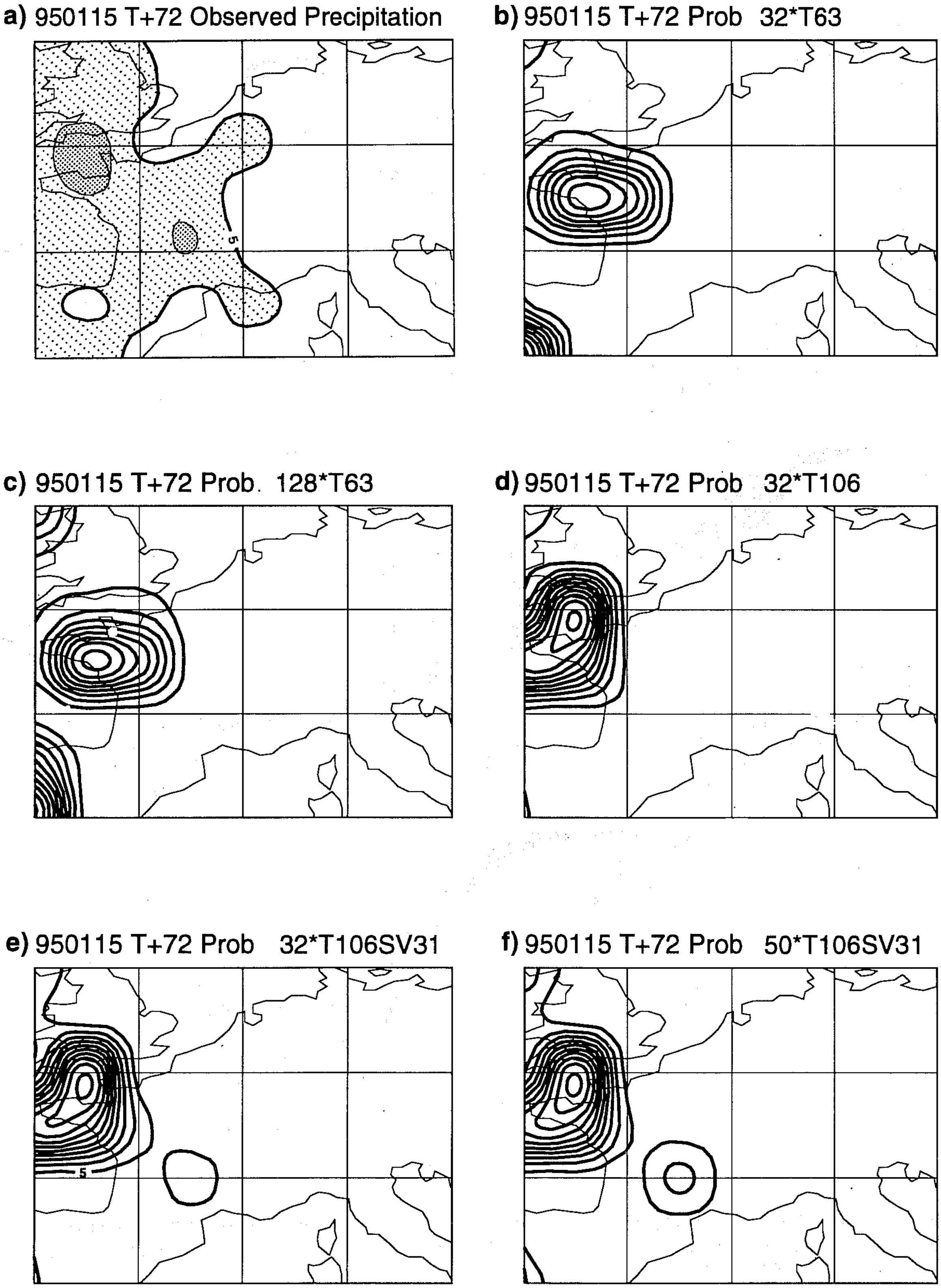
a) 950115 T+72 Observed Precipitation
5
b) 950115 T+72 Prob 32*T63
c) 950115 T+72 Prob. 128*T63
d) 950115 T+72 Prob 32*T106
e) 950115 T+72 Prob 32*T106SV31
5
f) 950115 T+72 Prob 50*T106SV31

Fig. 9 5.01.15 case, (a) precipitation observed between January 17 and 18, i.e. between forecast day 2 and 3 (shading identifies regions with more than 5 mm/day and 15 mm/day of rain), and probability of more than 15 mm/day of rain cumulated between forecast day 2 and 3 predicted by configuration (b) 32*T63, (c) 128*T63, (d) 32*T106, (e) 32*T106SV31 and (f) 50*T106SV31. Contour isoline for probabilities 10%.

