THE ECMWF ENSEMBLE PREDICTION SYSTEM: METHODOLOGY AND VALIDATION

F Molteni, R Buizza, T N Palmer and T Petroliagis
European Centre for Medium-Range Weather Forecasts
Shinfield Park, Reading

ABSTRACT

The European Centre for Medium-Range Weather Forecasts (ECMWF) Ensemble Prediction System (EPS) is described. Each ensemble comprises 32 10-day forecasts starting from initial conditions which have been perturbed away from the operational analysis. The perturbations are constructed from singular vectors of the short-range forecast trajectory. The calculation of these singular vectors, described in a companion paper (*Buizza and Palmer*, 1994), approximately determines the most unstable phase-space directions in the early part of the forecast period, and is estimated from the forward and adjoint tangent propagators of the ECMWF numerical weather prediction model. Relationships between the structures of these singular vectors at initial time and patterns showing the sensitivity of short-range forecast error to changes in the analysis are discussed. Scaling and phase-space rotation of the singular vectors are described, which together generate hemispheric-wide perturbations to the operational analysis for the EPS.

The validation of the ensembles is given firstly in terms of scatter diagrams and contingency tables of ensemble spread and control forecast skill. The contingency tables are compared with those from a perfect-model ensemble system. For some seasons there is no significant difference. Briar scores of the probability of climatological European flow clusters are described. It is shown that the scores are dependent on model errors. Finally, ensemble-member skill score distributions are discussed. Again these highlight the fact that the ensembles performed least satisfactorily during periods when model error was characteristically high.

Two cases are studied (one having large ensemble dispersion, the other having small ensemble dispersion). The case studies are used to illustrate the range of ensemble products routinely disseminated to ECMWF Member States. These products include clusters of flow types, and probability fields of weather elements.

1. INTRODUCTION

A major advance in operational numerical weather prediction (NWP) occurred in December 1992, when both the U.S. National Meteorological Center (NMC) and the European Centre for Medium-Range Weather Forecasts (ECMWF) started to produce and disseminate medium-range ensemble predictions (*Tracton and Kalnay*, 1993; *Palmer et al*, 1993). This event is the result of more than two decades of theoretical research and numerical experimentation, following the work of *Epstein* (1969), *Gleeson* (1970) and *Leith* (1974) which laid the theoretical and numerical foundations of a probabilistic approach to weather forecasting.

Since the actual state of the atmosphere at any time is known only approximately, a complete description of the weather prediction problem should be formulated in terms of the time evolution of an appropriate probability density function (PDF) in the atmosphere's phase space. Although this problem can be formulated exactly through the so-called continuity equation for probability (e.g. *Gleeson*, 1970), its practical solution is impossible for non-linear models with more than a few degrees of freedom. Even restricting the

attention to the evolution of the first and second-order moments of the atmospheric PDF (as in *Epstein*, 1969), one is still faced, for medium-range prediction, with a system of nonlinear equations which have no well-defined closure and which cannot be solved for the large numerical models currently used in NWP.

Ensemble forecasting appears to be the only feasible method to predict the evolution of the atmospheric PDF beyond the range in which error growth can be described by linearised dynamics. In ensemble forecasting, the PDF at initial time is represented through a finite sample of possible initial conditions. A nonlinear model integration is carried out from each of these states, and the properties of the PDF at any forecast time are assumed to be described by the sample statistics computed from the ensemble.

These statistics will approximate the correct PDF providing:

a) the sample of initial states has a realistic estimate of the probability distribution of analysis errors;

b) the phase-space trajectories computed by the numerical model are good approximations of atmospheric trajectories.

Requirement b) is also necessary for traditional, 'deterministic' NWP; hence, most of the recent research in ensemble forecasting has focused on point (a). However, as discussed below, systematic or regime-dependent model errors can severely affect the ability of the ensemble to forecast not only the first moment of the PDF, but also higher moments, such as the standard deviation.

Requirement a) poses a problem of considerable theoretical and practical difficulty. Firstly, the PDF of analysis error is poorly known; secondly, the number of independent directions in phase space spanned by this PDF (essentially the dimension of the NWP model) exceeds by many order of magnitude the maximum practicable ensemble size for a realistic NWP model. As demonstrated by early experiments in ensemble forecasting (see *Hollingsworth*, 1980), a sparse random sampling of phase space (even taking into account geostrophic and hydrostatic constraints) will not produce a realistic distribution of forecast states. For any given initial flow, only certain directions in phase space are associated with dynamical instabilities which will determine the growth of small perturbations (or errors) in the forecast.

Forecasts started from successive data assimilation cycles tend to diverge at a rate which is smaller but comparable to the actual error growth (*Lorenz*, 1982). The difference between the analysis at a given initial time and a very-short-range forecast verifying at the same time can therefore be considered as a growing perturbation consistent with our uncertainty in the initial conditions. This idea is exploited in the lagged-average forecasting (LAF) method proposed by *Hoffman and Kalnay* (1983), in which ensembles are composed of forecasts started from consecutive analyses. In this method, the ensemble size is limited by the number of available analyses in a relatively short time interval (typically not more than two days), and

the ensemble members cannot be considered as equally likely (at least in the medium range). These problems become less serious if one is mainly concerned with just the first moment of the sample PDF, namely the ensemble mean, or when longer forecast ranges are considered. In fact, LAF has been used in experimental programmes on extended-range ensemble predictions in most leading NWP centres (*Murphy*, 1988; *Tracton et al*, 1989; *Brankovic et al*, 1990; *Déqué and Royer*, 1992).

More recently, however, techniques to generate initial perturbation have been based on strategies (commonly used in dynamical systems theory) to identify those directions in phase space where dynamical instabilities are strongest. One possibility is to assume that errors in the initial conditions are dominated by those instabilities of the flow which have developed over a series of previous assimilation cycles. This assumption is the basis of the 'breeding' method proposed by *Toth and Kalnay* (1993), which corresponds to the computation of the vector associated with the largest Lyapunov exponent of the NWP model. The NMC ensemble prediction scheme is based on a combination of the breeding and LAF techniques (*Tracton and Kalnay*, 1993).

However, even assuming a white-noise spectrum for the initial error, which corresponds to an isotropic PDF in phase space at the initial time, the different amplification rates of perturbations along different axes would soon stretch the PDF principally along the directions of maximum instability during the early stages of the forecast period. In this way a particular phase-space direction, which perhaps was not necessarily associated with exceptional analysis error, may turn out to dominate the forecast error after a day or two. It would appear to be important to ensure that this direction was properly sampled by the ensemble of initial states.

As first shown by *Lorenz* (1965) in a meteorological context, for any finite time interval in which the dynamics of perturbations is assumed to be linear, the axes of maximum instability can be computed as the eigenvectors of a symmetric operator defined as the product of the linear propagator by its adjoint (see below for a more precise definition of these terms). In dynamical systems theory, this operator is sometimes referred to as the Oseledec operator (*Abarbanel et al*, 1991). In linear algebra notation, these eigenvectors are the singular vectors (SVs) of the linear propagator itself.

Singular vector growth can be much faster than either normal mode growth (for stationary flows) or Lyapunov exponent growth (for time-evolving flows) - see, for example, *Lacarra and Talagrand* (1988), *Farrel* (1990); *Borges and Hartmann* (1992); *Molteni and Palmer* (1993). Ensemble forecasting experiments in which unstable SVs computed from a 3-level quasi-geostrophic (QG) model were used to construct initial perturbations for a multilevel primitive equation (PE) model were carried out at ECMWF in the past four years (*Mureau et al*, 1993; *Palmer et al*, 1993).

This technique has proved to be more successful than those previously tested at ECMWF. However, the inconsistency between the vertical coordinates of the QG and PE model created difficulties in the vertical interpolation over high topography. Although this problem did not affect the strongly unstable SVs localised on the western side of the oceans, continental features often had a smaller growth rate in the PE than in the QG model (e.g. *Mureau et al*, 1993). Efforts were therefore directed towards the computation of SVs in a simplified PE environment, using an iterative Lanczos algorithm for the solution of the eigenvector problem. Preliminary results on this experimentation were reported in *Buizza et al* (1993); a more comprehensive description of the structure and dynamical properties of PE SVs can be found in a companion paper *Buizza and Palmer* (1994), hereafter referred to as BP.

Despite their rapid growth, the question of whether the SV structures, at initial time, have any correspondence to analysis error, has not been addressed. Determining analysis error statistics from conventional data assimilation techniques is difficult, and relatively little quantitative knowledge about flow-dependent 3-dimensional structure of analysis error statistics is available.

Recently, however, the development of adjoint models allows investigation of the component of analysis error which on any given day has the greatest impact on short-range forecast error (defined more precisely in section 2). This technique provides forecast sensitivity fields which can be directly compared with the SVs at initial time, since the forecast sensitivity fields do not necessarily indicate regions where analysis error was large, rather where analysis error has lead to significant forecast error. (For example, if analysis error was large in an area of weak dynamical instability, it may have a relatively unimportant contribution to forecast error.) We find (in section 2) that the structures of these forecast sensitivity fields share much of the same dynamical features of the SVs.

In December 1992, after almost nine years of experimentation in this field, ECMWF began to produce and disseminate real-time ensemble forecasts. For the first 16 months the ensemble system was run on an experimental basis on each Saturday, Sunday and Monday. From May 1994 they were run daily in full operational mode. Each ensemble comprises 33 10-day integrations of a reduced-resolution (T63L19) version of the operational (T213L31) forecast model, and post-processed products (including clusters of geopotential height fields and time-evolving PDF estimates for weather-related parameters) are disseminated to the Meteorological Services of ECMWF's Member States.

This paper describes aspects of the development of ECMWF's Ensemble Prediction System (EPS) and the results of the first year of its operational application. This introduction is followed by three main sections. In Section 2, the various components of the EPS are described, with emphasis on the properties of the initial perturbations and the post-processed products. In Section 3 we discuss results from the validation of

ensemble products, including estimates of the relationship between ensemble spread and skill, probabilistic verifications in terms of pre-determined flow types, and scores of individual ensemble members. In Section 4, two case studies are examined, characterised by different predictive skill and ensemble dispersion. These are used to illustrate the range of disseminated products. Finally, conclusions and plans are presented in Section 5.

