TECHNICAL MEMORANDUM

188

Ensemble Prediction

Palmer, T., F. Molteni, R. Mureau,
R. Buizza, P. Chapelet & J. Tribbia

Research Department

August 1992

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

Ensemble prediction

Document presented to the 20th Session of the Scientific Advisory Committee, 1992

ABSTRACT

A technique is described which explicitly uses information about the geostrophically-balanced dynamical instabilities of the flow to construct initial perturbations for ensemble prediction. Preliminary results from earlier experimentation are summarised. The technique is applied in the construction of a set of 24 15-day ensemble forecasts, where each ensemble comprises 32 individual integrations of the T63L19 Cycle 36 model.

The set of ensemble forecasts is validated using Brier and Ranked Probability Scores on the 500 hPa height and 850 hPa temperature fields. It is found that the ensemble forecasts are more skilful than the control (or deterministic) forecasts beyond about day 2. However, beyond day 7, the ensemble forecasts do not consistently beat a (probability) forecast of climatology. The more conventional spread/skill correlations are examined. These indicate that the ensemble spread is a good predictor of skill in the short-range. A number of practical probability forecast products are shown, concentrating on days 5-7 where they may have most impact on operational practice.

Improvements to the technique are described, including the calculation of the dynamical instabilities in the same primitive equation model used for the nonlinear integrations, and use of an estimate of initial error covariances as a constraint in the calculation of these instabilities.

Outlines of a real-time ensemble prediction experiment planned to begin in the winter of 1992/93 are given.

1. INTRODUCTION

Ensemble prediction is a potential method of estimating forecast predictability beyond the range, typically 3 days, in which error growth can be described by linearised dynamics. The ensemble forecast is a set of integrations of a deterministic numerical weather prediction (NWP) model; these integrations differ only in their initial states, reflecting uncertainties in such starting conditions. In essence, the ensemble prediction is an attempt to estimate the nonlinear evolution of the forecast error probability distribution function (PDF) through a finite sample of deterministic forecast integrations.

One of the fundamental difficulties in this approach lies in the choice of initial conditions for the individual members of the ensemble. From the definition above, these must reflect an adequate sampling of the probability distribution function of analysis error. However, the phase-space dimension of current NWP models is well in excess of a million; this can be compared with a maximum practicable ensemble size, currently less than one hundred. Hence, without some strategy for determining ensemble initial conditions, sampling error will be too large to make ensemble prediction useful.

Among the consequences of inadequate sampling are an underestimation in ensemble dispersion, and an overestimation of forecast predictability. In such a situation the individual members of an ensemble might each forecast a common but unrealistic flow type. An example of this can be seen in earlier studies (e.g. *Hollingsworth*, 1980) where perturbations to an operational analysis did not survive initialisation, though the problem remains endemic (e.g. *Brankovic et al.*, 1990).

During the linear phase of error growth, an initial isotropic error ball will evolve into an ellipsoid. Initial phase-space directions can be characterised according to whether they map onto the amplifying or decaying axes of the ellipsoid. Further evolution of the ellipsoid into the medium range will be nonlinear, and will ultimately reflect the multi-modal nature of the weather-regime structure of mid-latitude flow (*Molteni et al.*, 1990).

With a limited ensemble size it is important, for medium-range prediction, to sample those trajectories starting within the initial error ball, which evolve to significantly different large-scale flow patterns, such as weather regimes. Our results suggest that a promising strategy for picking the trajectories which define these fundamental medium-range forecast alternatives is: first to calculate the initial phase-space directions which are associated with the dominant major axes of the short-range forecast error ellipsoid, and then to integrate, nonlinearly, the members of the ensemble from the initial conditions associated with these directions. In estimating the forecast error PDF with such a procedure, larger weight should in principle be given to the control forecast from the centre of the error ball, since this trajectory will be representative of all the decaying (or weakly growing) phase-space directions, not chosen for explicit integration in the ensemble.

The techniques to calculate the growing perturbations from both composite and realistic initial states has been described in *Molteni and Palmer* (1992) using a 3-level quasi-geostrophic (QG) model. A limited study of the use of such perturbations in a set of ensemble forecasts carried out with the ECMWF T63 model was described by *Mureau et al.* (1992). Results from these two papers are summarised in section 2.

In practice, the initial error distribution is not isotropic in phase space. The spatially inhomogeneous nature of data coverage makes the structure of the PDF of initial error rather complex. Nevertheless, for realistic ensemble prediction studies, some account of these inhomogeneities must be taken. The study described in this paper uses the QG instability calculations, modulated by knowledge of the Optimal Interpolation (OI) analysis error variance fields, to generate a set of 24 ensembles of integrations of the ECMWF T63 model. Each ensemble forecast comprises 32 individual integrations. Details of the technique used to generate the initial conditions for the ensembles are described in section 3.

A second difficulty with the ensemble technique lies in the post-processing of forecast products. With tens of individual integrations, the choice of appropriate and manageable post-processed fields is problematic. There are two (somewhat differing) demands - firstly the need to verify objectively the ensemble forecast, and secondly the requirement to provide useful forecast products. Attempts to fulfil these different needs are described in sections 4 and 5 respectively.

The ensemble technique described in this paper is still preliminary. Two further developments are currently being tested: firstly, linearised calculations of the semi-major axes with the same primitive equation model used for the ensemble forecasts, and secondly the incorporation of initial error covariance estimates into the calculation of the semi-major axes. Some preliminary results are included in section 6. Plans for a real-time ensemble-forecast experiment are described in section 7.

2. SUMMARY OF EARLIER RESULTS

The formulation of the experimental procedure described in this paper follows studies of finite-time linear instabilities in a T21L3 QG model. Experiments were made to assess how these instabilities grow when interpolated onto the T63L19 grid and integrated nonlinearly in a primitive equation model beyond the range in which linearised dynamics is applicable. These earlier studies are documented in *Molteni and Palmer* (1992) and *Mureau et al.* (1992), results from which are briefly summarised here.

For chaotic systems, trajectories which are initially sufficiently close, diverge asymptotically at an exponential rate given by the positive Lyapunov exponents characterising the average predictability of the whole attractor set. The divergence rate for finite segments of a trajectory are given by the singular values of the 'propagator' $\boldsymbol{R}(t_1,t_2)$ of the linearised equations of the dynamical system, integrated over a nonlinear trajectory segment between time t_1 and t_2. For finite t_2-t_1, the divergence rates need not be exponential. The corresponding phase-space directions associated with the divergence rates are the singular vectors (SVs) of $\boldsymbol{R}$, equivalent to the eigenvectors of $\boldsymbol{R}^*\boldsymbol{R}$ (*Lorenz*, 1965; *Noble and Daniel*, 1977; *Lacarra and Talagrand*, 1988). Here "*" denotes the adjoint operation, defined in the work below with respect to a kinetic energy (QG model) or total energy (primitive equation model) inner product.

In *Molteni and Palmer* (1992), the optimal SVs were calculated in both barotropic and QG models, for a number of basic state flows, including both composite circulation patterns representative of observed weather regimes, and particular realisations of such regimes. The main conclusions can be summarised as follows:

i) For integration times up to about 15 days, the growth rates of the optimal SVs are up to an order of magnitude larger than the most unstable normal mode instability. The amplitude of such SVs can double in less than 12 hours.

ii) Optimal SVs in a barotropic model grew significantly faster in a weather regime which has a negative Pacific/North American (PNA) teleconnection index (anticyclonic flow in the NE Pacific), compared to one which had a positive index (cyclonic flow in NE Pacific), consistent with earlier studies on relationships between medium-range forecast skill and the PNA index (*Palmer*, 1988).

iii) Unlike in the barotropic model, the linear evolution of optimal baroclinic SVs on a smooth basic state with only planetary-scale structure has little upscale energy cascade from synoptic to planetary scales. However, such upscale energy transfer occurs in the baroclinic model when the basic state has synoptic-scale structure (using basic state averages over short periods of time or, more appropriately, unsmoothed time-evolving basic states). After a few days on a time-evolving basic state, initially small-scale linear baroclinic SVs develop a large-scale equivalent barotropic structure.

iv) If baroclinic SVs with initial amplitudes comparable to analysis errors are allowed to evolve non-linearly on a time-dependent basic state, then the non-linear self interactions also involve upscale barotropic energy transfer to large-scale equivalent barotropic structures. These interactions tend to damp the optimal linear growth described in iii).

Baroclinic SVs optimised for short time intervals, developed in an efficient and realistic way when superimposed on time-evolving basic states. This suggested a strategy for the use of SVs in non-linear ensemble predictions with the primitive equation model.

In a test of this strategy (*Mureau et al.*, 1992), four wintertime initial states were chosen, three at random, and one because of substantial development in the large-scale flow within four days, which the control forecast completely missed. A set of SVs was created using the QG model linearised about basic states taken from data close to the chosen initial dates. These SVs were interpolated onto the primitive equation model grid, and used as perturbations to the initial state. An ensemble forecast comprising 40 members was made from the perturbed initial states.

The dispersion of this ensemble was compared, for each date, with that from a second ensemble with initial perturbations constructed from 6-hour forecast errors (FEs) from days immediately preceding the initial date. All perturbations were normalised to have a 10m rms amplitude of 500 hPa height, averaged around the northern hemisphere.

The main findings of *Mureau et al.* (1992) were:

i) Throughout the (5-day) forecast period, the amplitude of the perturbations was noticeably larger using the SVs than using the FEs.

ii) In the case with substantial development, the dispersion of the ensembles using the FE perturbations did not indicate that the control forecast was likely to be poor. Indeed the synoptic development

in all elements of the FE ensemble was similar at day 4. By contrast, the dispersion of the ensemble using the SVs was notably larger for this case than the other three. A number of members of the SV ensemble were particularly skilful in predicting weather-related elements of the flow, such as 850 hPa temperature change.

iii) The overall rms dispersion of an ensemble integrated with the QG was better correlated with skill than the T63 ensemble. The T63 spread/skill correlation was improved if the envelope or maximum dispersion was used. It was concluded that these T63 problems were associated with inconsistencies between the QG and T63 models, particularly with regard to vertical resolution and orography.

iv) It was found that the evolution of perturbations, initially localised over the western Pacific, or western Atlantic, developed blocking-like structures several days later, over the eastern oceans. Case studies of blocking development in the real atmosphere have indicated similar development (*Uhl et al.*, 1992).

3. EXPERIMENTAL DESIGN

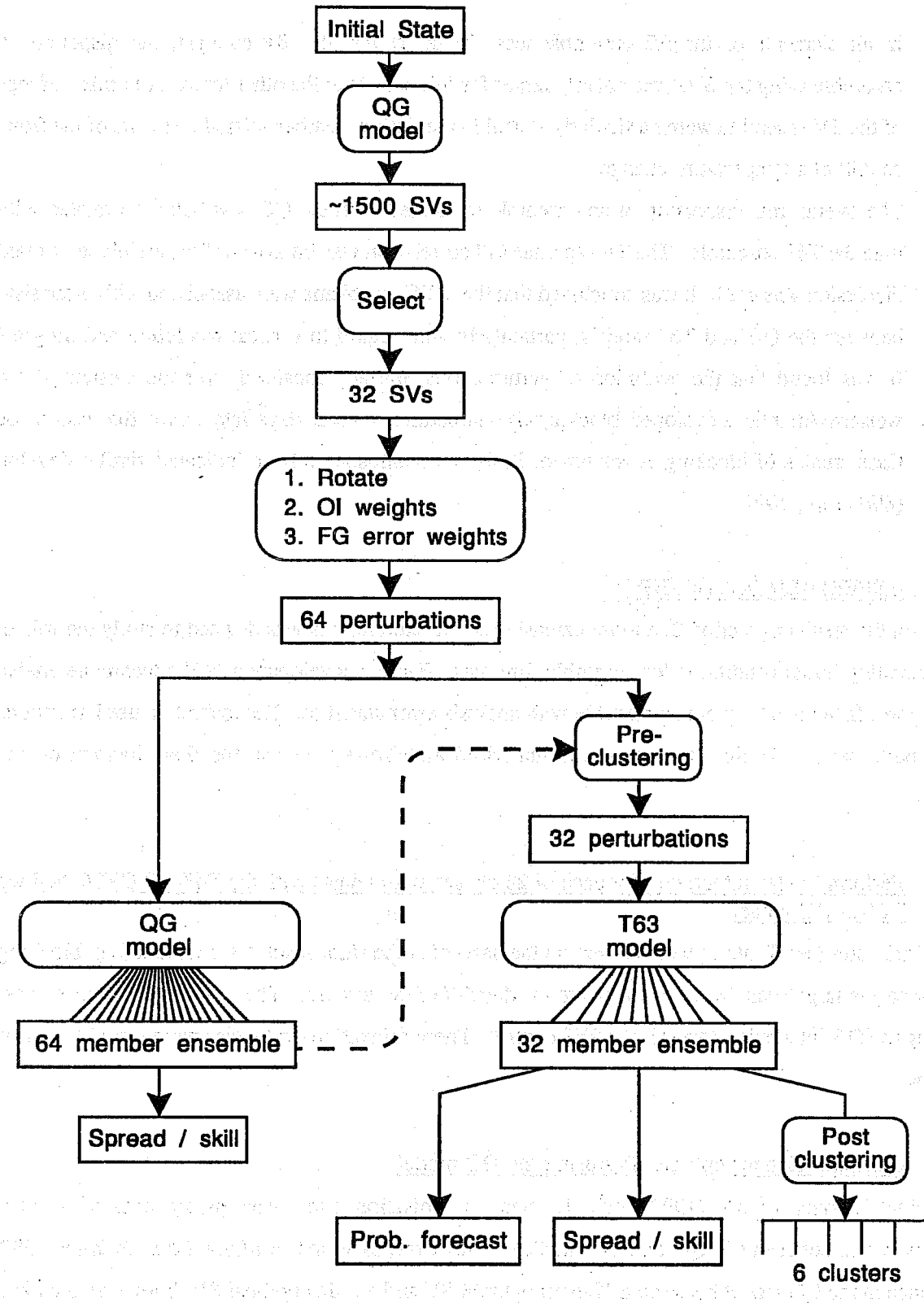
Based on the results in section 2, a more extensive set of experiments was devised to study the role of SVs for generating initial conditions for ensemble forecasts. Further development of the technique enabled the initial perturbations to be more compatible with analysis error statistics. The technique used to generate the initial perturbations for this study can be summarised as follows (see also the flow diagram on the next page):-

3.1 Perform 15-day integration for each of 24 chosen initial dates with the T63L19 CY36 and with the T21L3 QG model

The initial dates (see Table 1) were chosen on the basis of some significant development (e.g. blocking, cut-off low etc) in large-scale flow occurring around day 5-7 of the forecast. The control QG forecast was made relaxing the QG integration towards the T63 control. These relaxation tendencies were stored for use in step 6 below.

3.2 Calculate 12-hour optimal SVs using the QG model

Following *Mureau et al.* (1992), the 12 hour optimisation time was partly chosen to minimise inconsistencies between the QG and T63 models. However, as found in *Molteni and Palmer* (1992), the difference in the QG model between a 12-hour optimal SV and a 2-day optimal SV, both evaluated at day 2, is not very great. Using the QG model at T21L3 resolution, the entire spectrum of SVs was available. Some 500 hPa height perturbations from a selection of SVs are shown in Fig. 1 from data from 17 January 1989.



Flow diagram of Ensemble Prediction Procedure.

3.3 SVs with maxima over orography and over the Southern Hemisphere were rejected.

The rejection of SVs over orography was necessary because of severe inconsistencies between the QG and primitive equation model formulations. Rejection of SVs over the Southern Hemisphere was based on the assumption that the effects of initial errors in the Southern Hemisphere on the Northern Hemisphere forecast were relatively unimportant over the medium-range forecast period.

3.4 Set amplitude of 1st 32 remaining SVs to be consistent with the Optimal Interpolation (OI) estimate of analysis error

Details of the calculation are given in the appendix, but can be summarised as follows. Let E_i denote the 500 hPa height components of the set of SVs, then we find coefficients σ_i such that $\frac{1}{N}\sum_{i=1}^{N}(\sigma_i E_i)^2$ is as close as possible to a statistical estimate of initial error of 500 hPa height. The set σ_i are then given by a variant of standard linear regression analysis methods. In practice, since it is widely felt that the OI estimate of analysis error is not particularly reliable, the estimate of initial error variance was chosen to be the maximum at any grid point of the OI estimate, and the magnitude of recent short-range forecast errors.

3.5 Perform phase-space rotation to delocalise individual perturbations

As shown in Fig. 1 individual SVs can be somewhat localised. Such SVs are not representative of hemispheric-wide analysis error. Without some delocalisation, ensemble dispersion on a regional scale could be dependent on the time taken for SVs to disperse into the region of interest. In practice delocalisation was achieved by making a linear combination of SVs, equivalent to a phase space rotation (see appendix). 500 hPa height perturbations from a selection of normalised rotated SVs are shown in Fig. 2.

3.6 Make 64-member 15-day ensemble with the QG model

Positive and negative perturbations of the analysis A were chosen, so that the ensemble of initial states is given by $A+f_j$, $A-f_j$, $j=1,32$ (where A is the analysis and f_j is the perturbation). The relaxation tendencies from the control QG integration were applied as a prescribed time-dependent forcing for each of the perturbed QG integrations. The tendencies can be thought of as a crude "physics" parametrization, ensuring that the QG model fields were not too dissimilar to those in the T63 model.

3.7 Use cluster analysis on day 5 QG forecast fields to identify 32 representative perturbations

As it is currently impracticable to run a 64-member ensemble with the T63 model, it was necessary to choose a subset of representative perturbations. This was done on the basis of their impact on the medium-range forecast using the QG model. The Ward hierarchical clustering algorithm (*Anderberg*, 1973) was applied to the northern hemisphere fields to reduce the perturbations to 32.