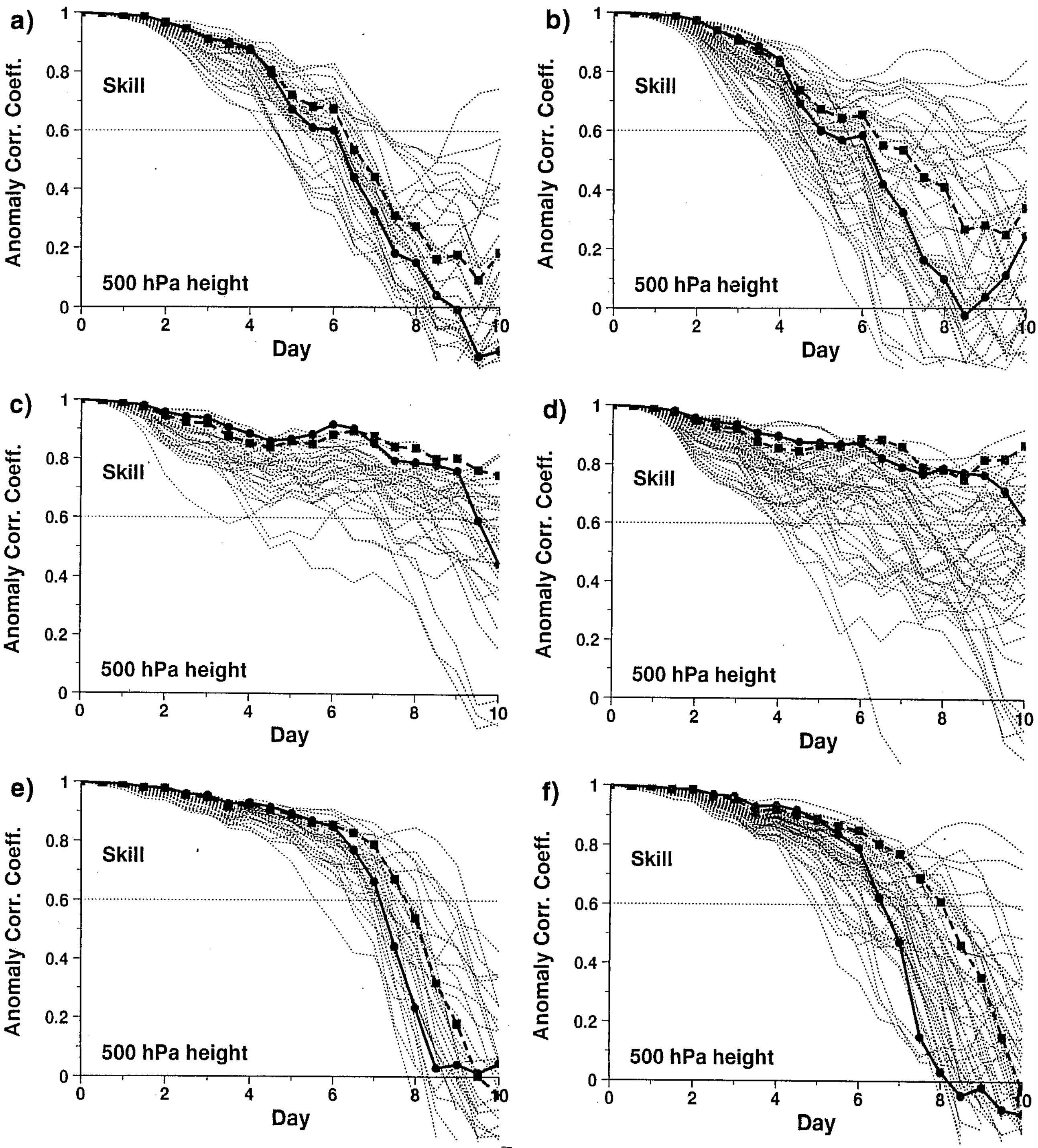

Fig.10 Anomaly correlation skill of the control forecast (solid), of the ensemble-mean (dashed) and of the perturbed ensemble members (dotted), of ensembles run in configurations 32*T63 (left panels) and 50*T106SV31 (right panels), with starting dates 96.09.21 (top panels), 96.09.22 (middle panels) and 96.09.24 (lower panels). Values refer to the 500 hPa geopotential height over Europe.