2. THE ENSEMBLE PREDICTION SYSTEM

a) Singular vector computation

The description and computation of SVs are described in detail in BP. The following is a brief summary of the formalism.

Let $L_P(t,t_0)$ be the integral propagator of the dynamical equations, linearised about a nonlinear trajectory portion of the same dynamical equations, so that

$$x'(t) = L_P(t,t_0)\, x'(t_0) \tag{1}$$

maps a small perturbation (a vector in the tangent space to the dynamical system phase space) at time t_0, along the trajectory, to a small perturbation at future time t.

Let L_P^* be the adjoint operator of L_P with respect to a total energy inner product <..;..> (see equation 2.4 of BP). Then the total energy of a perturbation at time t is

$$\|x(t)\|^2 = \langle x(t_0); L_P^* L_P\, x(t_0)\rangle \tag{2}$$

The propagator L_P is itself compounded from the action of the linearised versions of the Non Linear Normal Mode Initialisation (NNMI) procedure, the adiabatic part of the model equations and of the physical parametrizations. For the cases described in this paper, the latter are restricted to a simplified surface drag and vertical diffusion scheme (*Buizza*, 1994a).

In March 1993, a further operator was introduced to compute SVs whose energy growth was maximised for the northern extratropics. The action of the Local Projection Operator (LPO; see section 2.4 of BP) allows one to calculate perturbations growing at final time t over a prescribed area; in our case, the northern hemisphere (NH) from 30°N to the pole. Hence, introducing the symmetric LPO T in (2), the energy in the selected region is

$$\|x(t)\|^2 = \langle x(t_0); K\, x(t_0)\rangle \tag{3}$$

where

$$K = L_P^* T^2 L_P \tag{4}$$

Denote by $\mathbf{v}_i(t_0)$ a normalised eigenvector of K, and by σ_i^2 the respective (real positive) eigenvalue, then since any $x(t_0)$ can be written as a linear combination of the set $\mathbf{v}_i(t_0)$, it follows

$$\max_{x(t_0) \neq 0} (\|x(t)\|/\|x(t_0)\|) = \sigma_1 \tag{5}$$

The $\mathbf{v}_i(t_0)$ and the σ_i are called the singular vectors (SVs) and the singular values of the operator TL_P. The time interval $(t-t_0)$ is called the Optimisation Time Interval (OTI).

In the EPS, SVs are computed applying an iterative Lanczos procedure (e.g. *Strang*, 1986) to a T21L19 linear version of the Integrated Forecasting System (IFS) model developed at ECMWF and Météo-France (*Courtier et al*, 1991). Usually, 100 iterations of the algorithm are enough to give about 30 SVs with sufficient numerical accuracy. Table 1 summarises the principal characteristics of the SV computation.

In the studies referenced above the evolution of the SVs from $\mathbf{v}_i(t_0)$ to $\mathbf{v}_i(t)$ was found to be particularly non-modal. In particular BP noted:

a) There was an upscale energy transformation between initial and final time. At initial time perturbation energy peaked at about wavenumber 20, close to the truncation limit of the T21 model. Further calculations with a higher resolution (T42), made since BP, suggest that this peak lies closer to wavenumber 25 for OTI's of about 3 days. At optimisation time, SV energy peaked at about wavenumber 10, corresponding to synoptic-scale features of the general circulation. Since the basic state is an unsmoothed solution of the nonlinear equations, Rossby triad interactions between the perturbation field and the basic state can generate such upscale effects even though the calculations are linear.

b) At initial time, SV energy peaked mainly in the lower troposphere (approximately the baroclinic steering level). At optimisation time, SV energy peaked mainly at jet stream level. In BP it was shown that this could be understood in terms of Rossby wave-activity conservation. In particular, since the intrinsic frequency I of an SV increases from lower to upper troposphere, an SV localised in the lower troposphere at initial time and in the upper troposphere at final time, will gain energy E, through (approximate) conservation of E/I.

c) Energy growth occurred through both barotropic and baroclinic processes. In particular, in the lower troposphere, the horizontal SV structure at initial time had SW-NE oriented phase lines north of the jet, and NW-SE oriented phase lines south of the jet. There was strong westward tilt with height between lower and

middle troposphere at initial time. These phase tilts were much less strongly pronounced at optimisation time. The westward tilt with height is consistent with an upward directed group velocity component, required in order that wave activity can propagate from lower to upper troposphere during SV growth.

d) Singular vectors tended to be localised and concentrated near the principal regions where the vertical wind shear was large: over the west Pacific, over the west Atlantic, and over subtropical north Africa. In BP the geographical distribution of the dominant SVs was shown to agree qualitatively with a simple index of baroclinic instability constructed from the seasonal-mean flow.

b) Relationship of SVs to analysis error

The objective of SV analysis is to find perturbations (or phase-space directions) that could generate significant forecast errors in the short range and that should be sampled for the medium-range ensemble forecast. We have discussed above the particular structures associated with SVs at initial time. It can therefore be asked whether these structures correspond to any known features of analysis error. This question is not straightforward; as already noted, our knowledge of analysis error is poor. Moreover, it may not necessarily be the case that the largest amplitude analysis error structures generate the largest forecast errors.

One way of tackling this problem is to use the adjoint propagator $\boldsymbol{L}_P^*$ to map actual short-range forecast errors back to initial time. Such studies have begun at ECMWF and preliminary results are documented in *Rabier et al* (1993, 94). It is straightforward to show that if $\boldsymbol{E}$ is the day 2 forecast error, then $\boldsymbol{L}_P^*\boldsymbol{E}$ is the gradient of the energy norm of $\boldsymbol{E}$ at the initial time. In other words a perturbation to the analysis equal to $\boldsymbol{L}_P^*\boldsymbol{E}$ will optimise the projection of the perturbation at day 2 onto the observed error field. Note that this differs fundamentally from $\boldsymbol{L}_P^{-1}\boldsymbol{E}$, which (in a perfect model) is the analysis error. We refer to $\boldsymbol{L}_P^*\boldsymbol{E}$ as a sensitivity pattern.

Rabier et al (1993,94) note that whilst the day 2 error peaks in the upper troposphere, the sensitivity pattern peaks in the lower troposphere, similar to SV structure. They also note that the sensitivity pattern has a strong westward tilt with height through the depth of the troposphere, similar to SV structure. Finally, spectral analysis of the sensitivity patterns has revealed a broad energy spectrum with an indistinct sub synoptic-scale maximum, again comparable with SV structure.

Rabier et al (1994) have studied the skill of ten-day forecast where the analysis is perturbed with the $L_p^* E$ perturbation. They note that the skill can be substantially improved well beyond the 48 hour period that the error field and the adjoint propagator were derived from. In this way it appears that perturbations to the initial analysis made purely on the basis of linear calculations from the early part of the forecast period, may have a beneficial impact on the forecast during the later period where errors are growing nonlinearly.

At the time of writing, routine daily calculations of both sensitivity patterns and singular vectors using identical tangent linear propagators, are about to commence. A detailed statistical comparison of these sensitivity patterns with EPS perturbations will be made once a sufficiently comprehensive dataset has been built up.

A second method of comparing SV structures with possible analysis errors has been made by *Thépaut et al* (1993) using a 4-dimensional variational analysis (4DVAR). In a particular case study of the October 1987 storm, the analysis increments were estimated at the beginning of a 24 hour assimilation period, resulting from an additional observation at the end of the assimilation period. *Thépaut et al* noted the similarity between this increment, and the structure of some of the initial dominant SVs computed over this period.

Hence, many of the statistical features associated with the SV structures appear to be found in the sensitivity patterns, and in 4DVAR analysis increments. On the other hand, the precise patterns of SV structures cannot be directly used as perturbations for the ensemble system. Firstly, the SV structures take no account of the distribution observations; secondly, individual SVs are too localised to represent hemisphere-wide analysis errors. We discuss the means to overcome these objections in the next section.

c) Definition of the initial perturbations

As mentioned in the Introduction, for each initial date, an ECMWF ensemble comprises one 'control' forecast (a T63L19 forecast started from the operational analysis) and 32 perturbed forecasts. The initial conditions for the perturbed integrations are constructed by adding and subtracting to the operational analysis 16 orthogonal perturbations defined as linear combinations of SVs.

The methodology used in the EPS to define these linear combinations is a modification of the procedure described in *Palmer et al* (1993). Its aim is to create perturbations which cover most of the Northern Hemisphere (NH), and have an amplitude comparable (in any region) to the estimates of root-mean-square (rms) analysis error provided by the optimum-interpolation (OI) data assimilation. We will now describe the procedure in more detail.

The first step is the selection of 16 SVs among the first 32 computed by the Lanczos algorithm. This proceeds as follows:

1) The first 4 SVs are always selected.

2) For each SV, a localisation function is defined in three-dimensional grid-point space, equal to 1 wherever the local energy (per unit mass) of the SV field is greater than 1% of its maximum value over the grid, 0 elsewhere.

3) An overlap function is defined at each point as the sum of localisation functions of the first four SVs. In general, the overlap function gives the number of selected SVs which 'cover' any grid point.

4) Each subsequent SV (5th, 6th....up to 32nd) is examined in turn, and selected only if more than half of its total energy lies in regions where the current overlap function is less than 4. If this is the case, the localisation function for the new SV is used to update the overlap function.

Step (4) is repeated until 16 SVs are selected. The final overlap function gives the number of SVs with at least 1% of their maximum local energy (i.e., 10% of their maximum amplitude) at any location.