3.8 Make 32-member 15-day ensemble with T63 model, and archive selected fields

Only a small set of fields was archived because of storage constraints.

4. VERIFICATION

4.1 Spread-skill correlations

Perhaps the most familiar measure of the success or otherwise of an ensemble prediction is its ability to determine the skill of the deterministic control forecast through the overall dispersion of the ensemble.

Fig. 3 shows the correlation between the skill of the control forecast, and both the 64-member QG ensemble spread, and 32-member T63 ensemble spread as a function of forecast range. The correlations are calculated from 500 hPa height over the whole northern hemisphere, and are presented for both anomaly correlation coefficient (ACC) and RMS error.

Before discussing these correlations it should be remarked that, even for a perfect model the correlation between ensemble spread and skill of the control will not necessarily equal unity (*Murphy*, 1988). In using the ensemble spread to forecast the skill of the control there is an implied sampling error; the spread could just as well be used to forecast the skill of some other members of the ensemble. In a perfect model we can assume that one element of the ensemble is identical to the verifying analysis. Results from our ensembles suggest a maximum spread/skill correlation of 0.8 throughout the forecast range.

For ACC (Fig. 3a) correlations between day 1.5 and 4 are around 0.6 in both models. Thereafter the QG correlations drop to zero, whilst the T63 correlations only drop below 0.5 by about day 6. For RMS correlations (Fig. 3b), the QG model appears more skilful than the T63 out to about day 10. At day 3 the QG RMS correlations are not far from the maximum value attainable.

Inevitably, the curves in Figs. 3 a,b show some day-to-day noise. These are effectively filtered out in Fig. 3c, which illustrates the ACC spread/skill correlations for 5-day mean fields. Again the QG model correlations appear superior in the medium range, whilst the T63 model results are fairly independent of forecast range.

Tests (using 32-member QG ensembles) have shown that the apparent superiority of the QG model in the medium range does not arise from its larger ensemble size. Rather, we believe, it is associated with the consistency of the nonlinear integrating model with the initial perturbations. It is anticipated that the use of the primitive equation model to calculate the SVs in future experiments (see sections 6,7) will enhance the T63 spread/skill correlations above those shown here.

The measure of spread used in these calculations involves the calculation of the RMS distance from the control of the individual forecast elements. This could be referred to as an l^2 measure of spread. We have also calculated spread/skill correlations using the l^∞ (or envelope) measure, defined as the maximum distance between ensemble members and the control. The ACC l^∞ spread/skill correlation calculated on 5-day running mean fields is shown in Fig. 3d. It can be seen, beyond about day 5, that these correlations are not as high as those in Fig. 3c. However, in the short range, there is some improvement, particularly for the T63 ensembles (e.g. at day 3 the l^2 correlation is .62, whilst the l^∞ correlation is .68, corresponding to an 8% increase in explained variance). This result is consistent with earlier studies from *Mureau et al* (1992) again suggesting problems interpolating some of the QG SV perturbations onto the T63 grid.

4.2 Ranked Probability Score

In order to validate the ensemble forecasts without specific reference to the control, we have calculated Ranked Probability Scores (RPS) on 850 hPa temperature anomaly categories over Europe. Three categories, defined as less than -2 K below climatology, within 2 K of climatology, and greater than 2 K above climatology were chosen. Climatologically, each category is approximately equiprobable.

The RPS (*Epstein*, 1969) is a derivative of the Brier score, and is more appropriate then the latter in situations where the given categories can be ordered. The Brier score is obtained as the sum over the given categories of the mean square difference between the ensemble forecast probability for that category, and the *a posteriori* probability: 1 for the observed cluster, 0 for the others. (The control forecast probabilities equal 1 for the predicted category, 0 for the others.) The Brier score for climatology is given using the climatological probabilities for the forecast categories. The RPS is defined in a similar way to the Brier score, except that instead of being based on the categorical probabilities, they are based on the cumulative probabilities that the ensemble forecast lies in a category equal to or less than a given category. The mean square difference between forecast and analysis is then calculated on these cumulative probabilities. Compared with the Brier score, the RPS does not penalise a forecast which is one category in error as much as a forecast which is two categories in error.

Fig. 4-5 shows maps of RPS over Europe for days 5 and 7 respectively. In these Figures, the top map shows the RPS score for the ensemble forecast, the bottom shows the RPS score for the control forecast. Where the map is unshaded (RPS>44), the forecast is inferior to a probability forecast of climatology. Otherwise, the heavier the shading, the better the RPS score.

It can be seen at day 5, that whilst the RPS score from the ensemble forecast is almost everywhere better than climatology, the RPS score for the control is worse than climatology over about half the area shown.

Moreover, the ensemble forecast is more (RPS) skilful than the control almost everywhere. At day 7, whilst the ensemble RPS score is more skilful than climatology over most of Europe, the deterministic forecast is less skilful than climatology over most of the area.

The mean RPS scores, averaged over Europe (30-80N, 30W-30E), are shown in Fig. 6, as a function of forecast time. Consistent with the results above, the ensemble is always more skilful than the control, but more skilful than climate up to day 7 (beyond which the score has effectively asymptoted). Some improvement in the RPS score of both control and ensemble, relative to climatology, could be made if the forecasts were corrected by the mean bias of the model at each grid point. Such a bias correction, however, would not change the relative skill of the ensemble forecast compared with the control.

4.3 Categorical prediction of climatological "Grosswetterlagen"

The Ward hierarchical clustering algorithm was applied to 12 winters of 500 hPa height analyses to define a set of 12 basic climatological types over Europe. These 12 clusters together explain 41% of the total daily variance over the 12 winters. These clusters are illustrated in Fig. 7, together with the cluster frequency determined firstly from the analysis (top left hand of each panel) and secondly from a set of 15 120-day integrations of the T63 Cy 36 model (top right hand of each model). There are substantial biases in the model climatic frequency distributions. Note for example the model's overestimation of the frequencies of cluster 1 (zonal flow) and the Scandinavian ridge (cluster 12), and the underestimation of frequencies of east Atlantic blocking (e.g. clusters 8, 9, 10). These are consistent with known errors in the climate of the model.

Categorical forecasts were made from the 14 winter (DJF) ensemble cases. Fig. 8 shows the Brier score of the control (dashed line), the ensemble (solid line), and climatology, as categorical forecasts of these 12 clusters. The Brier score assessment for these clusters tests the ensemble in a different way to the RPS scores given above; firstly because it is not possible to order the clusters, secondly because the number of categories is much larger, and finally because only winter cases are chosen. (There is some evidence - *Ferranti et al.*, 1992 - that the model climatology of low-frequency variability over the Euro/Atlantic sector is more accurately represented in the transition seasons than in winter.) Results show that up to day 2, both control and the ensemble are essentially perfect. Beyond day 2, the ensemble forecast is superior to the control. However, at day 6 and beyond, neither forecast beats climatology.

We have studied the correlation between the skill of the most populated forecast cluster, and its population. Consider a forecaster who issues a deterministic forecast based on the cluster with largest ensemble population, but only provided that population exceeds some threshold (and issues no forecast if the largest ensemble population does not exceed this threshold). In order to determine what this threshold should be,

there are two criteria that he will have to trade off against one another. The first is that the forecasts he issues are as skilful as possible. The second is that he issues as many skilful forecasts as possible.

Fig. 9 shows these two criteria plotted for forecast days 3-6, as a function of the threshold. The light line shows the percentage of those issued forecasts that are correct. The heavy line shows the percentage of all the correct forecasts in the total sample of ensembles that will be issued. For example, for day 3, if the threshold was chosen to be 33, then every forecast issued would be correct, however, the forecaster would issue only 22% of the total number of correct forecasts, ie 78% of forecasts which turned out to be correct would remain unissued. For day 3, however, the forecaster could use a much lower threshold of 24, such that all his forecasts were still correct, and 90% of all correct forecasts would be issued. Alternatively, if he decided to use a threshold of 23, all the correct forecasts would be issued, though only 82% of the issued forecasts would be correct. (With a threshold lower than 23 all the correct forecasts would continue to be issued, though a smaller fraction of the issued forecasts would be correct.)

At day 4, the trade-off between the two criteria becomes more important. If the forecaster wanted his forecasts to be totally correct, he would choose a threshold of 28, but only issue 38% of correct forecasts. If he was prepared to tolerate a 70% accuracy of issued forecasts, then all of the correct forecasts would be issued. Similarly at day 5, a 70% accuracy of issued forecasts would ensure about 90% of correct forecasts were issued. At days 6 and 7 (not shown), the relationship becomes more difficult to interpret. For example, accepting 70% accuracy of issued forecasts at this range implies a minimum threshold of 29, which means that less than half of the correct forecasts would be issued.

Essentially these results confirm the Brier scores above, that useful skill (in predicting these clusters) is limited to about day 5. However, it is worth commenting that significant ensemble failures at day 7 occurred when the verifying analysis cluster was one whose climatological frequency was underestimated by the model. For example, at day 7, with a threshold of 27, only one of three issued forecasts would be correct. That correct forecast was for the zonal cluster 1, whilst the verifying analyses for the two incorrect forecasts were the "blocking" clusters 8 and 9. This suggests that model error limits the overall skill of the ensemble forecast towards the end of the medium range.

5. SOME ENSEMBLE PRODUCTS

Fig. 10 shows 850 hPa temperature from individual members of an ensemble at day 5 of the forecast from 3 January 1987. The ensemble includes not only the T63 control forecast (1st row, 1st column), but also the T106 operational forecast (3rd row, last column). The verifying analysis is shown in the last row, last column. In general, forecast dispersion on this European scale starts to become significant at about day 5. (For example, the forecasts in the third and fourth columns of the last row give a very different forecast of

temperature over western Europe.) Whilst the eye makes a reasonably good attempt at assimilating the common elements of such forecasts, it is necessary to process objectively the information from the ensemble in a more digestible form. The probability forecast product appears to be the most suitable tool. We give some examples of such forecast products below.

5.1 Probability forecasts at fixed locations

Fig. 11 shows forecasts of 850 hPa temperature confidence interval for three gridpoints of the forecast from 3 January 1987, throughout the forecast range (15 days). At any forecast time the probability contours (99%, 90%, 70% and 50%) bound a range of temperature values. So, for example, 70% of the ensemble forecast temperatures fall within the range (or possibly ranges, see below) bounded by the 70% contours. The dispersion of the ensemble can therefore be judged by the spreading of the probability contours with forecast time. Superimposed is the control forecast (solid line) and the verifying analysis (dashed line). It should be remarked that in the calculation of these probability estimates, a Gaussian smoother was applied, for each forecast time, over the range of values at each grid point. The width of this smoother was such that the confidence interval in the first couple of days of the forecast is artificially broad.

Fig. 11a indicates a region in which the forecast dispersion is relatively small up to day 5. Here the forecaster for this particular region could predict with confidence the cooling trend, within a possible error bar of about three degrees. Beyond day 5 the ensemble clearly indicates a warming trend, though the magnitude of this trend is now much more uncertain. For this region, up to day 5, the most likely evolution according to the ensemble is close to the control, and also close to the verifying analysis. Between day 5 and day 9 it appears that the most likely evolution according to the ensemble is more skilful than the control.

Fig. 11b shows the forecast evolution at a second gridpoint. Up to day 6 the falling, rising and falling temperature trend of the control forecast is supported by the ensemble. At day 7 there is a bifurcation in the probability distribution. This bimodal distribution suggests that there is no single "most likely" forecast evolution, but two distinct almost equiprobable possibilities - a continuation of the cooling trend, and a return to warmer conditions. In this case the control forecast followed the path of one of the forks, and the verifying analysis took the other.

By definition, a probability forecast cannot be considered wrong if the verifying analysis lies in a region of low probability. An example of this is shown in Fig. 11c. Between day 6 and day 9 at this gridpoint, there was a strong cooling trend in the verifying analysis. Only one member of the ensemble captured this development well, and this is indicated by the 99% contour line dipping down at day 9. In practice a

forecaster would need to examine more recent forecasts to see whether the probability of this strong cooling increased or decreased.

Fig. 12 shows the estimated forecast PDF at various points at day 6 of the forecast from 3 January 1987. The control forecast (dark bar) and verifying analysis (light bar) are also shown. For cities such as Madrid or London (or an oceanic point such as the north Atlantic), the PDF is reasonably gaussian with relatively small standard deviation, and the ensemble supports the control. By contrast, for Oslo and Berlin, the distribution is multi-modal. For Oslo, the control forecast indicates a temperature of about -9 C, whilst the ensemble forecast indicates a maximum probability of -11 C, a distinct secondary maximum of -16 C, and a low probability of colder temperatures. The verifying temperature for Oslo was about -19 C (falling to -23C at day 7). For Berlin, there is a secondary maximum at -19 C which, at day 6, does not appear to verify. However, on the next day, Berlin temperatures dropped to -18 C. In this case the ensemble forecasts for Oslo and Berlin gave useful indications for the forecaster, absent in the control.

5.2 Probability maps

The ability to give some warning of possible extreme weather is clearly an important test of the ensemble. Fig. 13 shows a map giving the probability at day 7 that the 850 hPa temperature is either at least 10 K warmer than climatology, or at least 10 K colder than climatology. The anomaly of the control and the verifying analysis is also given. Note that whilst the control forecast missed the westward extension of the cold pool over parts of Scandinavia, the ensemble prediction indicates a 10% probability of extreme cold conditions for this area. As above the forecaster would have to wait for tomorrow's forecast to assess whether this small risk of extreme temperature was increasing or decreasing.

5.3 Probability forecasts of flow-dependent clusters

In this section we give examples of the application of clustering algorithms to the ensemble members, to produce a reduced set of possible forecast alternatives. Whilst probability forecasts based on this type of cluster analysis are likely to be more useful in an operational environment than those based on the climatological "Grosswetterlagen" clusters above, it is less straightforward to derive meaningful Brier Scores from them, since the climatological probabilities of the clusters are flow dependent.

In one approach we have applied the Ward hierarchical clustering technique over the European region (30-70N, 30W-30E) using 850 hPa temperature fields, reducing the fields to four clusters, which on average account for about 50% of the original variance within the ensemble.

The four clusters at day 5 of the forecast from 3 January 1987 are shown in Fig. 14, with the cluster population shown in brackets in the top right hand side of the cluster map. The rms error of the cluster is

shown at the bottom right side of each cluster map. In this case (not always true!) the error of the clusters increases as the population decreases, so that cluster 4, with the cold tongue, has just one member and largest rms error.

In this case, the synoptic difference between clusters 1 and 2 is not very great. Collectively this pattern is overwhelmingly more probable than the "northward-pushing warm tongue" of cluster 3, or the "south-westward plunging cold tongue" of cluster 4. The forecaster would be able to issue a forecast based on clusters 1 and 2 with reasonable confidence, though he should keep an eye on later forecasts for evidence that either cluster 3 or 4 was in fact developing.

Overall, at day 5, one of the four clusters was more skilful than the control forecast in 18 of the 24 cases. At day 7 one of the four clusters was more skilful than the control in 22 of the 24 cases. The histograms in Fig. 15 show the percentage of times a cluster is more skilful than the control, compared with the cluster probability (based on 6 clusters of 500 hPa height over Europe). For days 3 and 5 the relationship is clear, the more populated the cluster, the greater the chance it will be more skilful than the control, though for day 7 no relationship is apparent.

At every step in the Ward clustering algorithm, the increment in the squared distance summed over all individual members within each cluster, is as small as possible. (As such, the technique might well merge a cluster with small population into a cluster with large population, rather than merge together two clusters with large populations, even though the distance between the centroids of the large-population clusters was smaller). A second method has been studied (based on 500 hPa height fields) in which the increment in the maximum distance between elements in a cluster is as small as possible at every step. In this method the clustering algorithm is halted when this maximum distance exceeds 170 metres (approximately the distance between randomly chosen 500 hPa height fields). Moreover, in order to ensure consistency between different forecast days, the distance between two forecasts, used in the cluster algorithm, is taken to be the maximum distance from days 4 - 7.

An example of this method is shown in Fig. 16 using the ensemble from 15 April 1990. Here three separate clusters are identified. They are shown at days 5 and 7. Here the first two clusters have equal populations of 15 members, the third has a much smaller population. Compared with the verifying analyses for days 5 and 7, the second cluster is quite skilful with cut-off low at day 5, and tilted trough/ridge system over Europe at day 7. The cluster with just 3 members is noticeably less skilful.

6. FUTURE DEVELOPMENTS

6.1 Calculation of SVs in a primitive equation model

The calculation of the SVs in the QG model enabled the whole spectrum to be estimated by conventional matrix inversion algorithms. However, interpolation of the SVs into the full primitive equation model destroyed some of the phase relationship that generated optimal growth. Problems associated with interpolation were particularly severe in the vicinity of orography whose representation in the QG and primitive equation models is completely different.

Calculation of the primitive equation SVs by matrix methods is not practicable, largely because of computer memory requirements. However, we have obtained from the Numerical Algorithms Group, a pre-release copy of the Lanzcos algorithm for finding the dominant eigenfunctions of symmetric operators. The algorithm does not require the matrix elements of R^*R, but works by iteration on an initial random vector. The structure of the Lanczos algorithm is such that the extreme eigenvalues tend to emerge after only a few iterations (*Buizza*, 1992).