Before the introduction of the LPO, the selection process was also used to eliminate SVs located in the southern hemisphere. The selection criterion was modified by requiring that more than 50% of the SV energy was located in the NH and in regions with small overlap function. During the boreal winter, most of the SVs computed by the Lanczos algorithm were in the NH, and therefore it was always possible to find 16 of them satisfying this criterion. In the following seasons, as the areas of maximum instability moved to the other hemisphere, computing enough SVs to be able to select 16 of them in the NH *a posteriori* would have been increasingly inefficient from computational point of view (as shown in *Buizza*, 1994b); therefore, the NH LPO had to be directly incorporated in the SV computation.

Once 16 SVs have been selected, an orthogonal rotation in phase space and a final rescaling are performed to generate the ensemble perturbations. Let $\boldsymbol{V}$ be the matrix whose columns are the 16 selected SVs, $\boldsymbol{R}$ a 16x16 orthogonal matrix and $\boldsymbol{D}$ a diagonal matrix of scaling factors; the matrix $\boldsymbol{P}$ containing the ensemble perturbations is computed by first defining

$$\boldsymbol{P}' = \boldsymbol{VR} \tag{6}$$

and then:

$$P = P'D. \tag{7}$$

Let $p_i' = \{u_i', v_i', T_i'\}$ be one of the orthonormal perturbations defined by (6), in terms of zonal and meridional wind and temperature respectively. Moreover, let e_u, e_v, e_T be the OI estimates of rms analysis error for these variables. The continuous function

$$f_i = [\overline{(u_i'/e_u)^8 + (v_i'/e_v)^8 + (T_i'/e_T)^8}]^{1/8} \tag{8}$$

where the overbar represents a mean over grid-point space, gives an estimate of the maximum local ratio between the perturbation amplitude and the estimated analysis error.

The rotation matrix R is defined in such a way to minimise the cost function

$$CF = \sum_{i=1}^{N} f_i^2 \,. \tag{9}$$

Since CF is not a simple quadratic function of the independent variables, the minimization cannot be reduced to the solution of a linear problem. Instead, we perform the minimisation iteratively by constructing R as the product of a series of 2x2 elementary rotation matrices.

In practice, the purpose of the phase-space rotation is to generate perturbations which have the same globally-averaged energy as the 'original' SVs, but a smaller local maximum and a more uniform spatial distribution. The iterative algorithm has proved effective in performing this task, despite the fact that the highly non-linear nature of the cost function may generate more than one minimum in phase space.

Once the rotation has been performed, the perturbations are rescaled in order to have a realistic local amplitude. The non-null elements of the diagonal matrix D in (7) are given by:

$$d_{ii} = \alpha / f_i \tag{10}$$

where α is a constant factor which represents the maximum acceptable ratio between perturbation amplitude and analysis error. On the basis of experimentation preceding the operational implementation of the EPS, a value of $\alpha = \sqrt{2}$ has been adopted.

An example of a rotated perturbation is given in Fig 1. This can be compared with the localised SV structures found in BP.

d) Statistical properties of ensemble perturbations

In this subsection we briefly discuss some of the statistics of the ensemble perturbations in the four calendar seasons of the year from December 1992 to December 1993. The initial date of the first and the last ensemble in each season is as follows:

winter: 19 December 1992 - 19 March 1993;

spring: 20 March - 18 June 1993;

summer: 19 June - 17 September 1993;

autumn: 18 September - 18 December 1993.

In total, 39 ensembles are included in each season.

An analysis of the statistics on the growth and distribution of the SVs was discussed in BP and is not repeated here.

Fig 2a-c shows the rms amplitude of the zonal component of wind at model level 11 (approximately 500 hPa) for the ensemble perturbations taken from winter, summer and autumn respectively. (The equivalent field for spring is shown in Fig 3b.) For winter (Fig 2a), the rms amplitude is maximised at approximately the regions where the vorticity maxima of the dominant SVs occur (see Fig 5 of BP), i.e. over the western Pacific and Atlantic basins, and over subtropical north Africa.

The impact of the application of the LPO can be seen comparing Fig 2a with rms amplitude distributions for other seasons. As discussed above, the LPO is designed to find SVs in which energy growth is maximised in the extratropical northern hemisphere. In particular, the LPO severely damps the rms amplitude of the subtropical north African perturbations. In common with the distribution of dominant singular vectors through the annual cycle (Fig 5 of BP), the rms amplitude of perturbations is more zonally asymmetric in winter compared with other seasons. There is a tendency for the oceanic regions to have higher rms amplitude, particularly in spring and autumn.

In Fig 3, we show, for the spring season, the rms amplitude at three model levels, corresponding to the lower, mid and upper troposphere. In common with the SV structure itself, perturbations over the Atlantic and Pacific have maximum amplitude in the lower troposphere. For the west Pacific maxima in particular, the strong westward tilt with height of the perturbations can be seen.

The wind amplitudes shown here (less than 1 m/s) are everywhere much smaller than the OI error estimates: in practice, the OI estimate of temperature error puts the strongest constraint on the perturbation amplitude. The rms temperature component of the perturbations (not shown) is largest around the 700 hPa level in all areas, with maxima between 1 and 1.5 °K. This can be expected from the hydrostatic relationship, given

the perturbation geostrophic wind maximum in the lower troposphere. These values are indeed comparable to the OI analysis errors, except on the eastern borders of North America and Asia, where the high density of radiosondes reduces the OI estimate to 0.6-0.7 °K.

e) Description of products for Member States

In this sub-section, we briefly describe ensemble products which are routinely disseminated to the National Meteorological Services of the ECMWF Member States. These are a subset of the diagnostics available from the ensembles.

i. "Stamp maps"

With an ensemble size of 33 forecasts per day, it is just about feasible for the human eye to assimilate qualitatively, information from each individual forecast. The set of 500 mb height forecast maps over Europe can be plotted on a single sheet of paper of A4 size or similar (e.g. Fig 13 in Sect. 4a). Although the size of an individual map is clearly minimal, it conveys the principal features of the synoptic-scale flow. One should not underestimate the processing power of visual cortex (*McIntyre*, 1988). In particular, the human eye is able to perform a subjective clustering which may be more relevant to the user than the objective methods discussed below. Moreover, the eye can readily spot whether the synoptic development of one or two individual ensemble members is unusual, and therefore worthy of further investigation. As ensemble sizes increase, this type of visual inspection will become less effective. Hopefully, objective probabilistic analyses will mature at a sufficient rate to compensate for this.

ii. Clusters of 500 hPa height trajectories

To condense the number of flow patterns predicted by the ensemble members into more basic varieties, a cluster analysis on the 500 hPa height fields produced by the 33 individual forecasts is performed. As in *Brankovic et al* (1990) and *Palmer et al* (1993), we have used Ward's hierarchical clustering algorithm, which has been applied to the flow over the European area (defined as 30°N-75°N, 20°W-45°E).

In the studies mentioned above, the clustering procedure had been applied at individual forecast times (or forecast intervals if time-averages were analyzed). For operational implementation, it was felt that information for more than one forecast time should be provided, though avoiding potentially confusing situations in which the grouping of ensemble members varied at different forecast ranges. It was therefore decided to cluster portions of forecast trajectories rather than instantaneous fields. This was done by defining the 'distance' between two ensemble members as the rms difference between height fields in the forecast interval from day 5 to day 7.

As in any hierarchical clustering, it is necessary to choose a criterion to select the 'best' partition of the ensemble members. This was based on an upper limit for the internal variance of the clusters (the mean square distance between individual members and their respective cluster centroids). Initially, during the winter season, this upper limit was 50% of the total sample variance. This choice guaranteed a good representation of the variability within the ensemble, but had the disadvantage of creating a number of very similar clusters when the ensemble dispersion was small; in this way, a large number of clusters did not imply a less predictable flow. From spring onwards, an absolute (rather than relative) limit for the internal variance of the clusters was adopted; this limit, which must vary with the seasonal cycle, was set equal to the monthly-average forecast error variance at day 3 derived from operational forecasts in previous years. The rationale behind this choice is that two medium-range forecasts can be grouped in the same cluster if their difference is of the order of a short-range forecast error.

The trajectory clustering performs well when there is a clear divergence of forecast trajectories in phase space, associated with possible transitions between large-scale regimes. In cases where the large-scale flow is more persistent, and the difference between ensemble members is mainly due to propagating baroclinic waves, trajectory clustering may lead to very smooth centroids in which the differences observed at individual forecast times are poorly represented. Of course, each clustering option (such as the choice of the clustering area and the time window, or the criterion for the 'best' number of clusters) has advantages and disadvantages; the usefulness of objective clustering would be greatly increased if these choices were made at the 'consumer' (i.e., operational forecaster) level rather than at the 'producer' level.

iii. Probability "plumes"

In order to give an assessment of the ensemble dispersion occurring throughout the forecast range at a particular location, "plumes" showing the time-evolving probability that the 850 hPa temperature lies within intervals of 1 K width are disseminated. The probabilities are computed assuming that each ensemble member is equally likely, and are expressed as percentages of the maximum possible value (which corresponds to all 33 ensemble forecasts being in a 1 K interval). A Gaussian smoother is applied to the sample frequencies to produce smooth probability estimates.

This assumption of equal likelihood of all ensemble members requires further discussion. If we assume that the analysis error is isotropic in phase space, and if we were able to sample all the unstable directions (growing singular vectors), then the control integration should have a weight over the perturbed forecasts given by the ratio of the total phase space dimension, to the dimension of the unstable subspace. In this way, the weighting given to the control would implicitly take into account the number of decaying directions, which have little impact on ensemble dispersion. However, in BP it was noted that the dimension of the unstable subspace exceeded the 16 dimensions used for constructing the ensemble

perturbations. In this way it is not safe to assume that the control is representative of all the phase-space dimensions not explicitly sampled. Hence it is not safe to give the control excessive weighting. For simplicity we have taken the opposite extreme where the control is as likely as the perturbed members. Note that this is a somewhat conservative assumption, since, compared with a weighted control, a given forecast probability would correspond to a smaller ensemble dispersion.

Unfortunately, in regions of steep orography, the T63L19 850 hPa temperature can be locally inconsistent with the operational forecast values (produced with a T213L31 model), even when both forecasts have essentially identical synoptic-scale features. In order to allow the forecaster to assess whether inconsistencies in low-level temperature between operational and ensemble forecasts are due to differences in synoptic flow, plumes of 850 hPa temperature were supplemented with plumes of 500 hPa geopotential height. The width of the height categories is 2.5 dam.

Examples of probability plumes are shown below in Fig 20.

iv. Probability maps

Two dimensional fields representing probabilities of rainfall, 10 m wind speed and 850 mb temperature anomalies for specific forecast days are also postprocessed and disseminated. As with the plumes, the probabilities are calculated on the basis that each ensemble member is equally likely. The rainfall categories are: > 1 mm/day, > 5 mm/day, > 10 mm/day and > 20 mm/day. The wind speed categories are: > 10 m/s and > 20 m/s. Finally the temperature anomaly categories are: < -8°K, < -4°K, > 4°K and > 8°K. Examples of probability fields for rainfall and temperature categories are given in Sect. 4 (see Figs 15 and 19).

3. OBJECTIVE VALIDATION OF ENSEMBLE PREDICTIONS.

a) Relationship between ensemble spread and forecast skill

One of the principal uses of an ensemble forecast is to provide an estimate of the confidence in a prediction; the larger the ensemble dispersion, the less reliable is the forecast by any one member. From this basic notion, it is often assumed that the ensemble spread can be taken as a predictor of the skill of the control forecast. However, even in a perfect environment, spread will not be perfectly correlated with skill (*Murphy*, 1988; *Barker*, 1991). Consider a well-sampled PDF integrated with an error-free model. When spread is small, the control forecast trajectory is constrained to be close to the verifying analysis trajectory; however, when the spread is large, the control forecast is not constrained to be far from the verifying trajectory. Hence for large spread, the control forecast could be skilful if, by chance, it happened to be close to the verification trajectory. As a result, when discussing relationships between ensemble spread and control forecast skill, we compare results with those from a hypothetical perfect-model ensemble.

Fig 4 shows scatter diagrams of skill and spread for the northern hemisphere for all four seasons. Here the ensemble spread is taken as the 75 percentile of the distribution of the rms 500 mb height difference between the perturbed ensemble members and the control. The day 7 rms error of the control is taken as the skill value. (Spread/skill relationships have also been studied using anomaly correlation as the measure of distance between either control or verifying analysis, but are not discussed here for brevity. Results do not depend strongly on the measure used.) For each diagram, the distribution is divided by the median value, and the number of elements in each quadrant is shown in the figure (non-bracketed numbers). These give a 2x2 contingency table for high/low spread, high/low skill cases. (Note that by using the median to define the categories, the contingency table is necessarily symmetric).

For all four seasons, and for both chosen areas, the diagonal entries are notably more populated than the off-diagonal entries. Whilst the off-diagonal entries are not negligible, we noted above that even in a perfect model environment we would expect the off-diagonal elements to be non-zero. We have estimated a "perfect-model" contingency table by taking, at random, one member of each ensemble to be a verifying analysis and averaging the results over several possible realisations. The contingency table for this perfect model verification is given in parentheses in Fig 4. It can be seen that for autumn the two distributions are identical. For winter there is a difference of just one member. Indeed only for summer would a chi-square test indicate that the two distributions might not be similar (though by eye, this scatter diagram does not appear particularly poor). The linear correlation coefficient for these scatter diagrams ranges between 0.6 and 0.7. Values compare very well with perfect model estimates of 0.6 for the longer forecast time (day 12) reported in *Barker* (1991). The spread/skill relations appear superior to those obtained by statistical means (e.g. the day 7 NH values between 0.3 and 0.4 reported by *Molteni and Palmer*, 1990).

It is interesting to note directly from the scatter diagrams, that in general there is a smaller range of control skill values for ensembles with small spread, than for ensembles with large spread. This is particularly true in winter and summer, and was anticipated by the remarks at the beginning of this section. We have studied sub-regions of the NH using these measures of spread and skill. Results are not shown here for brevity, but give comparable contingency relations to those in Fig 4.

A second method of analysing the relation between forecast skill and ensemble dispersion is illustrated in Fig 5 which shows scatter diagrams for the European region (30N-75N, 20W-45E) based on forecast probability plumes for 500 hPa geopotential height at 513 grid points distributed uniformly throughout the region. Results for all four seasons are shown. The spread is estimated from the forecast time a chosen probability contour first vanishes (for example, from the plume shown in Fig 20b, the 30% contour first vanishes at about day 9). The mean forecast time averaged over all the grid points is then taken as the appropriate measure of spread (the smaller the spread the longer the forecast time). The chosen probability

contour used to estimate ensemble dispersion varies with season. For winter it is taken as the time at which the 20% contours disappears; for summer it is taken from the 30% contour, and for autumn and spring we use the 25% contour. This variation in spread diagnostic is reasonable; ensemble spread is generally weaker in summer so that the 30% contour more generally extends into the medium range, and is therefore a reasonable predictor of medium range skill. By contrast in winter the 30% contour may vanish closer to the end of the short range, and therefore would not be a good indicator of medium-range skill. The skill measure of the control forecast is taken as the mean rms 500 hPa height error of the forecast averaged over days 5-7.

Generally it can be seen that the clearest relation between plume dispersion and forecast skill occurs in spring and summer, with autumn the poorest. The number of elements in each of the four quadrants (with boundaries as the seasonal mean skill and dispersion) is shown at the corner of each quadrant as in Fig 4. (Unlike the earlier diagram, estimation of a perfect model contingency table is not given.) As for the NH spread/skill statistics, the contingency tables are dominated by diagonal entries. The poorest season, with most off-diagonal elements, is autumn.

Focusing on the autumn season, there are four ensemble forecasts (shown circled in Fig 5) for which the relationship between spread and skill appears to break down. We have studied these in more detail. Two particular facts are worthy of note. Firstly, for all four cases, the verifying analysis had either a deep trough or a cut-off low over Europe. Secondly, in three of the four cases, the operational T213L31 model was noticeably more skilful than the T63L19 control. This latter remark should be seen in the context that over the full set of forecasts described here, the higher resolution model was not found to be more skilful than the T63L19 model beyond about day 5 (see Section 3c). It is therefore possible that the failure of these cases may be attributable to T63L19 model error, rather than deficiencies in the initial perturbations.

b) Probability of synoptic flow patterns

In this sub-section, we investigate the quality of the information provided by the ensembles in terms of probability of alternative synoptic flow patterns, concentrating on the Euro-Atlantic region during winter. As mentioned in Sect. 2b, clusters of 500 hPa height for the European region are computed and disseminated for every ensemble. The probability of occurrence of each cluster is taken to be proportional to the cluster population. Although clusters computed from individual ensembles have the advantage of explaining a large fraction of the ensemble variance with relatively few centroids, the fact that the number and pattern of these centroids vary every day makes an objective verification of the cluster probabilities impossible with standard skill tests.

For this purpose, as in *Palmer et al* (1993), we have therefore used a 'fixed' hierarchy of clusters, computed by applying the Ward algorithm to (instantaneous) daily analyses of 500 hPa height in 12 winters (1979/80 to 1990/91). In this case, the clustering area covers both the Atlantic and Europe, from 45°W to 45°E and from 30°N to 80°N. Three hierarchical clustering levels, including 12, 8 and 4 clusters, have been selected for the verification; for brevity only results from the 8-member clusters are shown.

There are two advantages in using a set of clusters which are representative of the climatological distribution of atmospheric states. Firstly, it is straightforward to compare the skill of the ensemble probabilities against the climatological probabilities. Secondly, if one also knows the frequencies of these clusters in the climatology of the numerical model used for the ensemble forecasts, one can estimate whether model systematic errors (which are reflected in the differences between observed and modelled frequencies) are likely to affect the ensemble probabilistic predictions of certain particular flow types.

The climatological distribution of T63L19 model states has been estimated from a set of 15 120-day integrations with observed sea-surface temperature as boundary conditions, started on 1, 2 and 3 November 1986 to 1990; these integrations are identical to those described by *Brankovic et al* (1994) apart from the fact that they were performed with an updated version of the T63L19 model (the so-called cycle 46, which was used in the EPS for about six months). The 500 hPa height fields corresponding to the last 90 days of each integration were classified in one of the 8 clusters computed from the analysis sample, according to their similarity with observed composite fields belonging to the clusters.

Fig 6 shows the 8-cluster centroids, with observed and modelled climatological frequencies. The frequencies of clusters 5, 6, 7 and 8, corresponding to flows with a strong ridge or blocking high over the Atlantic or northern Europe, and (for clusters 5 and 7) a trough or cut-off low over southern Europe, are severely underestimated in the long-term climatology of the T63L19 model. For any of the three cluster sets used in this verification, the observed and modelled climatological frequencies were significantly different at the 99.5% confidence level according to a chi-square test.

For each ensemble in the winter season, the control forecast, the 32 perturbed forecasts and the verifying analysis have been classified into one of the 8 'climatological' clusters, from forecast time $t = 0$ to $t = 10$ days at 12-hour intervals. We therefore have three probability distributions $P_a(j,t)$, $P_c(j,t)$, $P_e(j,t)$, where j is the cluster index and the subscripts a, c and e indicate the analysis, the control and the full 33-member ensemble respectively.

Let $j_a = j_a(t)$ be the index of the cluster in which the analysis is classified at forecast time t. Then

$$P_a(j_a,t) = 1, \quad P_a(j \neq j_a,t) = 0. \tag{11}$$

Similarly, the control forecast probability P_c is 1 for the predicted cluster and 0 for the others, whereas for the ensemble, P_e is proportional to the number of ensemble members classified in each cluster.