The Lanczos algorithm has been coded into the Integrated Forecasting System (IFS), and SVs of the T21L19 primitive equation have been calculated. Fig. 17 shows the amplification factor for the spectrum of growing SVs evaluated at the optimisation time, here 12, 24 and 36 hours. In order to avoid estimating non-meteorological SVs, the Lanzcos code was applied to N^*R^*RN where N is the normal mode initialisation operator (here applied to the first 5 normal modes in the vertical). It is interesting to note that there are over 60 SVs with an amplitude doubling time of 1 day or less.

Fig. 18 shows a number of streamfunction perturbations at about 500 hPa associated with SVs optimised for 24 hours, calculated for a trajectory from 17 January 1989. It can be seen that the structure of the SVs are qualitatively similar to those in the QG calculations shown in Fig. 1. Both calculations generate perturbations near the entrance of the Pacific and Atlantic storm tracks.

Fig. 19 illustrates a problem encountered with the primitive equation calculations. It shows the second SV at levels 17, 18 and 19 at time 0 and after 24 hours of integration with the forward tangent model. It can be seen that the perturbation amplitude is concentrated in a very thin layer near the surface of the model reversing in sign between levels 19 and 18. If these low-level structures are integrated in the full IFS model, they rapidly lose amplitude through damping by the boundary layer physics. Since the version of the tangent model used to calculate these SVs is adiabatic, they are not damped by the Lanczos calculation. Work is currently underway to introduce a boundary layer scheme into the tangent model.

On the basis of results obtained so far, it is likely that future ensemble prediction will be made using the IFS system both for the nonlinear trajectory calculations, and for the calculation of the SVs. It is hoped that the consistency obtained in the two calculations will lead to a superior forecast product. Full details of the SV calculations in the IFS system using the Lanczos algorithm are given in *Buizza* (1992).

6.2 Incorporation of analysis error covariance into SV calculation

In the experimentation described in the main body of this paper, the OI estimate of analysis error variance was used to determine the amplitude of the SV perturbations. With the transition to variational data assimilation, estimates of analysis error covariances are becoming available. It is possible to build into the eigenvector calculations, information about such covariances.

Let C be an estimate of analysis error covariance, then it is straightforward to show that this evolves at time t_1 to RCR^*. The semi-major axes of the ellipsoid at t_1 are given by eigenvectors of this product operator. At analysis time these are associated with the eigenvectors of CR^*R. The analysis can be formulated to give the SVs of R with initial normalisation constrained by C.

6.3 Product Development

We have given some examples of possible products in this paper. Many others, both for instantaneous and time-average states are possible. Probabilistic precipitation forecasts are examples of products being developed at present. Further work will refine the choice of products; for this feedback from Member State forecasters will be crucial.

One aspect we have not discussed in this paper is the time-evolution of the probability forecast estimates from ensemble forecasts initialised one day apart. One would hope that if a particular event was indicated with some small probability towards the end of the forecast range, then that probability would either increase or decrease fairly smoothly as more recent forecasts become available. The ability of the ensemble to avoid inconsistent predictions from forecasts initialised from consecutive days could well give it an important advantage over the conventional deterministic forecast. However, the extent to which this is true will have to be determined from future experiments.

6.4 Ensemble size

The number of growing phase-space directions indicated in Fig. 17 suggests that ensemble forecasts with more than a 150 members may be required to sample adequately the short-range forecast error PDF. As computer power increases, and numerical techniques become more efficient, it will therefore be necessary to consider whether it would be more beneficial to integrate the ensemble with a higher resolution model, or to increase the size of the ensemble. Whilst PDF estimates will in general become more reliable with

a larger ensemble size, estimates of the probability of extreme wind speeds and extreme precipitation rates may become more reliable with a higher resolution model.

6.5 Model resolution for SV calculations

If the SV calculations were performed with a higher resolution model, the enhanced number of degrees of freedom would probably increase the size of the growing spectrum, generating faster growing SVs with smaller initial scale. However, given the results of *Molteni and Palmer* (1992) (see also section 2 above), it is possible that such a high-resolution linear SV, optimised over a couple of days, say, would evolve with upscale energy transfer towards the synoptic patterns resolvable, at the end of the optimisation period, with the current T21 model. In practice, the growth of initial errors projecting onto the initial structure of such an SV would become nonlinear before the end of the optimisation time. From *Molteni and Palmer* (1992) these nonlinear self interactions also involve upscale energy transfer, which tends to damp the final SV amplitude compared with purely linear growth. As such the overall structure of the high resolution perturbation at the end of the two-day period, may be describable by the linear SV of a lower-resolution model.

These results suggest that the resolution of the model used for generating initial perturbations need not necessarily be finer than the scale necessary to represent forecast errors of interest. As such, for the medium-range, the T21L19 resolution may well be satisfactory. Further work is necessary to clarify these important issues.

6.6 Model development

Probability forecasts from ensemble predictions are especially susceptible to model systematic error. On the one hand, simple biases can be handled for by an *a posteriori* adjustment of probabilities. However, errors associated with failure to simulate certain types of flow regimes correctly are much more difficult to correct for. As such, development and testing of physical parametrizations will continue to be an essential component of research for ensemble forecasting.

7. A REAL-TIME EXPERIMENT

Starting winter 1992/93, an ensemble forecast will be run in real time. The SVs will be estimated using the Lanczos algorithm in the T21L19 IFS system. These will be projected onto a 6-hr forecast error covariance matrix, giving an amplitude for the individual SVs. Linear combinations of SVs will be used to delocalise the perturbations. It is anticipated that each ensemble will comprise between 30 and 40 integrations. The ensemble will be run for 10 days several times per week, with longer integrations from time to time.

8. REFERENCES

Anderberg, M.R., 1973: Cluster analysis for applications. Academic Press, New York.

Brankovic, C., T.N. Palmer, F. Molteni, S. Tibaldi and U. Cubasch, 1990: Extended-range predictions with ECMWF models: Time-lagged ensemble forecasting. Q.J.R.Meteorol.Soc., 116, 867-912.

Buizza, R., 1992: Unstable perturbations using the adjoint technique. ECMWF Technical Memorandum. In progress.

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

Farrell, B.F., 1989: Optimal excitation of baroclinic waves. J.Atmos.Sci., 46, 1193-1206.

Farrell, B.F., 1991: Stochastic perturbations in shear flow. Extended abstract. Proceedings of eighth conference of American Meteorological Society on atmospheric and oceanic waves and stability, pages 64-66. American Meteorological Society. Massachusetts pp 417.

Ferranti, L., F.Molteni, C.Brankovic and T.N.Palmer, 1992: The climate of ECMWF seasonal integrations. I: Extratropical circulations. In preparation.

Hollingsworth, A., 1980: An experiment in Monte Carlo forecasting procedure. ECMWF workshop on stochastic dynamic forecasting. ECMWF, 1980, 99pp.

Kalnay, E. and A. Dalcher, 1987: Forecasting forecast skill. Mon.Wea.Rev., 115, 349-356.

Lacarra, J.-F. and O.Talagrand, 1988: Short-range evolution of small perturbations in a barotropic model. Tellus, 40A, 81-95.

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

Lorenz, E.N., 1965: A study of the predictability of a 28-variable atmospheric model. Tellus, 17, 321-333.

Molteni, F. and T.N. Palmer, 1992: Predictability and non-modal finite-time instability of the northern winter circulation. (MP). Proceedings of ECMWF Workshop on New Developments in Predictability, 1992.

Molteni, F., S. Tibaldi and T.N. Palmer, 1990: Regimes in the wintertime circulation over northern extratropics. I: Observation evidence. Q.J.R.Meteor.Soc., 116, 31-67.

Mureau, R., F. Molteni, T.N. Palmer, 1992: Ensemble prediction using dynamically-conditioned perturbations. Proceedings of ECMWF Workshop on New Developments in Predictability, 1992.

Noble, B. and J.W. Daniel, 1977: Applied Linear Algebra. Prentice-Hall, Inc., USA, 477pp.

Palmer, T.N., 1988: Medium and extended-range predictability and stability of the Pacific/North American mode. Q.J.R.Meteorol.Soc., 114, 691-713.

Uhl, M.A., P.J. Smith, P.J., Lupo, A.R. and Zwack, P., 1992: The diagnosis of a pre-blocking explosively-developing extratropical cyclone system. Tellus, 44A, 236-251.

Wallace, J.M. and D.S. Gutzler, 1981: Teleconnections in the geopotential height field during the northern hemisphere winter. Mon.Wea.Rev., 109, 784-812.

APPENDIX 1

DEFINITION OF PERTURBATIONS AS LINEAR COMBINATIONS OF SVs

In order to define the amplitude and spatial structure of the N initial perturbations (in our ensembles, N = 32), we proceed as follows.

a) We select the N fastest growing SVs of the 12-h linear resolvent operator, eliminating those vectors which have more than 50% of their spatial variance over areas with latitude < 10°S or mean orographic height > 1000 m.

b) We assume that these N SVs are the leading eigenvectors of the covariance matrix $\mathbf{C}_0$ of the initial error at the QG-model levels, and neglect the variance explained by the other eigenvectors. In this way, we can write $\mathbf{C}_0$ as:

$$\mathbf{C}_0 = \mathbf{E}\,\boldsymbol{\Sigma}\,\mathbf{E}^t \qquad \text{A.1}$$

where $\mathbf{E}$ is the matrix which has the N normalised SVs $\mathbf{E}_i$ as columns, $\mathbf{E}^t$ its transpose (equal to its inverse) and $\boldsymbol{\Sigma}$ the diagonal matrix of the variances σ_i^2 explained by the SVs.

c) We assign appropriate values to the variances σ_i^2 using mean-square-error fields obtained from optimum-interpolation estimates and very-short-range forecast errors (see below for more details).

d) We construct a set of N perturbations (which may have either positive or negative sign) in such a way that the covariance matrix generated by the perturbations coincides with $\mathbf{C}_0$. If $\mathbf{P}$ is the matrix having the perturbations as columns, then it must be:

$$N^{-1}\,\mathbf{P}\,\mathbf{P}^t = \mathbf{C}_0 \qquad \text{A.2}$$

If we define the diagonal matrix $\mathbf{W}$ with elements:

$$w_i = \sqrt{N}\,\sigma_i \qquad \text{A.3}$$

from Eqs. A.1 and A.2 we have:

$$\mathbf{P}\,\mathbf{P}^t = \mathbf{E}\,\mathbf{W}\,\mathbf{W}^t\,\mathbf{E}^t \qquad \text{A.4}$$

Eq. A.4 is satisfied for any choice of

$$\mathbf{P} = \mathbf{E}\,\mathbf{W}\,\mathbf{R} \qquad \text{A.5}$$

where $\mathbf{R}$ is an orthogonal rotation matrix such that $\mathbf{R}\,\mathbf{R}^t = \mathbf{I}$. Since the 12-h SVs have spatially localised patterns, we can use the rotation matrix $\mathbf{R}$ to obtain perturbations with a more homogeneous spatial structure over the northern hemisphere.

In summary, once the SVs are computed from the linear resolvent operator, the final perturbations P_i depend on the values assigned to the error variances σ_i^2 and on the definition of the rotation matrix $\mathbf{R}$. We shall now discuss these points in more detail.

In order to define the σ_i^2, firstly we compute an upper bound for the error variance of 500 hPa streamfunction in physical space $\underline{x}$:

$$V(\underline{x}) = \max\,[\,OI(\underline{x}), FE(\underline{x})\,] \qquad \text{A.6}$$

where OI is an estimate produced by the optimum-interpolation analysis scheme, and FE is the mean-square error of 12-h forecasts in the 5 days preceding the initial date of a given ensemble forecast. From Eq. A.1, the variance in physical space corresponding to the diagonal of the $\mathbf{C}_0$ matrix is given by:

$$V_0(\underline{x}) = \Sigma_i\,\sigma_i^2\,E_i^2(\underline{x}) \qquad \text{A.7}$$

Appropriate values for σ_i^2 can be chosen so that $V_0(\underline{x})$ is as close as possible to $V(\underline{x})$.

In theory, this problem can be easily solved by least-square techniques. However, given the truncation of the SV series to only N = 32 terms and the uncertainties in the estimate of $V(\underline{x})$, direct least-square estimates of the variances often provided unrealistic values for some of the SVs. Therefore, a more empirical approach has been chosen, which is based on the concept of equipartition of variance between SV that cover overlapping areas. To do so, we fix an amplitude threshold ε, and define for each SV the localisation function:

$$\begin{aligned} H_i(\underline{x}) &= 1 \quad \text{if} \quad E_i^2(\underline{x}) > \varepsilon^2 \\ &= 0 \quad \text{otherwise} \end{aligned} \qquad \text{A.8}$$

From these functions, we compute an overlap factor which, for each point, counts the number of SVs with significant amplitude:

$$O(\underline{x}) = \Sigma_i H_i(\underline{x}) \qquad \text{A.9}$$

For each SV, a 'local' average of any spatial function $f(\underline{x})$ can be defined as:

$$< f(\underline{x}) >_i = \int H_i(\underline{x})\, f(\underline{x})\, d\underline{x} \,/\, \int H_i(\underline{x})\, d\underline{x} \qquad \text{A.10}$$

and a realistic value for σ_i is computed as:

$$\sigma_i = < |E_i| \sqrt{V(\underline{x})/N_i} >_i \,/\, < E_i^2(\underline{x}) >_i \qquad \text{A.11}$$

where

$$N_i = \max (< O(\underline{x}) >_i, 4) \qquad \text{A.12}$$

is an estimate of the number of perturbations covering the same area as E_i, which takes into account the truncation of the SV series.

Finally, the rotation matrix is defined by requiring that each perturbation P_i is a linear combination of up to 8 SVs. Since SVs with very similar eigenvalues often cover the same area with a phase shift of about 1/4 of the dominant wavelength, we divided the 32 SVs in 4 submatrices $\mathbf{E}_k$ as follows:

$$\mathbf{E}_k = \{ E_k, E_{k+4}, E_{k+8}, \ldots, E_{k+28} \}, \qquad k = 1,..,4 \qquad \text{A.13}$$

Then we computed the submatrices $\mathbf{P}_k$ of the final perturbations as:

$$\mathbf{P}_k = \mathbf{E}_k \mathbf{W}_k \mathbf{R}_8 \qquad \text{A.14}$$

where the rows of the 8x8 rotation matrix $\mathbf{R}_8$ are given by discretised orthogonal trigonometric functions of wavenumber 0 to 4.

Table 1: Flow characteristics of the 24 selected cases. Listed are the flow types at day 0 and day 5 over 2 areas (m: meridional flow; z: zonal flow; tr: transient flow)

		Atlantic/Europe		Pacific/USA	
870103	d0	m	(ridge)	m	(weak ridge)
	d5	m	(block at day 9)	m	(ridge)
870122	d0	m	(block)	m	
	d10	z	(transition)	z	
870225	d0	m		m	(ridge)
	d5	m		m	(block, transition)
880214	d0	m	(ridge)	z	
	d5	m	(ridge, varying)	m	(ridge, transition)
880306	d0	m	(ridge)	m	(ridge)
	d7	z	(transition)	m	(ridge)
881111	d0	z		z	
	d5	m	(trans. to ridge)	z	
881202	d0	m	(block over Europe)	z	
	d5	m	(trough over Europe)	m	(ridge over USA)
890117	d0	z		z	
	d5	tr		z	
890127	d0	m	(ridge over Europe)	m	(block over USA)
	d5	m	(ridge over Europe)	m	(ridge over PAC)
890205	d0	z		m	(block)
	d5	m	(block, transition)	m	(block)
890301	d0	z		m	(block)
	d5	m	(ridge, transition)	m	(block)
890325	d0	m	(ridge)	m	
	d5	m	(amplifying ridge)	m	
890425	d0	z	(trough over Europe)	m	
	d5	tr	(cutoff low, then m)	m	
891110	d0	m	(trough)	z	
	d5	m	(block, transition)	z/m	(weak ridge)
891204	d0	m	(block)	z	(weak ridge)
	d5	m	(weak ridge)	m	(intense ridge)
900117	d0	z		m	(ridge)
	d5	m	(ridge, cutoff low)	z	
900311	d0	m	(weak ridge)	m	(trough)
	d5	z	(weaker ridge)	m	(wavenumber 7)
900145	d0	z		m	
	d5	m	(cutoff low)	z	
901015	d0	m	(trough/ridge)	z	
	d5	m	(block)	z	
901027	d0	m	(block over Europe)	m	
	d5	m	(ridge over Atlantic)	m	
901213	d0	m	(ridge/trough)	m	(weak ridge)
	d5	m	(block)	m	(intense ridge)
910106	d0	z	(block over Europe)	z	
	d5	m	(block)	z	
910117	d0	m	(intense ridge)	m	(ridge)
	d5	m	(block)	m	(block)
910219	d0	m	(cutoff low)	m	(intense ridge)
	d5	m	(ridge, transition)	m	(block)