Using these probabilities, the seasonally-averaged Brier score for the ensemble probabilistic forecast of Euro-Atlantic clusters can be defined as

$$B_e(t) = \overline{\sum_{j=1}^{M} [P_e(j,t) - P_a(j,t)]^2} \tag{12}$$

where M is the number of clusters and the overbar represents the average over the 39 winter ensembles. An average Brier score $B_c(t)$ can also be defined for the control forecast by substituting P_c to P_e in (12).

As a reference, one can compare these scores with the Brier score $B_{cl}(t)$ of a climatological forecast, obtained by using the observed climatological frequencies $P_{cl}(j)$ of the clusters instead of the predicted probabilities. In theory, B_{cl} should be independent from forecast time, and given by

$$B_{cl} = 1 - \sum_{j=1}^{M} P_{cl}^2(j) \tag{13}$$

although in our verification a weak time dependence arises because of sampling problems.

Using the definition of P_a, one can rewrite (12) as:

$$B_e(t) = 1 + \overline{\sum_{j=1}^{M} P_e^2(j,t)} - 2\overline{P_e(j_a,t)} \tag{14}$$

where $\overline{P_e(j_a,t)}$ is the average probability of the verifying cluster. The second term on the right hand side of (14) depends only on the smoothness of the probability distribution, and is necessarily less than 1 unless just one cluster is assigned a non-zero probability. For the control forecast this is always the case, so that

$$B_c(t) = 2[1 - \overline{P_c(j_a,t)}] \tag{15}$$

If we make the assumption that on average any ensemble member (including the control) has the same probability of predicting the correct cluster, then $\overline{P_e(j_a,t)} = \overline{P_c(j_a,t)}$, and B_e must be lower than B_c provided that the ensemble members are distributed in more than one cluster. In other words, the difference in Brier

score between the ensemble and the control forecast does not reflect a greater probability of predicting the correct cluster (on average), but rather the capacity to provide an *a priori* estimate of this probability.

Another useful comparison can be made between B_e and the theoretical score that the ensembles would achieve if the predicted probabilities were an exact estimate of the probability of occurrence of each cluster. The score of this hypothetical 'perfect' ensemble is given by

$$B_{ep}(t) = 1 - \overline{\sum_{j=1}^{M} P_e^2(j,t)} \qquad (16)$$

In the absence of model systematic errors, the ensemble probability distribution should asymptote to the observed climatological distribution, and therefore B_e and B_{ep} should tend to B_{cl}. On the other hand, the Brier score of the control forecast should asymptote to 2 B_{cl} (this behaviour is analogous to that of the mean-square error of a deterministic vs. ensemble-mean forecast).

Curves of B_e, B_{ep}, B_c and B_{cl} are shown in Fig 7a for the 8-cluster verifications. Firstly, we notice that the ensemble score B_e crosses the climatological score at fc. day 8. The non-monotonic growth of B_e makes it difficult to assess whether the score has reached saturation at day 10. However, beyond day 8 the difference between B_e and B_{cl} is very small.

Before comparing the ensemble with the control forecast and the theoretical 'perfect ensemble', it is worth commenting on the non-monotonic behaviour of the control score B_c, which is also reflected in B_e. This implies (see (15)) that the average probability of a correct cluster prediction by the control forecast does not decrease monotonically with forecast time. It is likely that such behaviour is not entirely due to sampling problems.

Consider what happens when the atmosphere makes a transition from the initial cluster (i.e. the cluster at $t = 0$) to another cluster. The control may perform in three different ways:

- it makes the correct transition at the correct time;
- it makes the correct transition but at a different forecast time;
- it makes a transition to a different cluster.

In the second case, the Brier score for that particular forecast will vary from 0 to 2 and then back to 0, showing a 'return of skill' which may not be so evident in scores like rms error. Indeed it appears that this type of error is particularly frequent in the early medium range; in the short range, cluster transitions are usually well predicted, while transitions in the late medium-range are usually associated with 'irreversible'

errors. More quantitative arguments, based on statistical models of error growth and distributions of transition times, confirm that the probability of an error in the time of the first transition must have a maximum in the medium range.

Hence, at least a plateau in the Brier score of the control forecast should be expected. This argument, however, cannot be applied to the ensemble; although the transition times of individual members may differ from the analysis, the ensemble probability of transition should peak at the correct time (unless systematic deficiencies are present in the model variability or in the construction of initial perturbations). Indeed, the Brier score B_{ep} of the 'perfect ensemble' is always strictly monotonic in Fig 7a.

Let us now return to the comparison between B_e, B_{ep} and B_c as illustrated in Fig 7a. Ideally, the ensemble score should be as close as possible to B_{ep} and substantially lower than B_c in the medium range. In reality, the B_e curve is roughly equidistant from B_{ep} and B_c, and shows a non-monotonic growth as B_c does. This result indicates that the ensemble follows the control forecast too closely, especially in the day 3-to-5 range. This may occur because of quasi-systematic model deficiencies or because of a non-optimal behaviour of the perturbations.

It is reasonable to assume that systematic model error will be mainly felt when the real atmosphere resides in those clusters whose frequency is severely underestimated in the long-term climatology of the model. We have therefore recomputed the Brier scores excluding those ensembles in which the verifying analysis was classified in clusters 5 to 8 of the 8-cluster partition (Fig 6) for more than half the time beyond forecast day 3. The average Brier scores for the remaining 22 ensembles are shown in Fig 7b.

While the B_{ep} score is only marginally changed, consistently with a perfect model assumption, the score of both the ensemble and the control forecast are substantially improved in the medium-range. The B_e curve shows a plateau rather than a return of skill. However, B_e remains nearly equidistant from B_{ep} and B_c, indicating that the ensemble remains too 'supportive' of the control forecast even in non-blocked flows.

In conclusion, the Briar score indicates that the ensemble probability distribution estimate is skilful in the medium range at least up to about day 8. The score is significantly affected by model error; however, it is also likely that there may be deficiencies in the initial perturbations, leading to insufficient spread.

c) Skill scores of ensemble members, control and operational forecasts

We conclude this section on objective validation by showing time series of the ensemble distributions of conventional skill scores. Anomaly correlation (AC) and rms error of 500 hPa height have been computed over various areas for all the individual forecasts in each ensemble as well as for the T63 control and the T213 operational forecast. Here, scores are illustrated for the whole NH (north of 20°N) and for Europe (as defined in Sect. 2d.ii). Figs 8-9 show for each ensemble the two extreme scores (dashed lines), the 25% and 75% percentiles (solid lines) and the median (dotted line). In addition, the score of the T63 control is indicated by a full circle, and the score of the operational forecast by a square.

The diagram for day-7 AC over the NH in winter is shown in Fig 8a. The AC of either the control or the operational forecast was usually between 0.6 and 0.8, except for two periods of poor performance at the end of December 1992 and February 1993. In all cases, the AC of these two unperturbed forecasts was lower than the AC of the best member of the ensemble (which was usually around 0.8), and in most of them lay between the 25% and 75% percentiles. On the other hand, the score of the best ensemble member dropped considerably during the two periods mentioned above, indicating that none of the perturbed forecasts managed to stay close to the actual atmospheric trajectory.

The corresponding distributions for rms error (not shown) support the results above. However, rms errors tend to give a more favourable view of the performance of the EPS with respect to the operational forecast; this reflects the tendency of the T213 model to produce more intense features which, if out of phase with the verifying analysis, have a stronger negative impact on the rms error than on the AC.

Fig 8b shows the wintertime AC statistics for Europe, again at fc. day 7. The smaller verification area generates a wider range of scores spanned by the ensemble members. In all but three cases, the scores of the unperturbed forecasts are within the range of the perturbed integrations. The three exceptions refer to the operational forecast, and in two of them all the ensemble members were more skilful.

The two periods of poor performance found on the hemispheric domain are also reflected in the scores for Europe. In addition, poor scores in the unperturbed and in a large majority of the perturbed forecasts can be seen on 23-24 January and 8 March. In all these cases, the synoptic situation over Europe at fc. day 7 showed a marked split of the westerly flow with blocking highs over northern Europe and/or cut-off lows over the Mediterranean region, characteristic of clusters that were inadequately simulated by the model in climatological mode. Again, this highlights the possibility that model error may have contributed to the failure of some of the ensembles.

The three panels of Fig 9 show the day-7 AC distribution over Europe in the other seasons. EPS scores were least satisfactory in spring; one can see in Fig 9a two fairly long periods (respectively at the beginning and at the end of the season) in which none of the anomalies predicted by the ensemble members was strongly correlated with the observed one. Still, even in these periods, the 'best' ensemble member was often considerably more skilful than the operational forecast.

The EPS performance over Europe during summer and autumn was consistently good; except for few and fairly isolated exceptions, the most skilful ensemble member had a fairly high AC, of the order of 80%. As in the other seasons, the scores of the control and operational forecast were within the 25-75% percentile band in the majority of cases.

Indeed, from Figs 8-9 we can deduce the number of times when at least one member of the ensemble was more skilful than either the control, or the operational forecast. With respect to the control it can be seen that there is only one case (number 23 in summer) where the control is clearly more skilful than any member of the ensemble distribution. There are more cases where the operational forecast was more skilful than any ensemble member (1 in winter, 2 in spring, 1 in summer, 3 in autumn). On the other hand there are 2 cases where the operational forecast was worse than any ensemble member.

Based on these skill score distributions, it is possible to estimate the percentage of times the verifying analysis trajectory lay significantly "outside" the ensemble distribution. This can be done by counting the number of cases where the AC of the most skilful ensemble member ε_s was less than some prescribed threshold AC_T. Results are given in Table 2 for day 7 over Europe, and based on $AC_T = 0.6$. Cases where the skill of ε_s was less than 0.6 constitute major ensemble failures. For comparison, the percentage of corresponding operational forecasts with $AC < 0.6$ is shown.

The most numerous ensemble failures occurred in spring. In particular, the fact that in over a quarter of springtime ensembles, the best day 7 member had an $AC < 0.6$ is problematic. However, this period was also especially poor for the operational model. By comparison, summer and autumn ensembles performed more satisfactorily; in only 5% of cases was the verifying analysis substantially outside the ensemble distribution at day 7.

The impact of model systematic error on these ensemble statistics has been studied for the winter season. Specifically, the percentage of ensembles whose best member has $AC < 0.6$ has been recomputed excluding the cases where the verifying analysis was in clusters 5-8 (see section 2b) at least half the forecast period.

The computed value reduced to 6% (1/18 as compared with 3/39 for the full set) for the 0.6 threshold. Again it appears that model error may play a role in accounting for some of the poorest ensembles. On the other hand, it should be recognised that sample sizes are fairly small.

It is curious to note that there is no obvious correspondence between the seasons with either high or low skill scores on the one hand, and either best or poorest spread/skill relationships on the other. For example over Europe the springtime spread/skill contingency tables were relatively good (see section 2a), despite having the worst overall skill scores. Similarly autumn spread/skill contingency tables over Europe were relatively poor, despite having satisfactory skill scores.

4. CASE STUDIES

In this section, we study in some detail two particular ensemble forecasts associated with the development of strongly meridional flows over Europe. We shall focus on the medium-range performance of the ensembles between days 5 and 7. These examples illustrate situations where the ensemble appeared to perform satisfactorily. In the first, the ensemble spread was large and the control and operational forecasts were poor; in the second, the ensemble spread was small and the operational and control medium-range forecasts were skilful. The case studies are used to illustrate the range of products disseminated to the National Meteorological Services of ECMWF member states.

a) 30 October - 6 November 1993

In discussing results from this case study, we take into account the fact that three consecutive ensemble predictions are made each week. Initial conditions (T=0) are shown for the first of these consecutive forecasts, and verification is made for day 7 of the first ensemble, day 6 of the second ensemble, and day 5 of the third ensemble (collectively referred to as T=7 days). A smooth evolution of probability distributions from consecutive ensembles is a desirable property of a successful ensemble prediction scheme (contrasting with the often discontinuous behaviour of consecutive single deterministic forecasts).

Fig 10 shows the analysed 500 hPa height fields. At T=0 (Fig 10a), a meridionally-oriented dipole is already well established over western Europe and the eastern Atlantic. This pattern evolves slowly through to T=7 (Fig 10b) with low geopotential height values dominating most part of Europe, including two minima over the Iberian peninsula and over southern central Europe. The region of high geopotential height has evolved northward, with a maximum over Scandinavia by T=7. The region where rainfall from the verification exceeded 10 mm/day is shown superimposed on the height field in Fig 10b. This verification is in fact taken from a 24 hour operational forecast from 6 November. It has been checked, where station data is available, that this estimate is accurate. The largest rainfall rates are associated with the low height centre over northern Italy.

Day 7, 6 and 5 operational forecast height and precipitation fields verifying at T=7 are shown in Fig 11. Over northern Europe, the jet at both days 7 and 6 is too zonal, resulting in significant precipitation over the west coast of Norway. Over southern Europe, the position of a cut-off low is forecast for the straits of Gibraltar at day 7, Italy at day 6, and northern Spain at day 5. This results in considerable inconsistency in rainfall prediction over southern Europe.

Fig 12 shows the dispersion over Europe for the three consecutive ensemble forecasts, as measured by the anomaly correlation between each perturbed member of the ensemble forecast and the control. It can be seen that there is an overall tendency towards increased predictability between the 30 October ensemble and the 1 November ensemble. For the first ensemble, 50% (16 out of 32) of ensemble members have an anomaly correlation of less than 0.6 with the unperturbed control forecast at day 7. For the second and third ensembles (again at day 7), this percentage reduces to 25 (8 out of 32) and 6.25% (2 out of 32) respectively. Hence not only is the day 5 prediction for T=7 more reliable than the day 7 or 6 forecasts because of shorter lead time, but also the atmosphere is evolving towards an intrinsically more predictable phase between 30 October and 1 November.

Fig 13 shows an example of 500 hpa height "stamp maps" at T=7 described in section 2e (i) , for the day 6 ensemble. A forecaster studying this set of fields has a clear impression of the disturbed nature of the flow over Europe, and of the variety of solutions offered by the ensemble. Over southern Europe, many of the members indicate a cut-off low, though the positioning is uncertain (compare, for example, members 7 and 25). Over Scandinavia, some members show significant troughing (e.g. members 14 and 23) while many of the rest show ridging (e.g. members 12 and 27).

Fig 14 shows an objective clustering applied to these fields, using the method described in section 2e (ii) (though for illustrative purposes, the clustering is only applied to the fields in Fig 13). For this ensemble, four clusters were chosen. We show the rms and anomaly correlation skill of the clusters at the bottom of the diagram. The majority of ensemble members are grouped in clusters 1 and 2. These mainly differ in the position and intensity of the trough over southern Europe. Either of these clusters is more skilful than the operational forecast. The 3rd and 4th clusters together only contain 8 elements. The 3rd is associated with the north European trough over the west coast of Scandinavia, whilst the final cluster (the most skilful) has a more dominant ridging over the whole of northern Europe and the north-east Atlantic.

Fig 15 shows rainfall probability maps for T=7, for the three ensemble forecasts (as discussed in section 2e (iv), this is given by the ensemble fraction in which rainfall accumulated between days 6 and 7 November exceeded 10 mm/day). It can be seen that the probability distributions evolve much more smoothly from day 7 to day 5 than do the operational precipitation forecasts. For the region over southern

central Europe with strong rainfall rates reported from station data (see Fig 12), the ensemble probability increases monotonically, whilst for Ireland, rainfall probabilities are decreasing. The operational prediction of rain over western Norway is not strongly supported by any of the ensembles.

Overall, this is a case where forecaster confidence would have been relatively low.

b) 13 -20 November 1993

The analysed 500 hPa height for the second case study (two weeks after the first case study) is shown in Fig 16. Superimposed on the height contours for 20 November are regions where the 850 hPa temperature anomaly either exceeded 4K, or was less than -4K. At T=0 (Fig 16a), the flow is zonal across the Atlantic and western Europe, with a reversed gradient height dipole at about 40E. At T=7 day (Fig 16b), this dipole has amplified with major height anomaly centres over Scandinavia and central Europe. Associated with the flow, 850 hPa temperatures are anomalously cold over much of Europe, though anomalously warm in the extreme north.

For brevity we shall only discuss the first of the three consecutive ensemble forecasts from this (high predictability) period. The ensemble dispersion from T=0 is shown in Fig 17. By comparison with the previous case, ensemble spread is relatively small, and hence this forecast period appears to be relatively predictable. Indeed, virtually all of the day 7 forecasts for T=7 have an anomaly correlation of at least 0.6 with the control forecast.

The two largest clusters (8 members each) from this ensemble are shown in Fig 18. (As can be seen they have an anomaly correlation of 89% and 93% respectively with the verifying analysis and differ in the positioning of the meridional height dipole.) It is interesting to note that over western Europe, the flow type evolved from a largely zonal flow, to a largely meridional flow over the 7 days of the forecast. Despite this, the change in flow was estimated to be highly predictable by the ensemble forecast.

Finally, in Fig 19 we show the probability that the temperature anomaly either was greater than 4K, or was less than -4K. Consistent with the relatively weak ensemble dispersion, there is fairly strong agreement that over much of central Europe the probability of relatively cold temperatures is high. Similarly, over northern Scandinavia, the probability of relatively warm temperatures is also high. The two categories are not entirely exclusive; for example, over the Faroe Isles there is a small probability of both cold and warm temperature anomalies.

In general, this is a case where the European forecaster could be reasonably confident in a medium-range prediction. However, ensemble spread is not uniform over Europe. In Fig 20a we show 850 hPa

temperature probability for the ensemble from 13 November plumes for three locations: Longyearbyen (Spitzbergen), Paris and Funchal (Portugal). For comparison, the control forecast (solid) and verifying analysis (dashed) are also shown.

The three plumes indicate varying local predictability. For example the 30% contour breaks at day 5 in the first plume, day 9 for the second plume, and continues to the end of the forecast range for the third plume. As discussed in section 3, this measure of ensemble spread is generally correlated with the skill of the control forecast, and indeed for these examples the control and verifying trajectories are closest for the most predictable location.

5. CONCLUSIONS

The initial state of the atmosphere is known only approximately. A complete weather prediction must therefore be cast in terms of a probability distribution of forecast states. Ensemble prediction is a practical means of estimating this probability distribution for the medium range, where error evolution has started to become nonlinear.

Since initial error can project onto many possible phase space directions, some sampling procedure is necessary for the choice of initial perturbations for the ensemble forecast. Here we use the singular vectors calculated from forward and adjoint primitive-equation tangent models linearised about the short-range forecast trajectory. The calculation of singular vectors, described in the companion paper *Buizza and Palmer* (1994), allows one to estimate the initial directions in phase space that have maximum growth during the early parts of the forecast period.

In order to justify the use of these singular vectors as potential analysis perturbations, relationships between their structure and those obtained from analysis sensitivity studies were compared. The latter give the structure of perturbations to the analysis which, at some short-range forecast time, point along the gradient of the chosen energy norm for the forecast error.

To make realistic perturbations from the singular vectors, it was necessary firstly to find a set which covered a reasonable proportion of the northern hemisphere, secondly to ensure that a given perturbation was not too localised in physical space, and finally to determine the amplitude of the perturbations from the operational analysis error variance estimates.

33-member ensemble forecasts using the T63L19 version of the operational model have been made routinely since December 1993. Ensembles from the first year of experimental trials have been validated in terms

of contingency tables of spread/skill relationships and Briar scores of cluster probabilities. Distributions of ensemble-member scores have also been studied.

Results show a good relationship between ensemble skill and spread, which for one season was no worse than that obtained in a perfect model environment. On the other hand, there was evidence that model error contributed adversely at other times. In particular, underprediction of ensemble spread during cases where the verification developed strong ridging and/or cut-off low behaviour was shown to be consistent with model climatology problems.