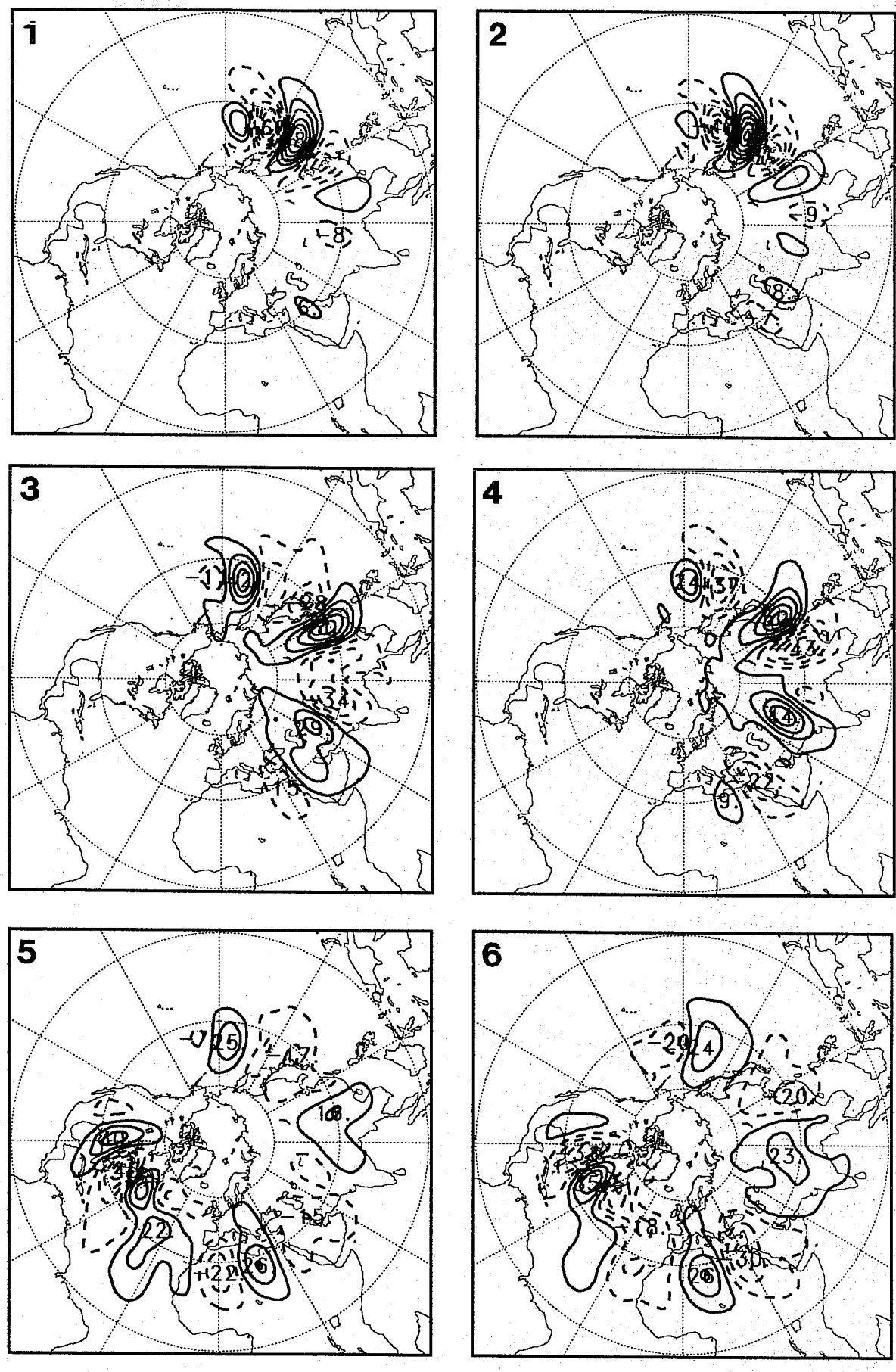


Fig. 1 Example of 500 hPa height perturbations from 12-hour singular vectors (numbers 1,2,3,4,9,10) calculated from the Quasi-Geostrophic model from data for 17 January 1989. Contour interval 10m.

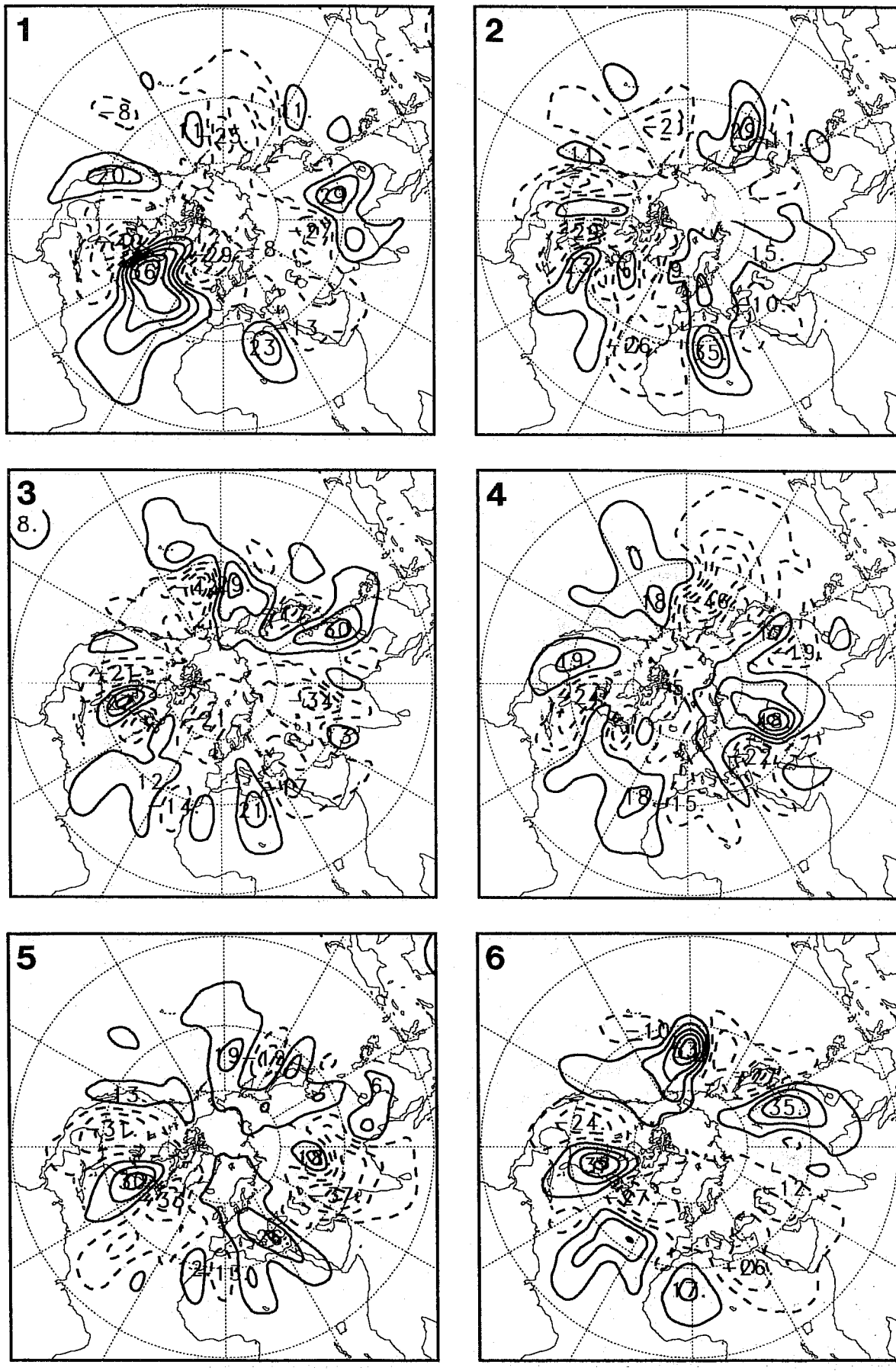


Fig. 2 Example of initial perturbations in 500 hPa height calculated from the singular vectors for 17 January 1989. Contour interval 10m.

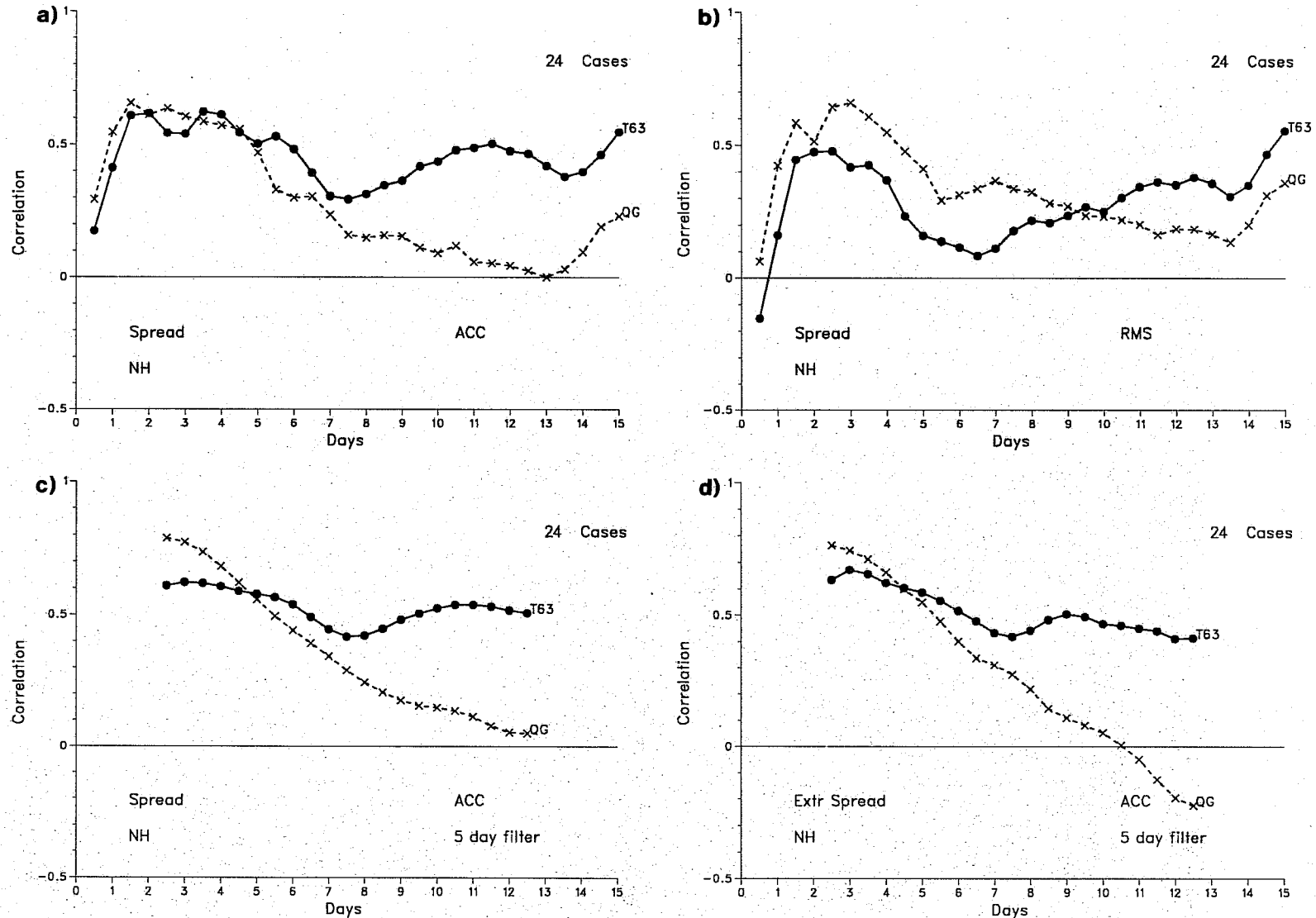


Fig. 3 Correlation between the skill of the control forecast and the spread of the ensemble, calculated from 500 hPa height over the Northern Hemisphere. The correlation is an average over 24 ensembles. a) using an rms measure difference, b) using anomaly correlation measure of difference, c) anomaly correlation measure from five-day mean fields, d) envelope spread from five-day mean fields.
Dotted line = T63 model, dashed line = QG model.

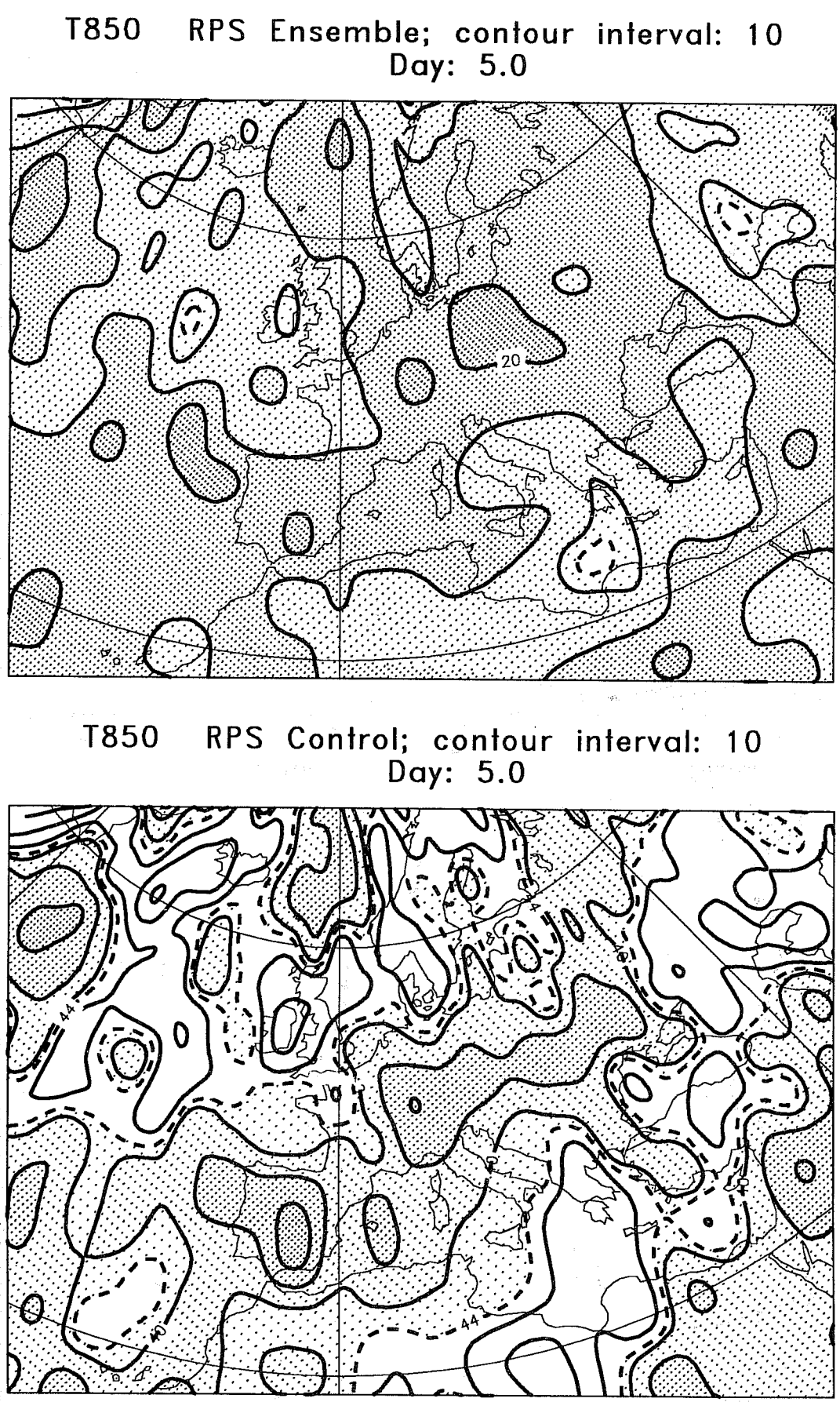


Fig. 4 Spatial distribution of Ranked Probability Scores for day 5. Based on three equiprobable categories for 850 hPa temperature. Top diagram ensemble forecast. Bottom diagram control forecast. RPS contours 10, 20, 30 ... shown as solid lines. RPS contour of 44, associated with the skill of a climatological probability forecast, shown as dashed line. Forecast is superior to climatology where shaded. The heavier the shading the more skilful the forecast.

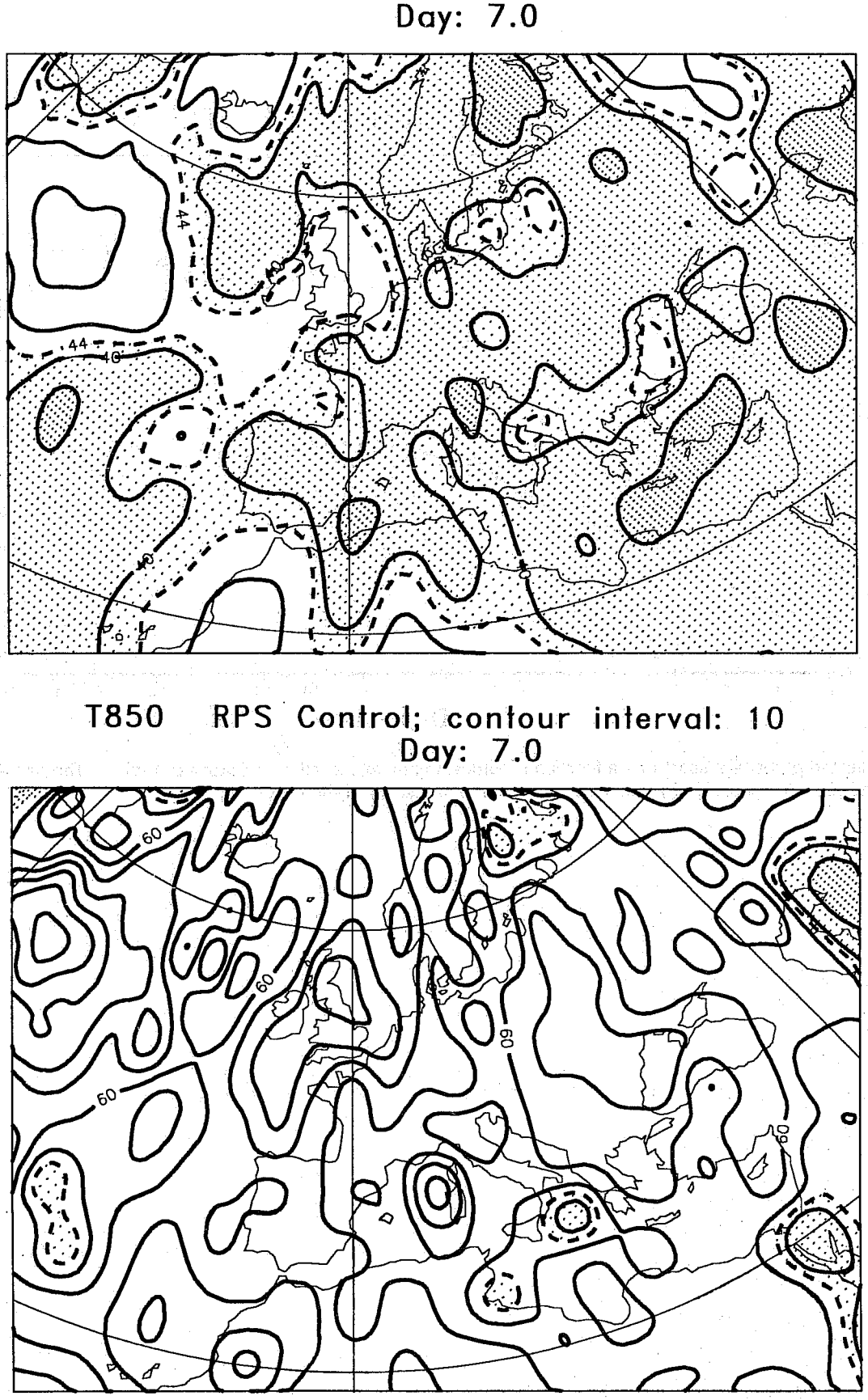


Fig. 5 As Fig. 4 but for day 7.