On the other hand, it is likely that the initial perturbations chosen can be further improved, and this will lead to larger ensemble dispersion in the medium range. Firstly, according to BP, there may be many more unstable directions than are presented sampled. However, increased ensemble sizes may have to await more powerful computers. Secondly, with the advent of 4-dimensional data assimilation, constraints based on flow-dependent analysis error statistics could be incorporated more directly into the calculation of initial perturbations. In this way, instabilities developing not only during the forecast, but also during the data assimilation cycle, implicit in the breeding method of *Toth and Kalnay* (1993), can be incorporated into the initial perturbations.

Despite a couple of decades of theoretical work in the area of probabilistic weather prediction by dynamical methods, routine operational ensemble forecasting is new. There is much work to be done, not only in the construction of initial perturbations, but also in the development of user-oriented products, some of which were illustrated in two case studies. Initial reaction to ensemble forecasts have been generally favourable, both in the United States (*Tracton and Kalnay*, 1993), and in Europe, and it appears that ensemble prediction will become a routine and established part of the practice of weather forecasting.

Table 1: Characteristics of the SV computation

Horizontal res.	T21
Vertical res.	L19
OTI	36 hours
LPO area	*lat* ≥ 30° *N*
Iterations	100
no. of acceptable SVs	30:35
NNMI	5 gravest modes
Physics	surface drag and vertical diffusion

Table 2: Left hand column. Percentage of ensemble forecast where anomaly correlation of the most skilful member for day 7 over Europe was less than 0.6. For comparison, the percentage of operational forecasts whose skill was less than 0.6 is also shown.

	Ensemble	Operational forecast
WINTER	8	38
SPRING	28	56
SUMMER	5	49
AUTUMN	5	44

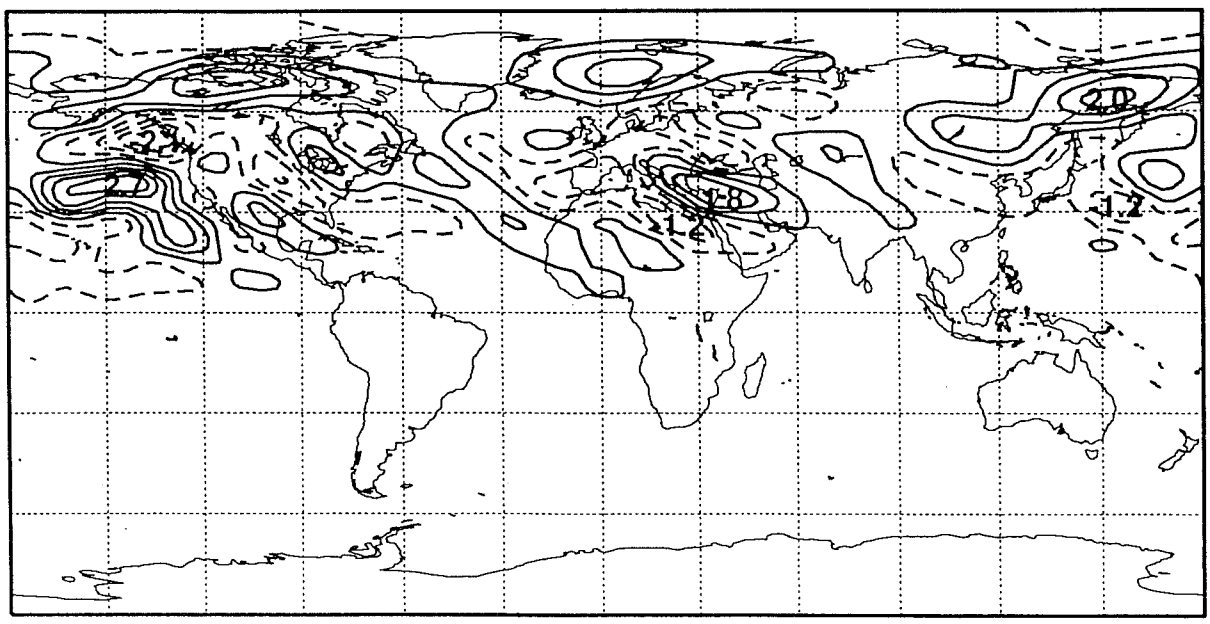


Fig 1 An example of an initial perturbation of zonal wind (contour interval 0.5 m/s) produced from the SV analysis of BP, with the scaling and phase-space rotation described in section 2c.

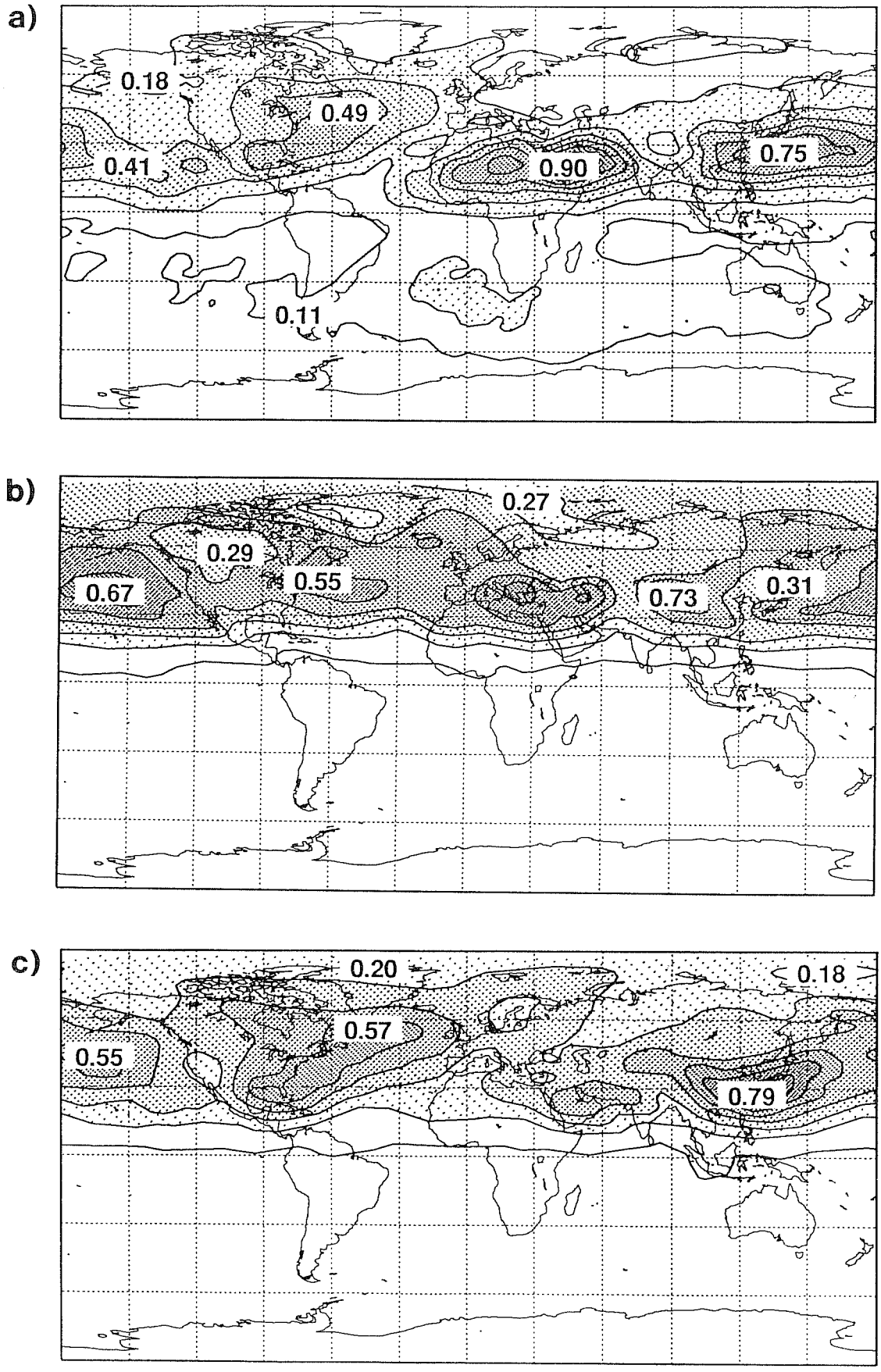


Fig 2 Rms amplitude of the u-wind component of the initial perturbations, at model levels 11 (contour interval: 0.1 m/s). a) winter 1992/93 b) summer 1993 c) autumn 1993.