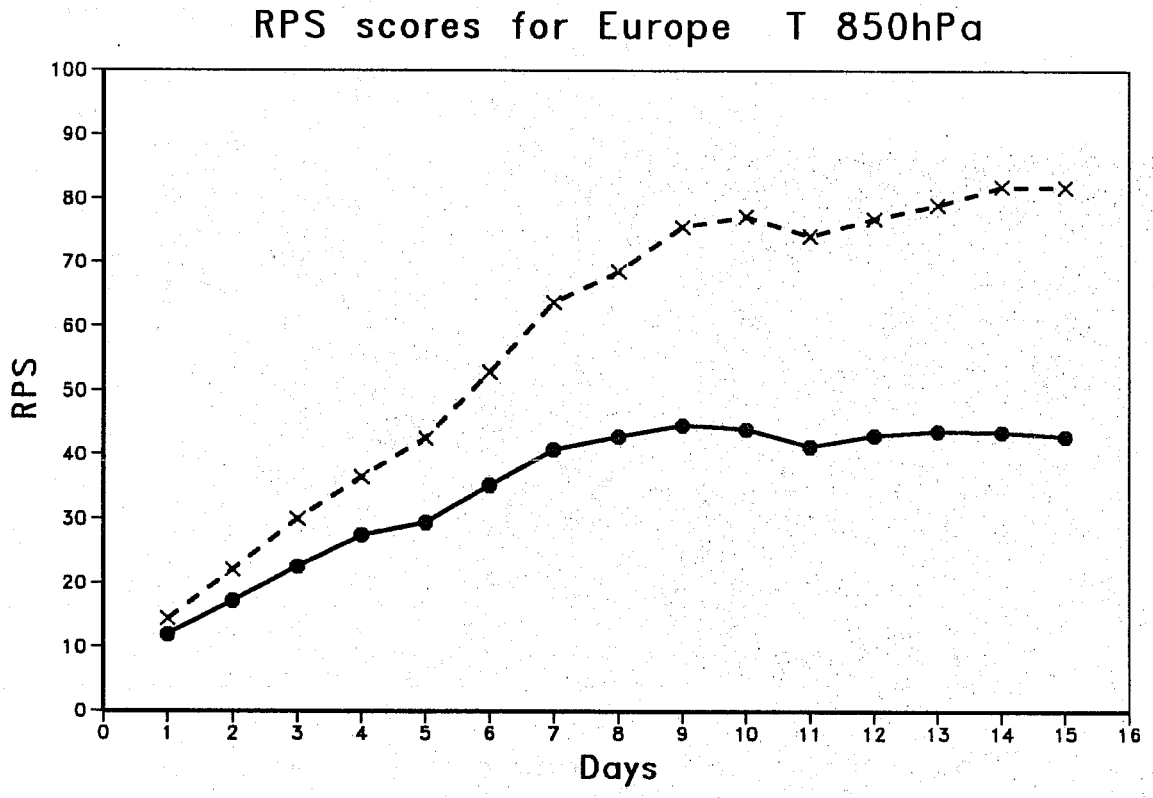


Fig. 6 Ranked probability scores as a function of forecast time, averaged over European region. The dashed curve indicates the score of the control, the solid curve the score of the ensemble.

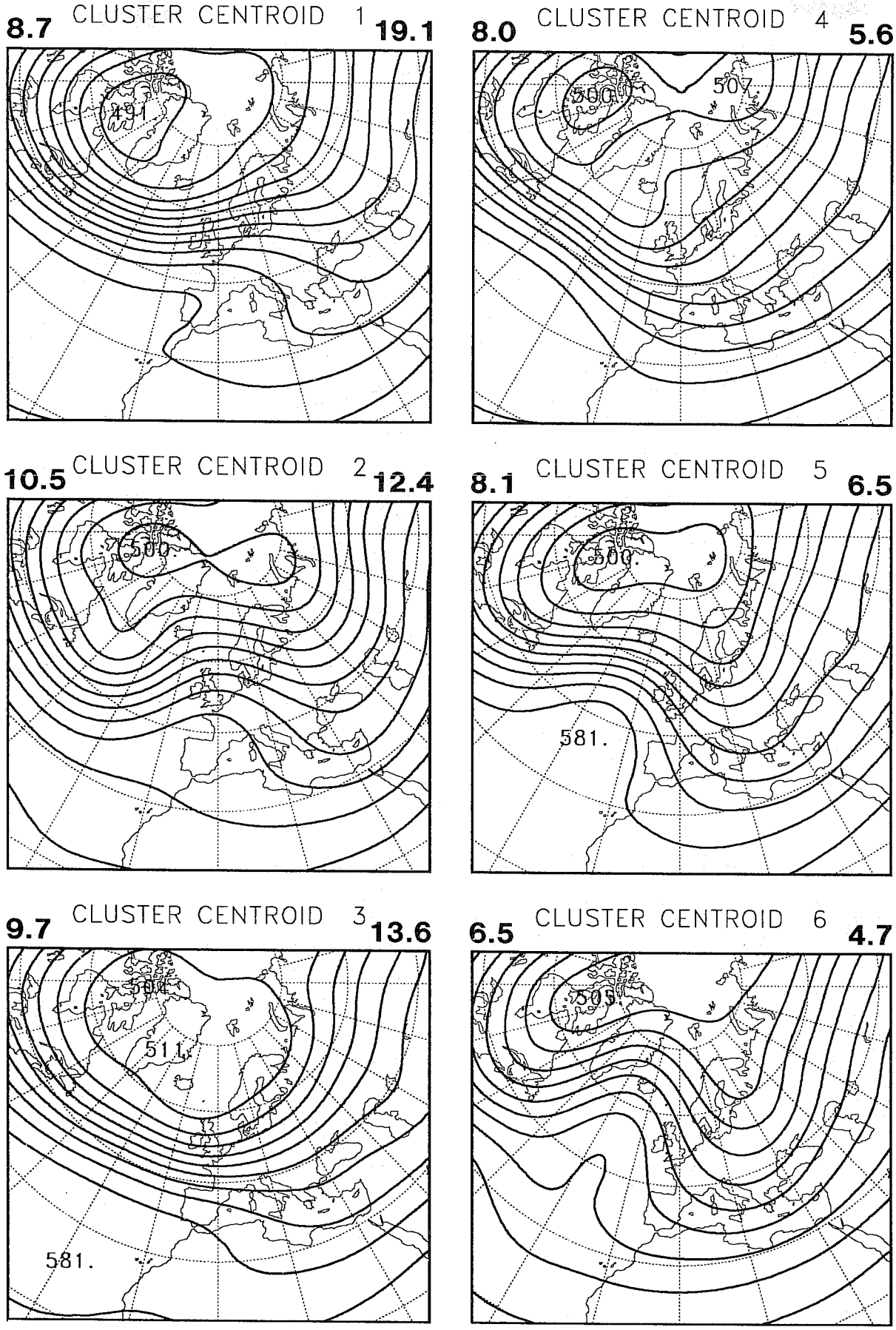


Fig. 7 Clusters of 500 hPa height calculated from 12 winters of data. Above each panel is the observed frequency (top left) and simulated frequency from a set of 120-day integrations with the T63 model (top right).

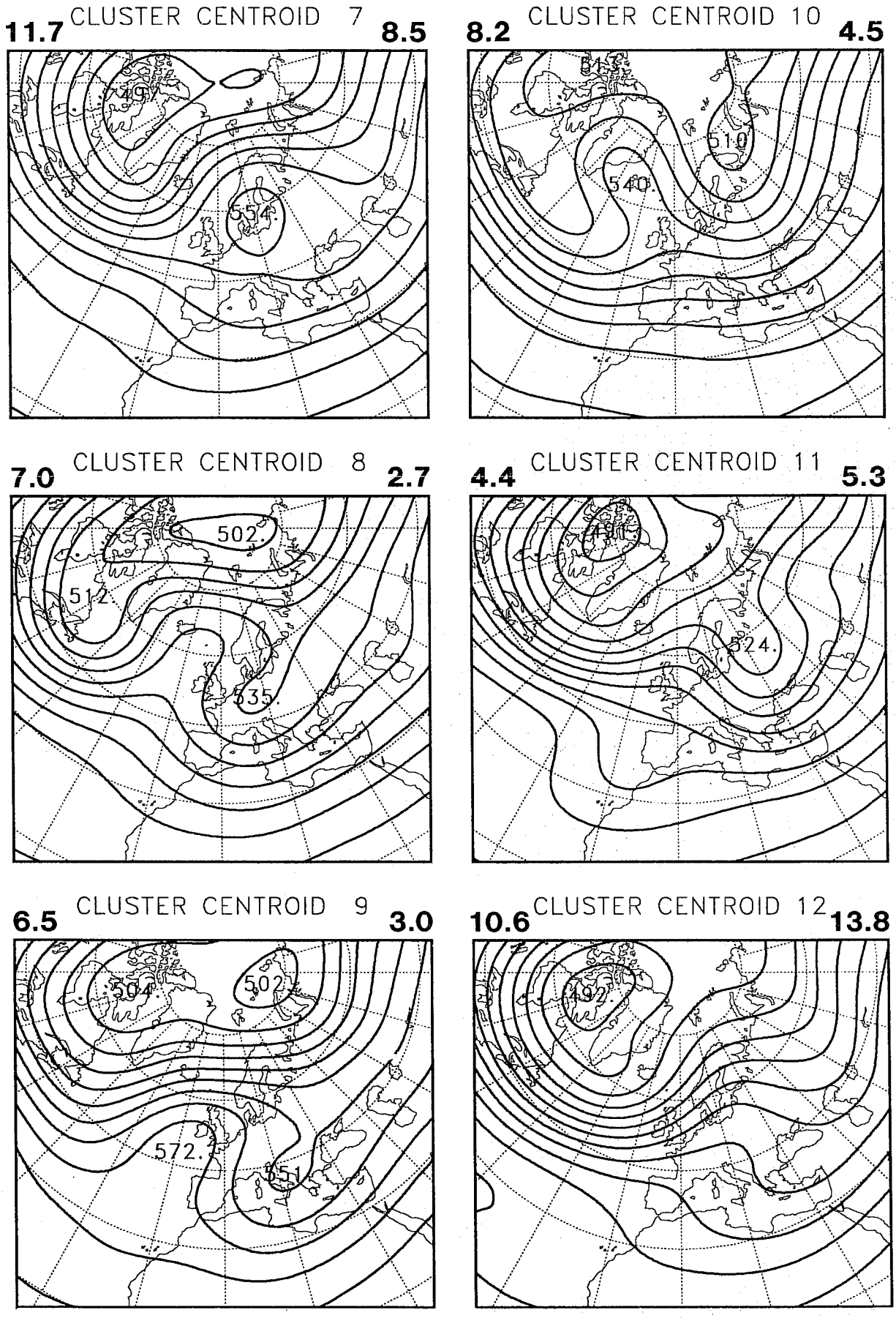


Fig. 7 **continued....**

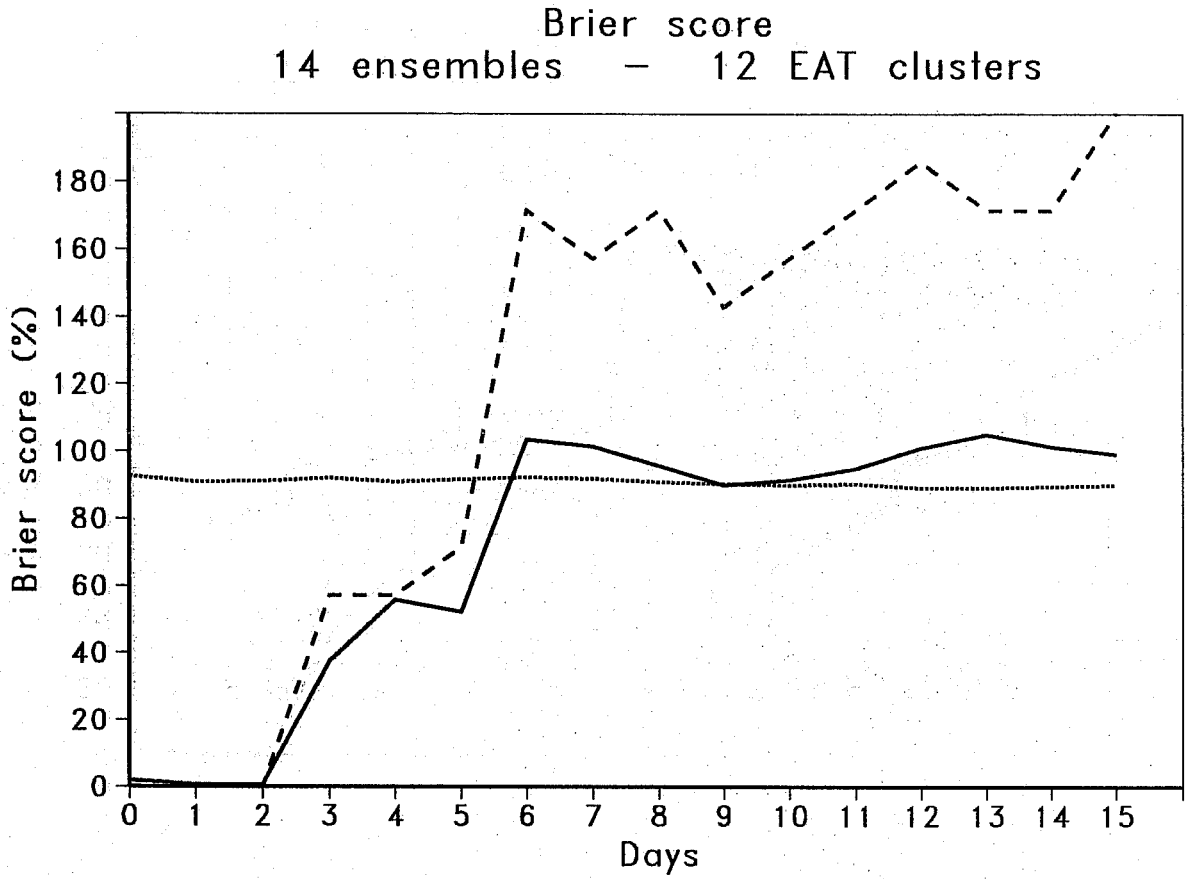


Fig. 8 Brier score of control forecast (dashed), ensemble forecast (solid) and climatology (dotted) from clusters shown in Fig. 7.

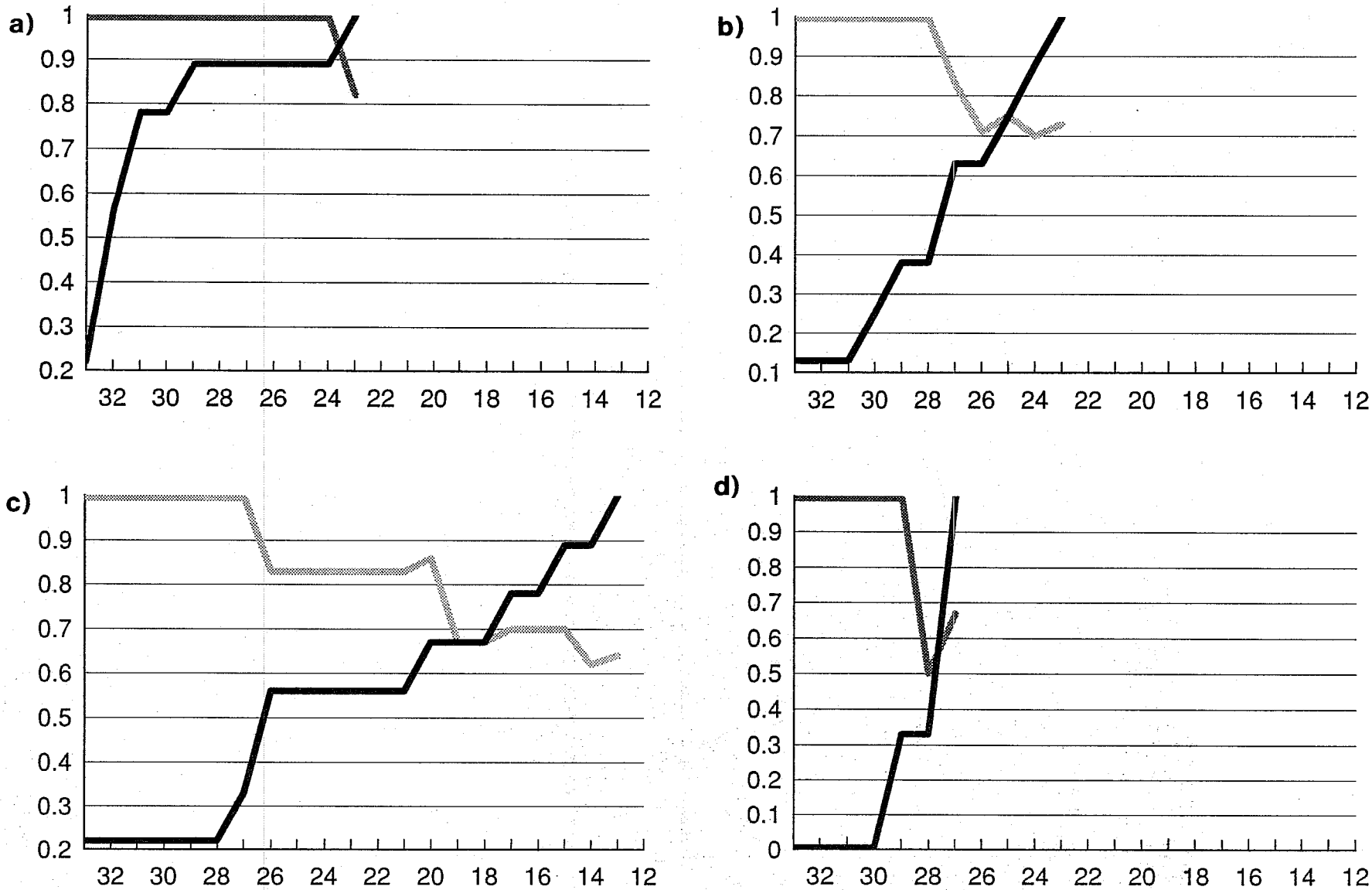


Fig. 9 Assessment of clustering based on the hypothetical situation that a forecast of the most populated cluster is "issued" providing its population exceeds some threshold. The light line shows the percentage of those issued forecasts that are correct, as a function of threshold. The heavy line shows the percentage of correct issued forecasts amongst the total sample of correct forecasts in all ensembles, also as a function of threshold. a) day 3, b) day 4, c) day 5, d) day 6.

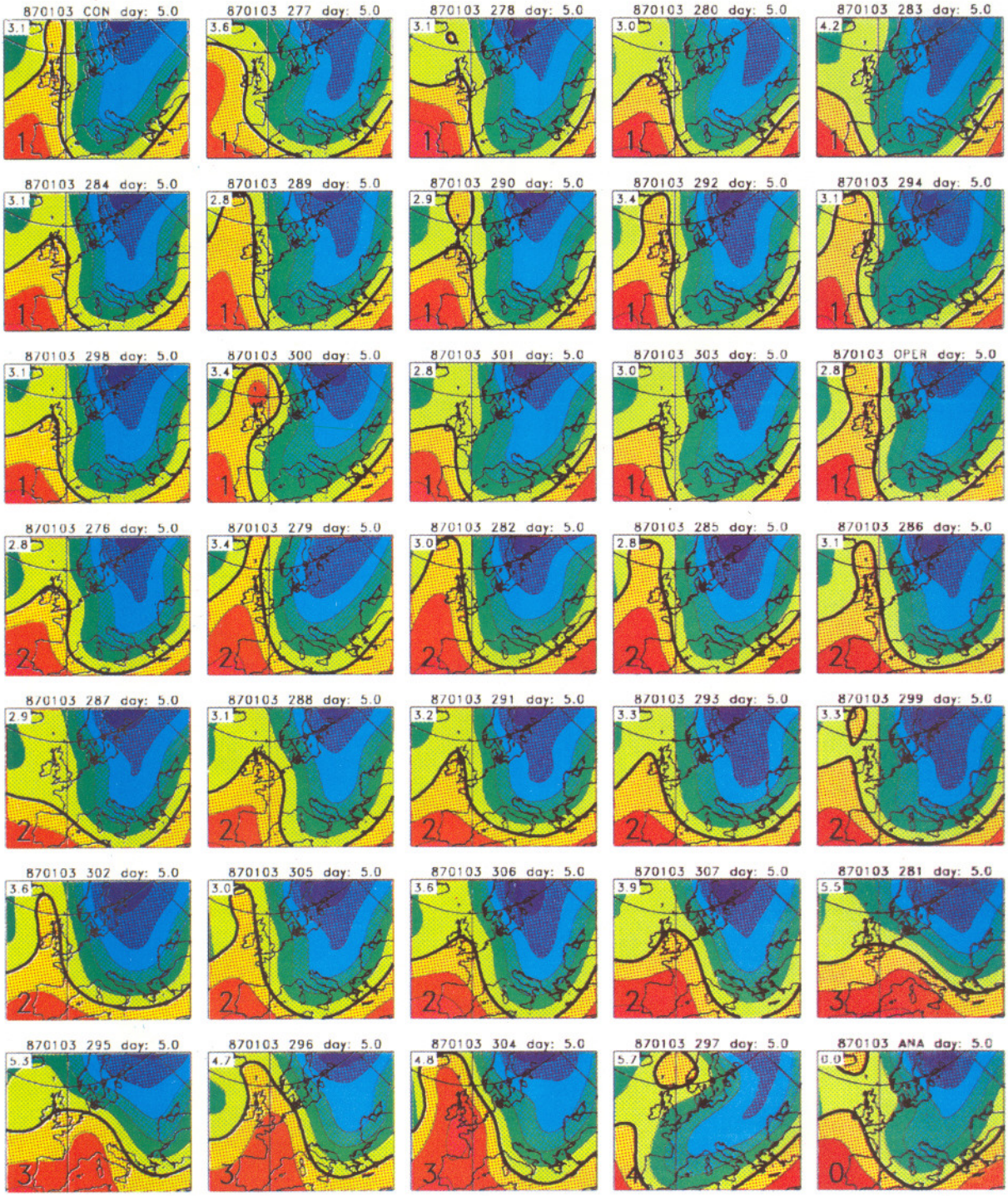


Fig. 10 850 hPa temperature from an ensemble of T63L19 integrations at day 5, initialised on date of 3 January 1987. The black contour is 0C, and the (colour) contour interval is 5C.

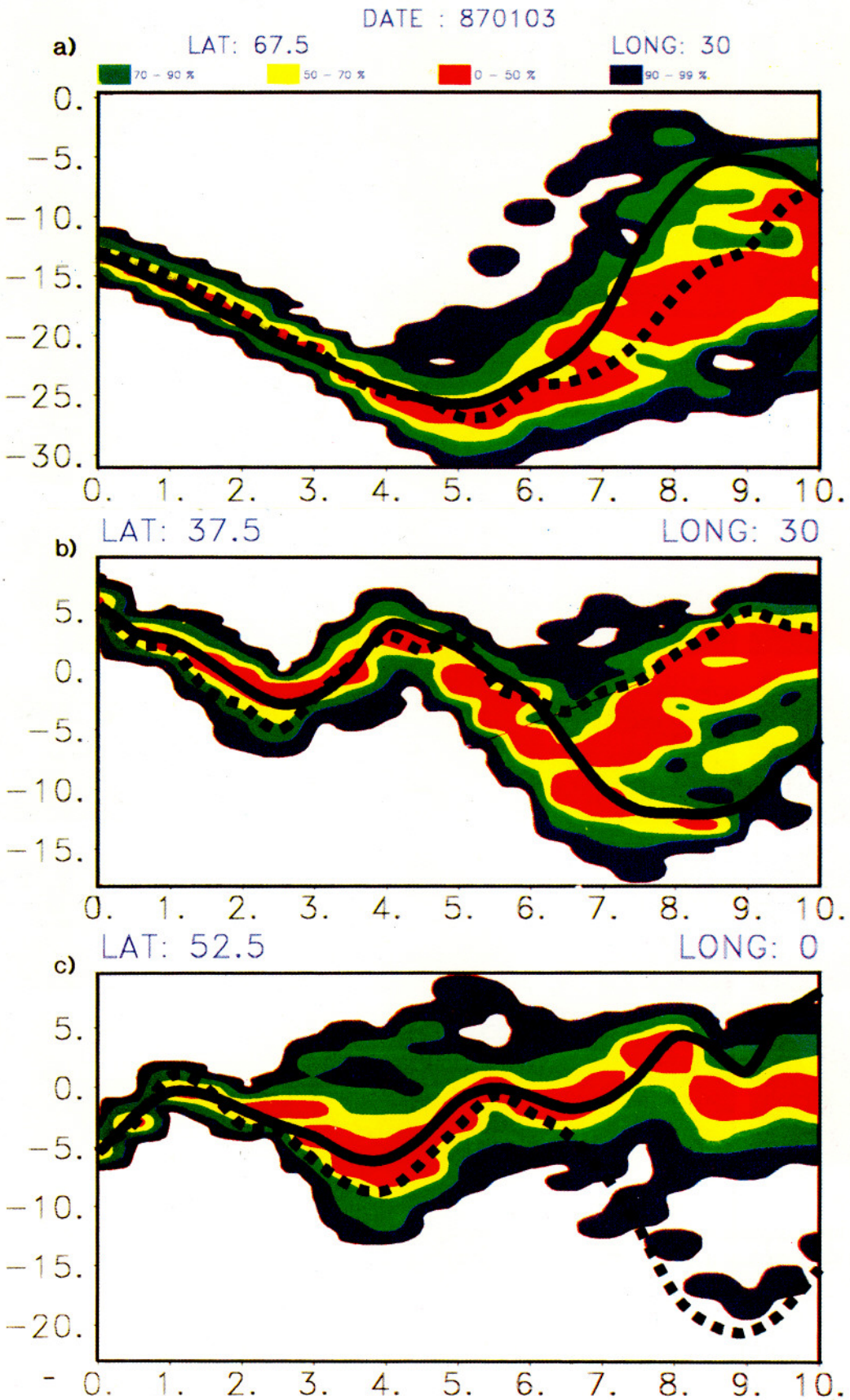


Fig. 11 a-c. Confidence intervals for the ensemble forecast from 3 January 1987 of 850 hPa temperature. Shown at three grid points throughout the forecast range. Contours shown are 99%, 90%, 70% and 50%.

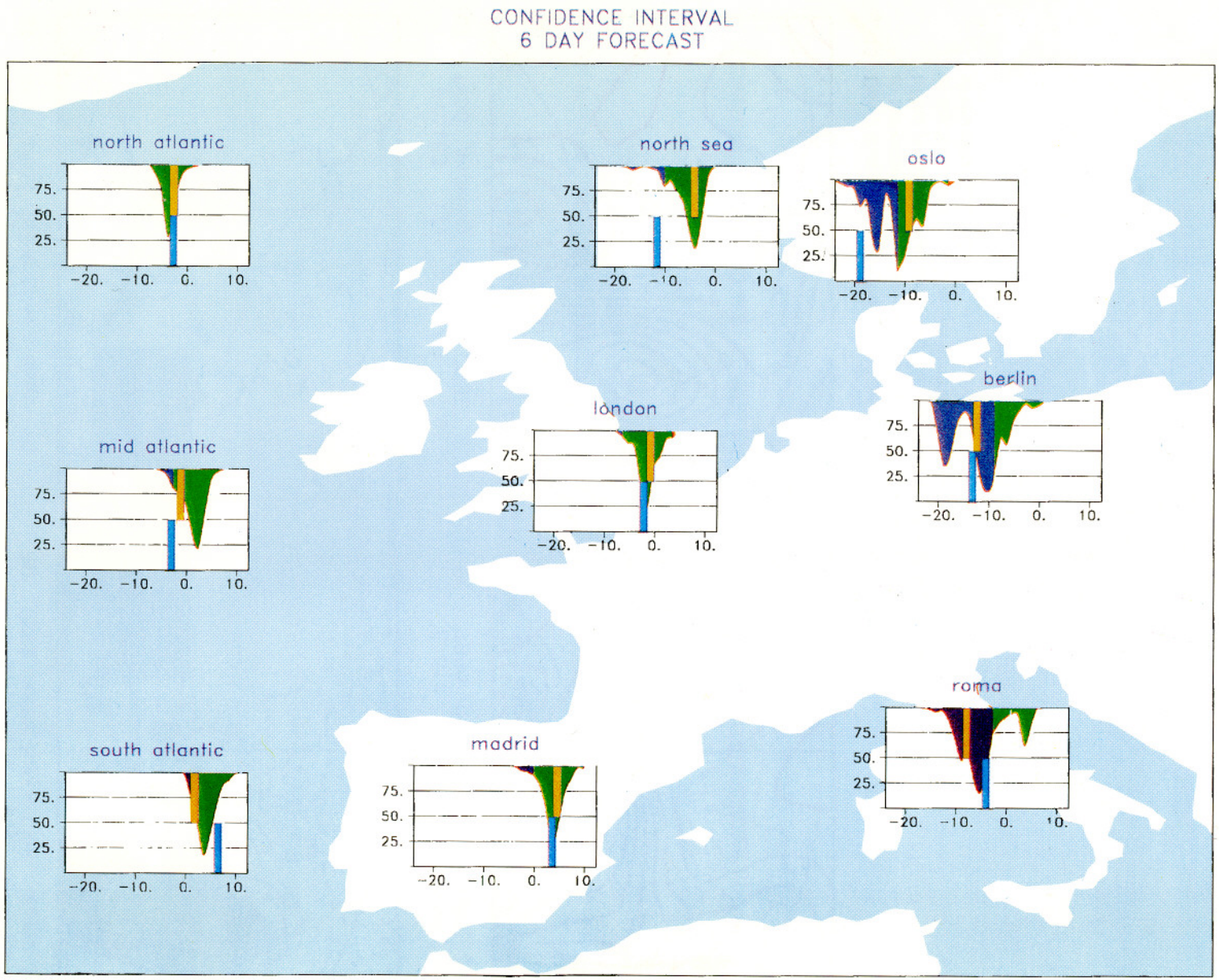


Fig. 12 Confidence intervals for day 6 of the ensemble forecast from 3 January 1987 of 850 hPa temperature at a variety of cities and oceanic points. The control forecast (dark bar) and verifying analysis (light bar) are also shown.

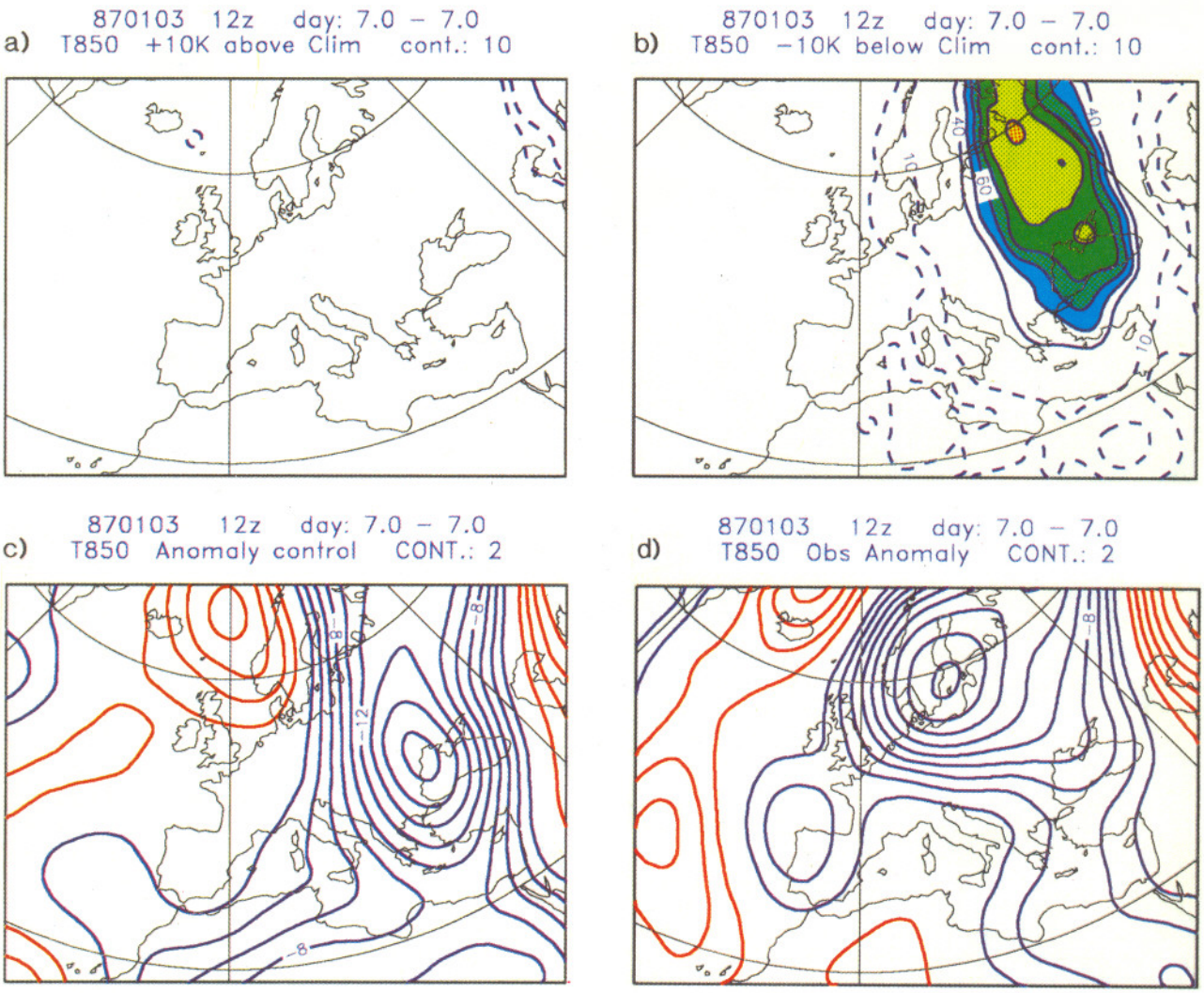


Fig. 13 Day 7 forecast of 850 hPa temperature from 3 January 1987.
a) the probability that temperatures are at least 10K above climatology
b) the probability that temperatures are at least 10K below climatology
c) the forecast anomaly of the control
d) the observed anomaly

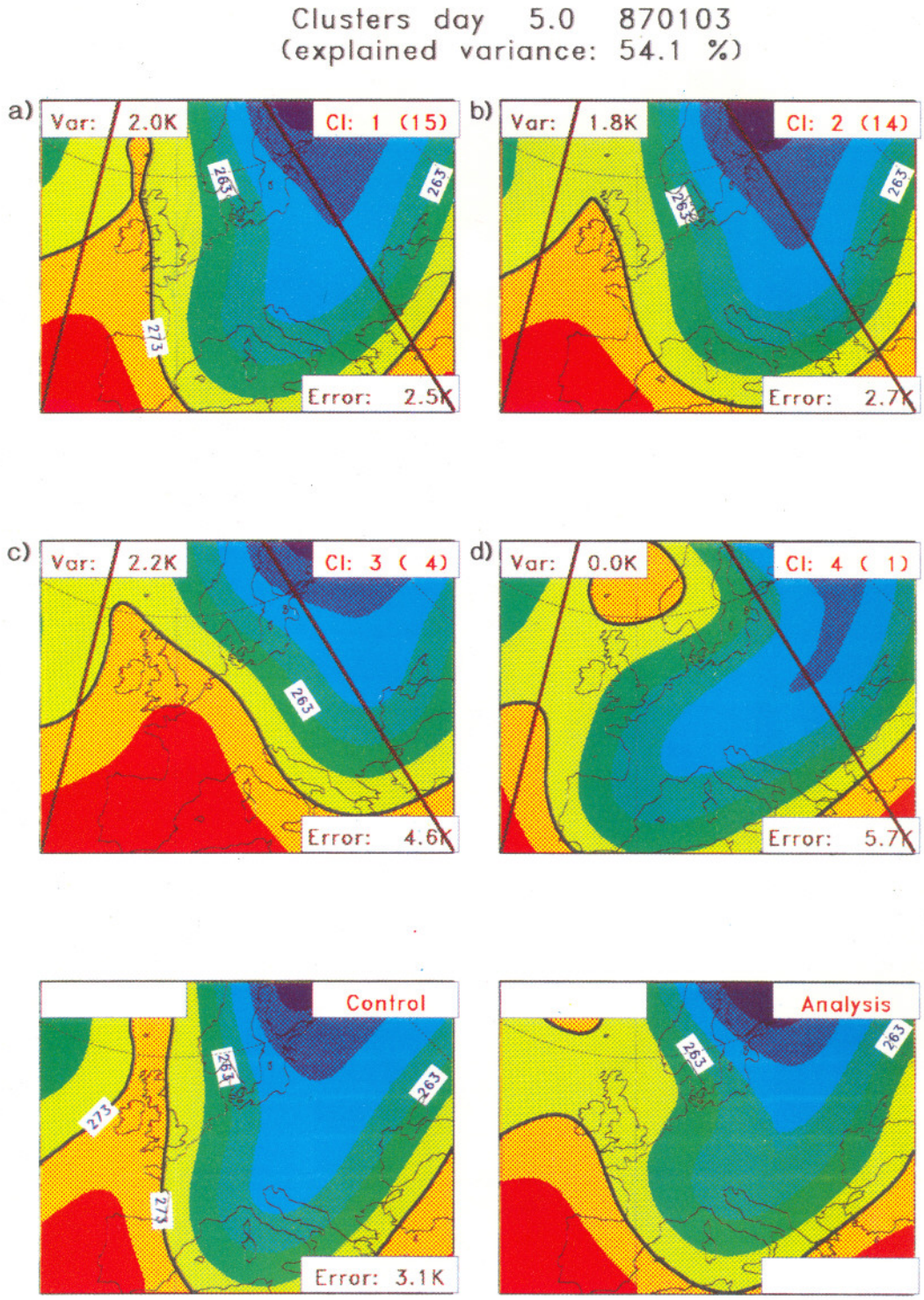


Fig. 14 a-d Four day 5 clusters of 850 hPa temperature defined by reducing the variance of the ensemble shown in Fig. 10 using the Ward hierarchical clustering algorithm. The probability associated with each cluster can be obtained from the clusfer density, respectively 15/32, 14/32, 4/32 and 1/32.

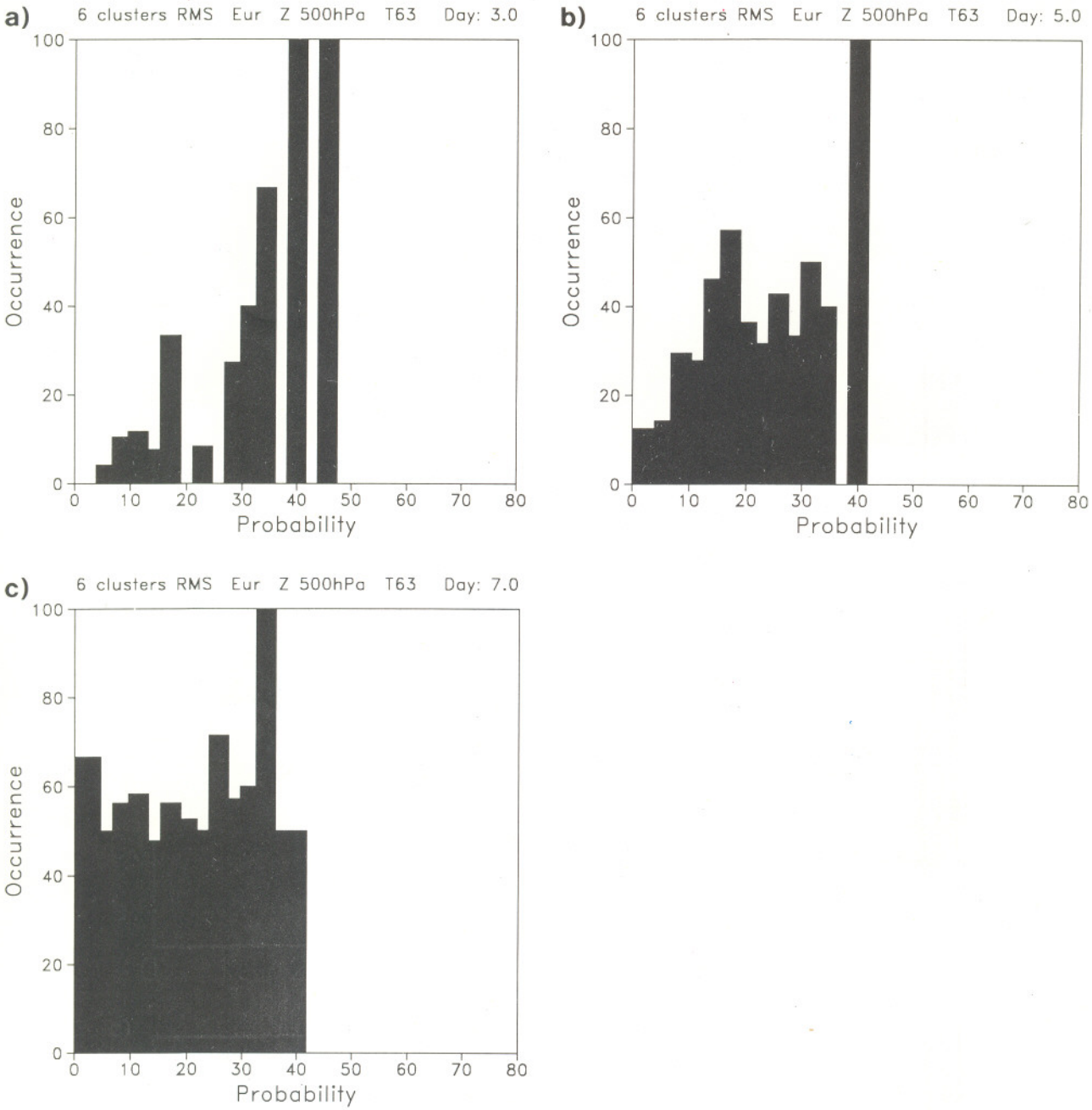


Fig. 15 Histograms showing the number of occasions that a cluster was more skilful than the control, as a function of the fractional population of the cluster: a) day 3, b) day 5, c) day 7.

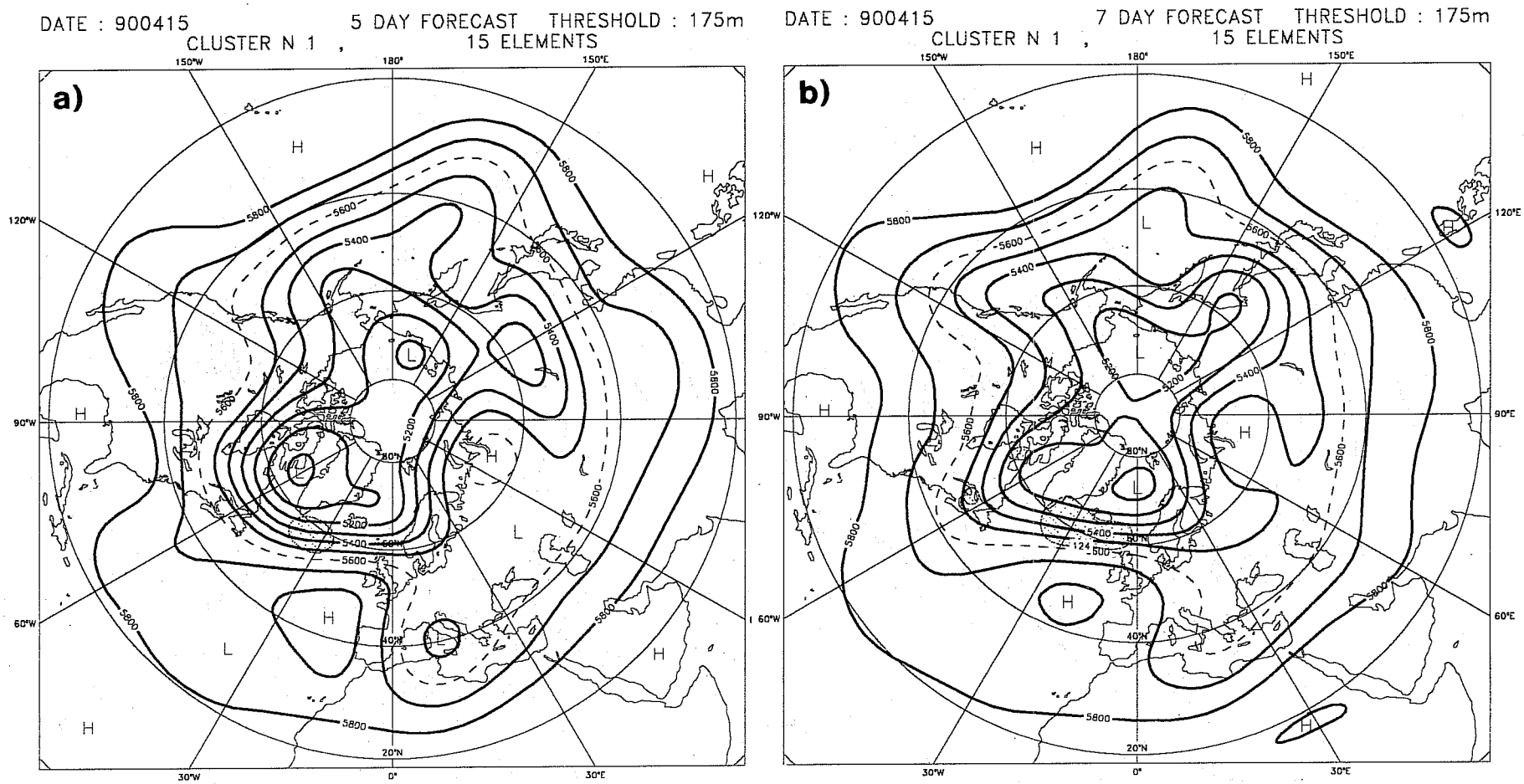


Fig. 16 Evolution of 500 hPa height clusters from day 5 and 7 based on forecast from 15 April 1990. The clustering was based on the maximum distance between forecasts in the range 4 to 7.
a)-b) cluster 1 (15 elements). c)-d) cluster 2 (15 elements). e)-f) cluster 3 (3 elements). g)-h) verifying analysis.
1st panel on each page is for day 5, 2nd panel day 7.

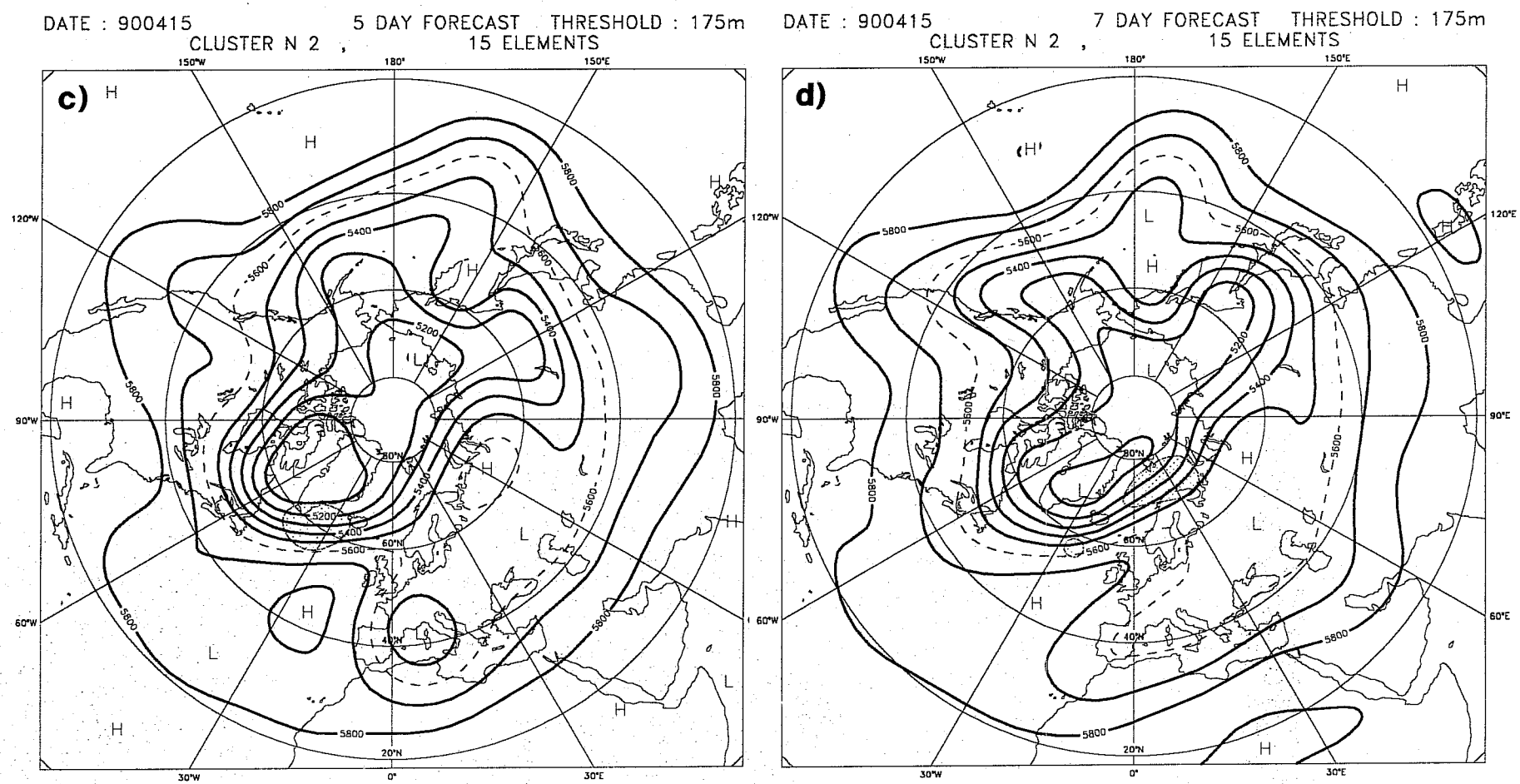


Fig. 16 continued....

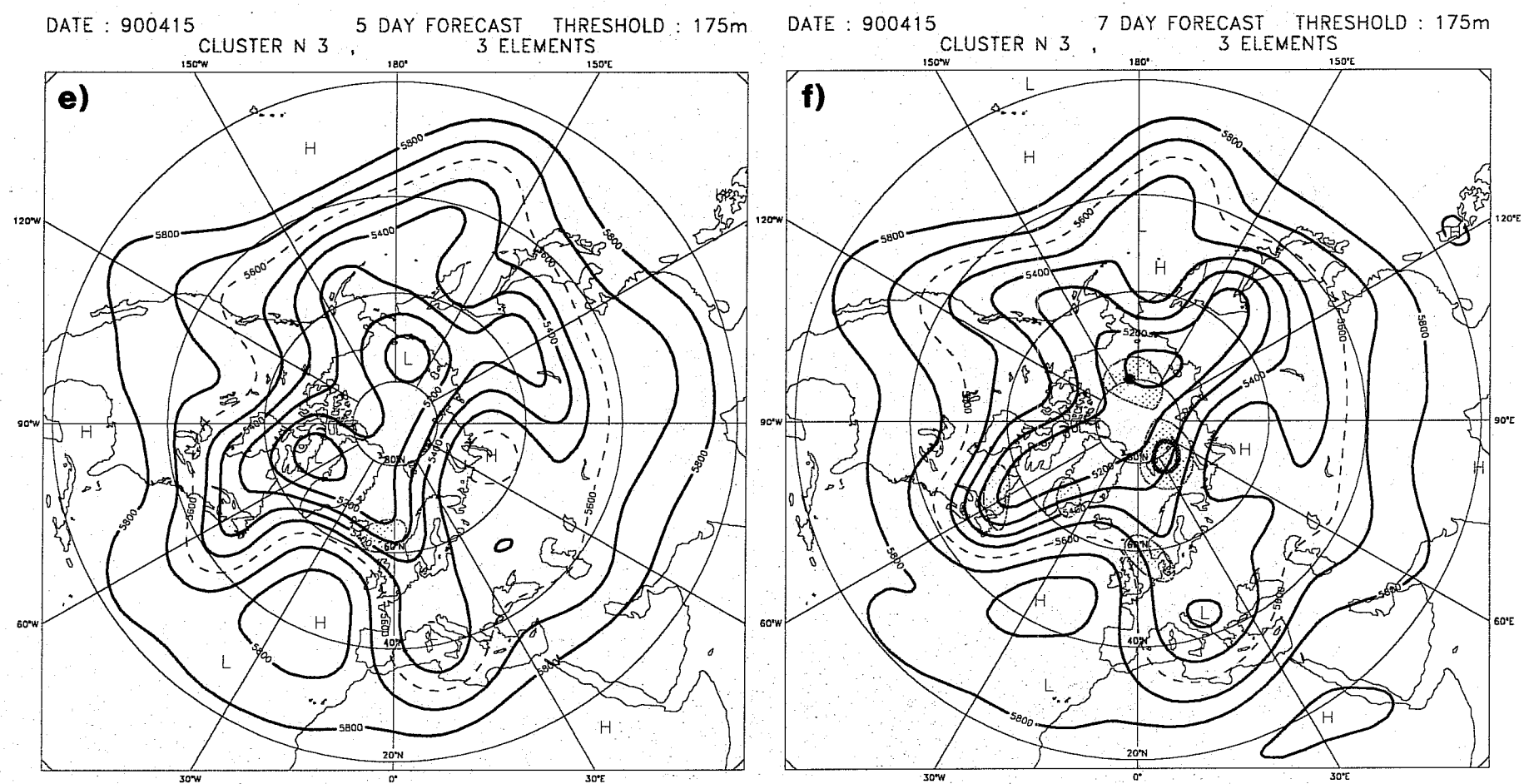
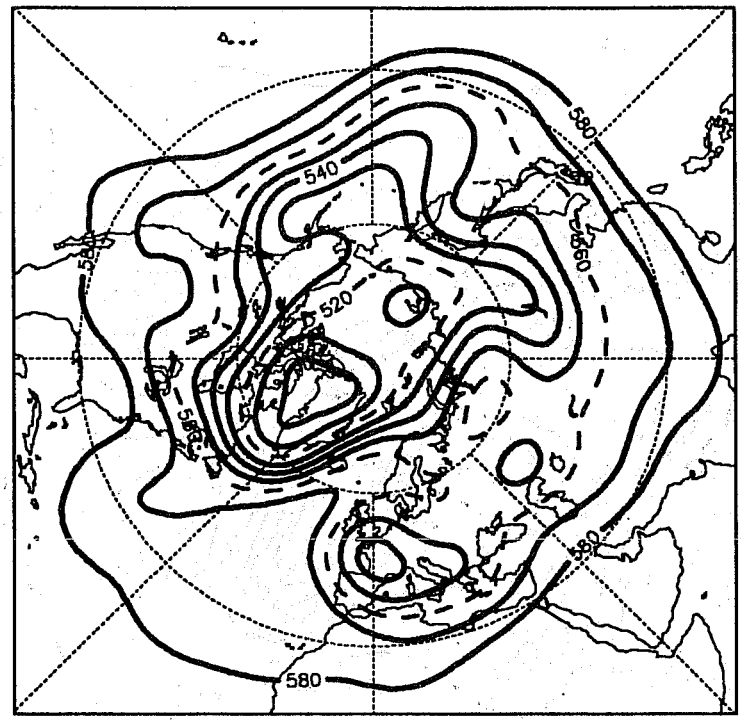


Fig. 16 continued....

900415 900420 12Z Analysis day: 5.0
Z cont. int.: 10

g)



900415 900422 12Z Analysis day: 7.0
Z cont. int.: 10

h)

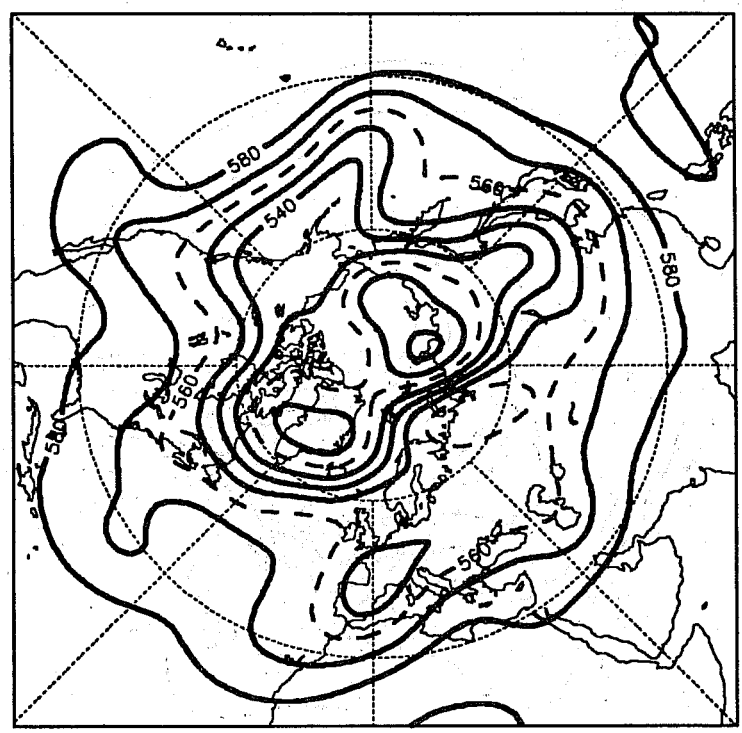


Fig. 16 continued....

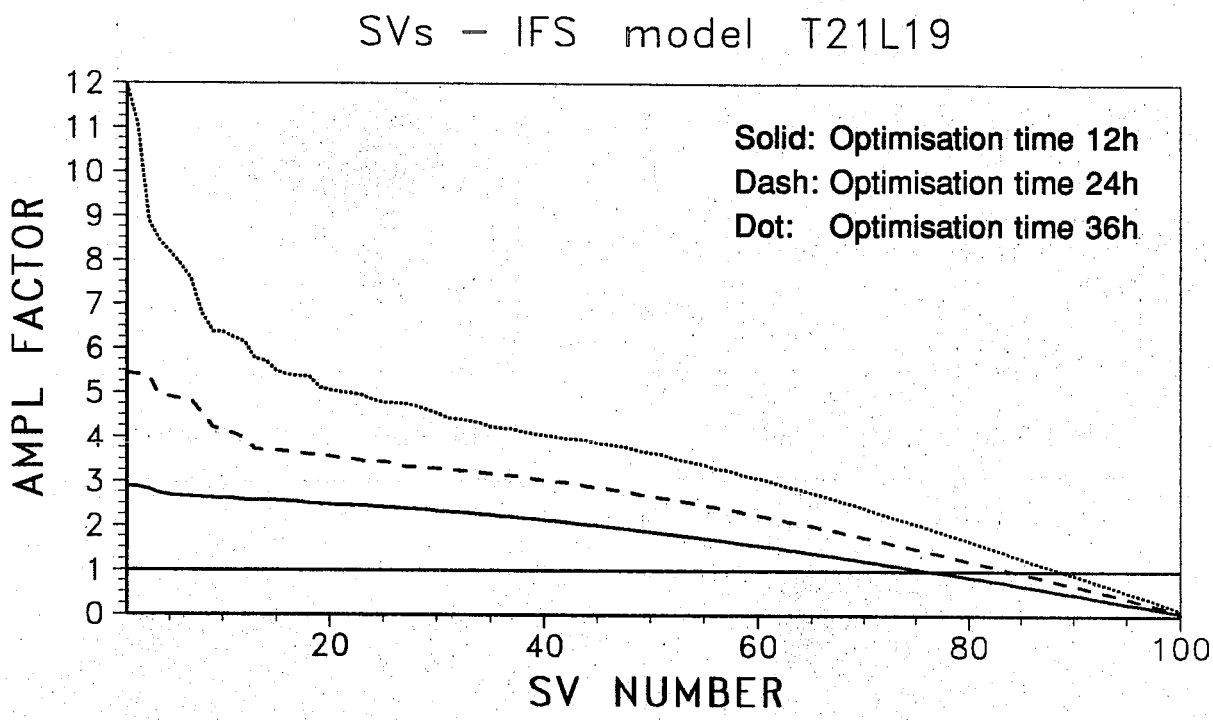


Fig. 17 Amplification factors of singular vectors in the IFS (T21).

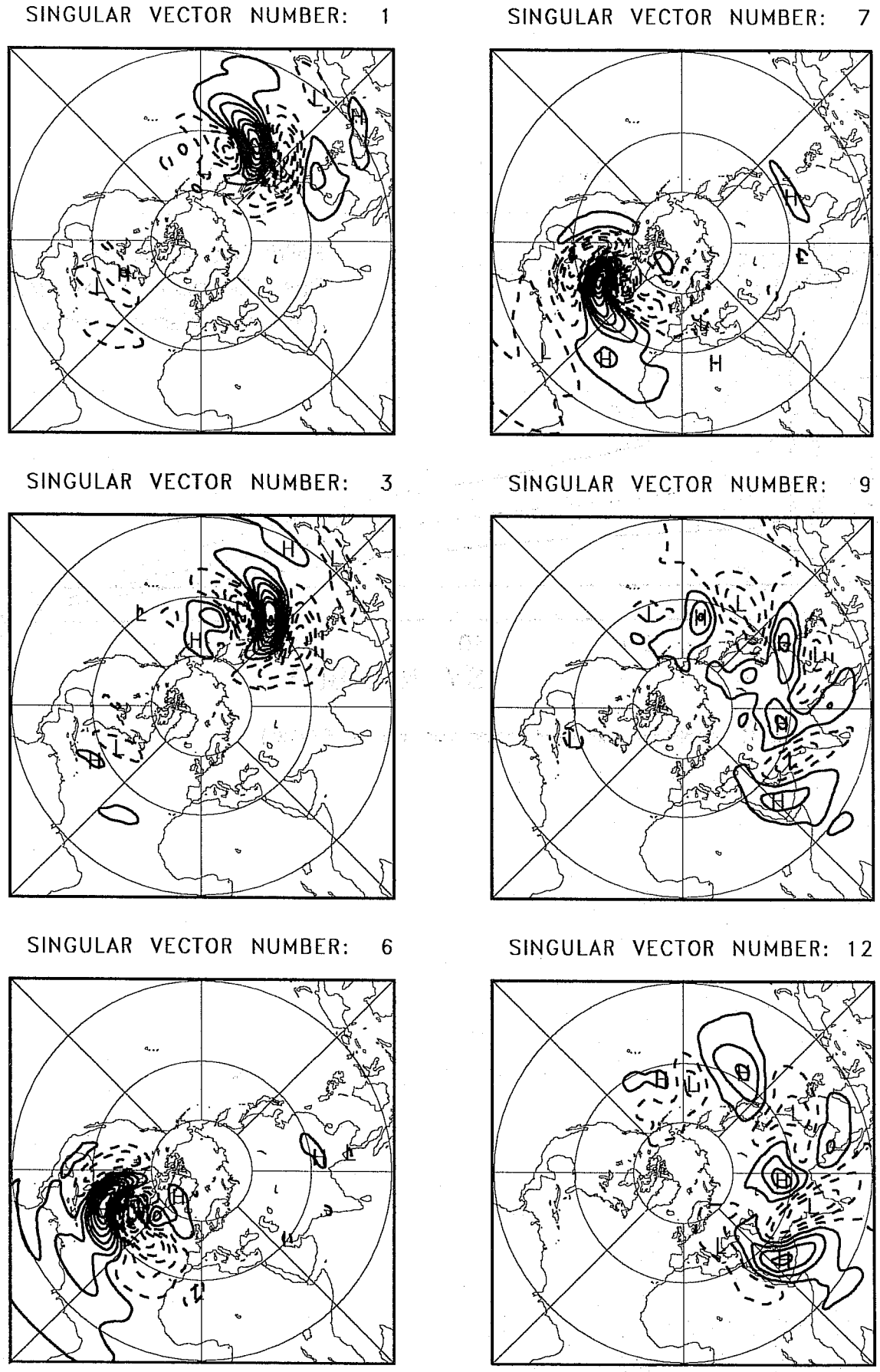


Fig. 18 Streamfunction of 24-hour singular vectors (1,3,6,7,9,12) at model level 11 (about 500 hPa) in the IFS model from data for 17 January 1989 (cf Fig. 1).

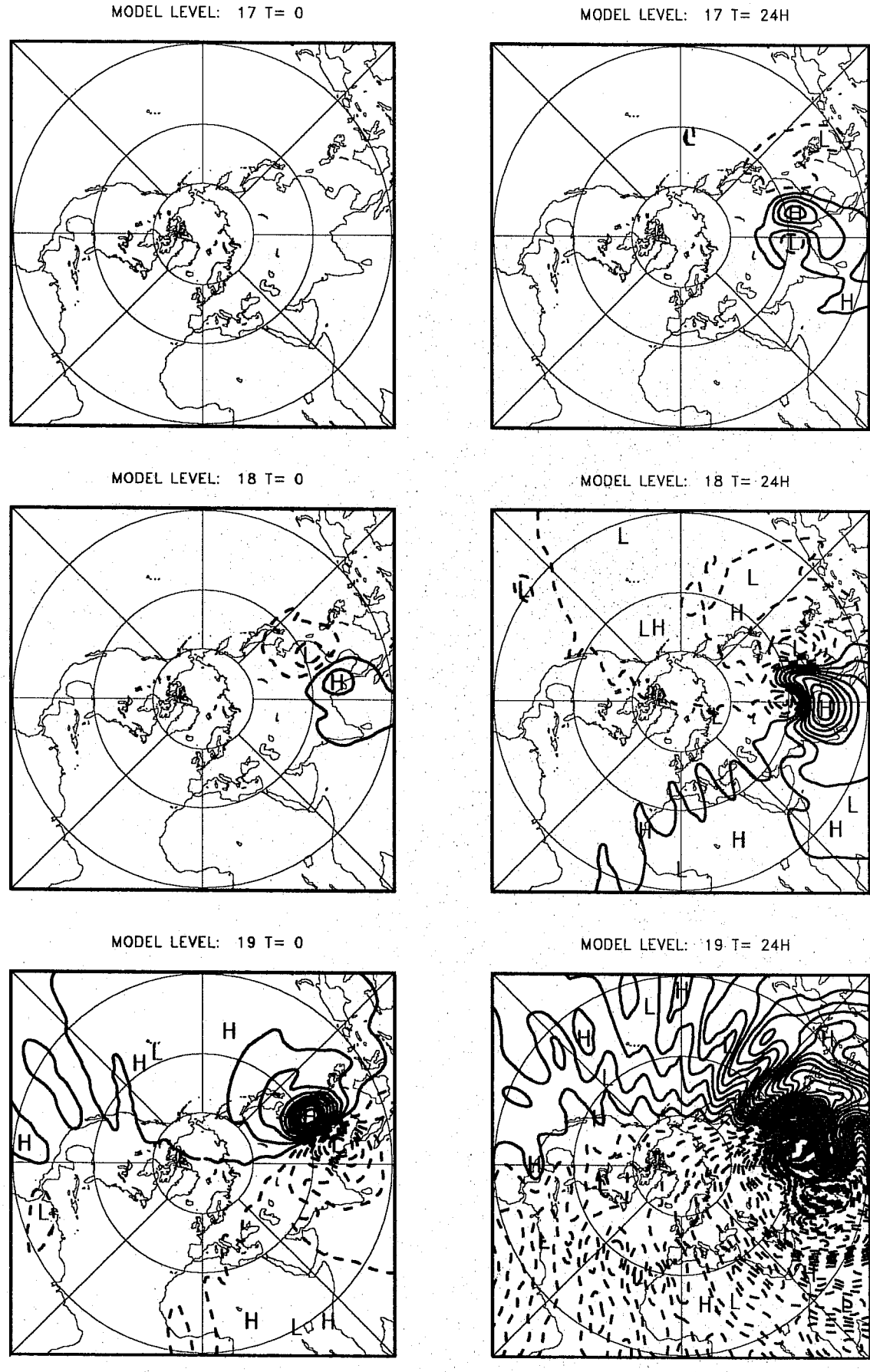


Fig. 19 Time evolution of second 24-hour singular vector (streamfunction) at model levels 17, 18, 19 in the IFS from data for 17 January 1989. The left-hand side panel shows the perturbation of the initial state, the right-hand side panel after 24 hours.