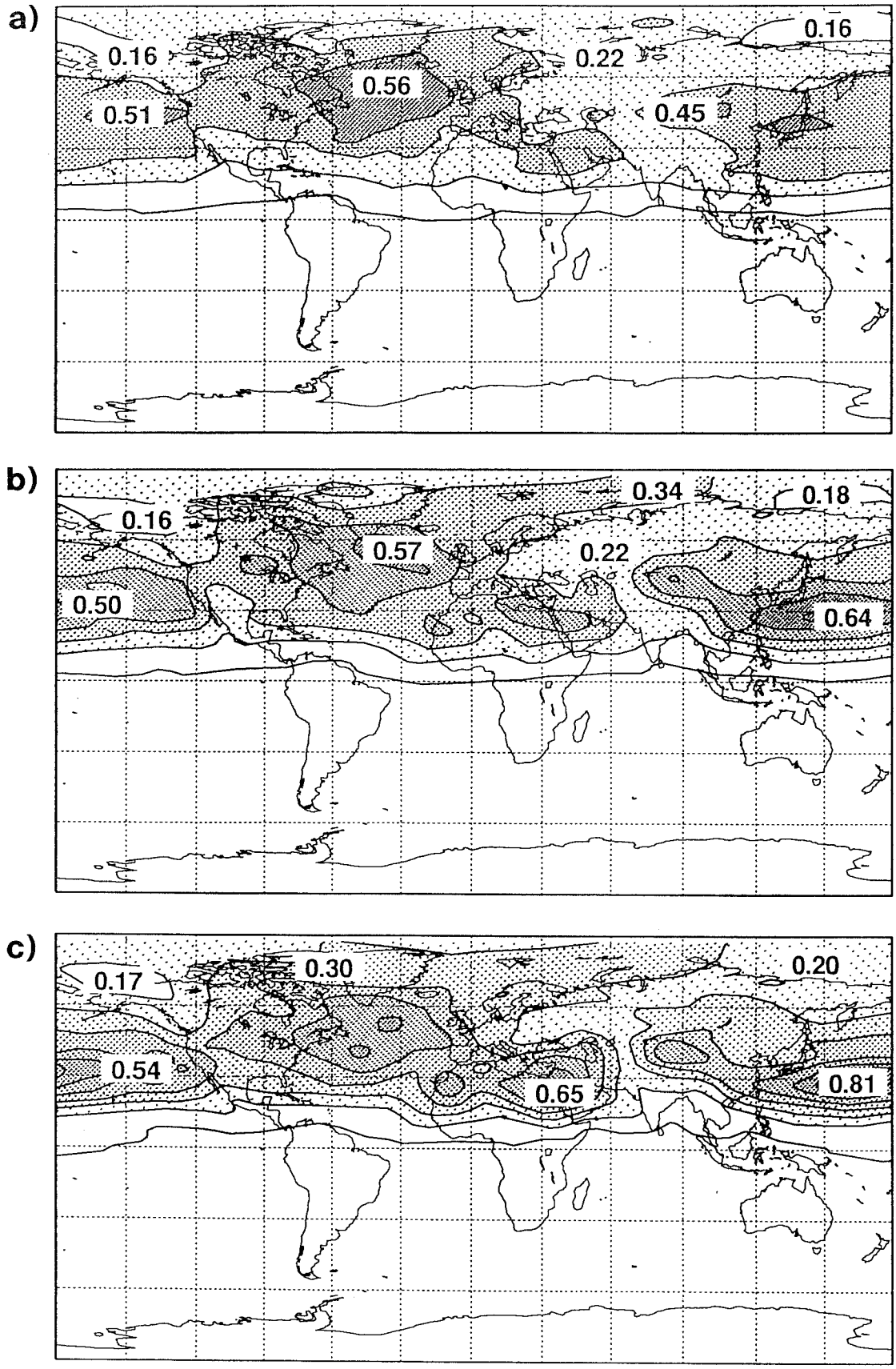


Fig 3 Rms amplitude of the u-wind component of the initial perturbations for spring at a) model level 9, b) model level 11, c) model level 13.

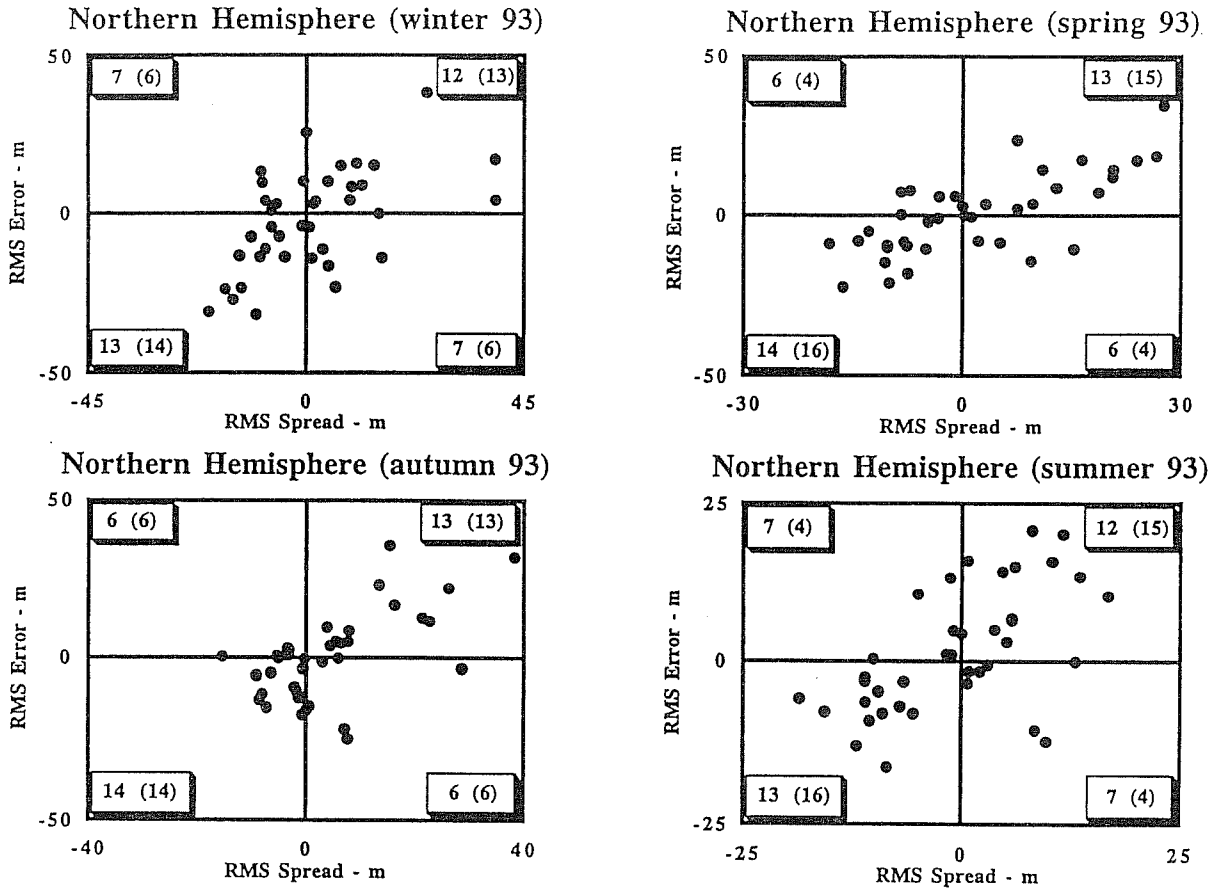


Fig 4 Scatter diagram between day 7 NH rms error of control forecast (ordinate) versus day 7 NH rms spread between ensemble members and control forecast (abscissa). Panel a) for winter 1992/93. Panel b) for spring 1993, c) for summer 1993 and d) for autumn 1993.

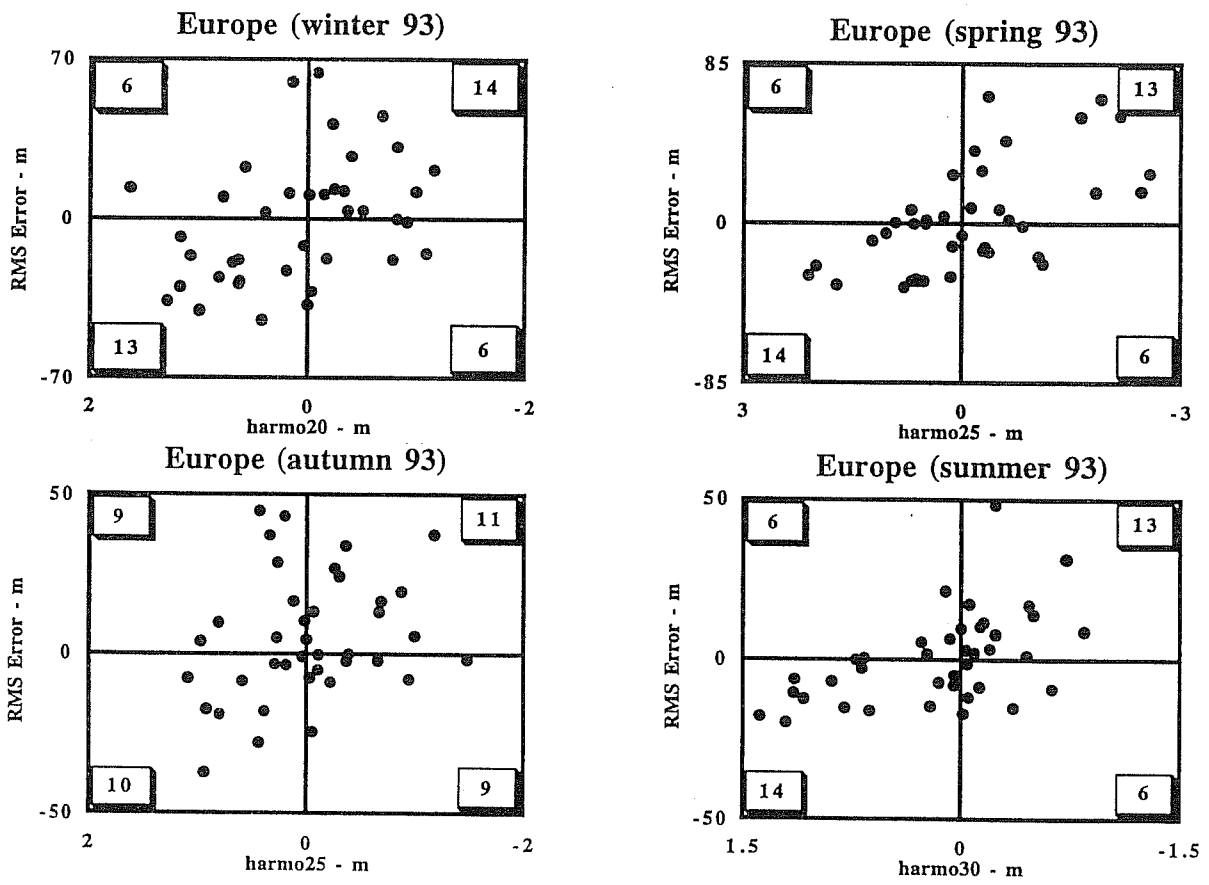


Fig 5 Scatter diagram between day 5 to day 7 mean European rms error of control forecast (abscissa) versus the forecast day value on which a chosen probability contour (of the 500 hPa height plumes - see text for details) first vanishes (ordinate). Panel a) winter 1993, b) spring 1993, c) summer 1993 and d) autumn 1993.

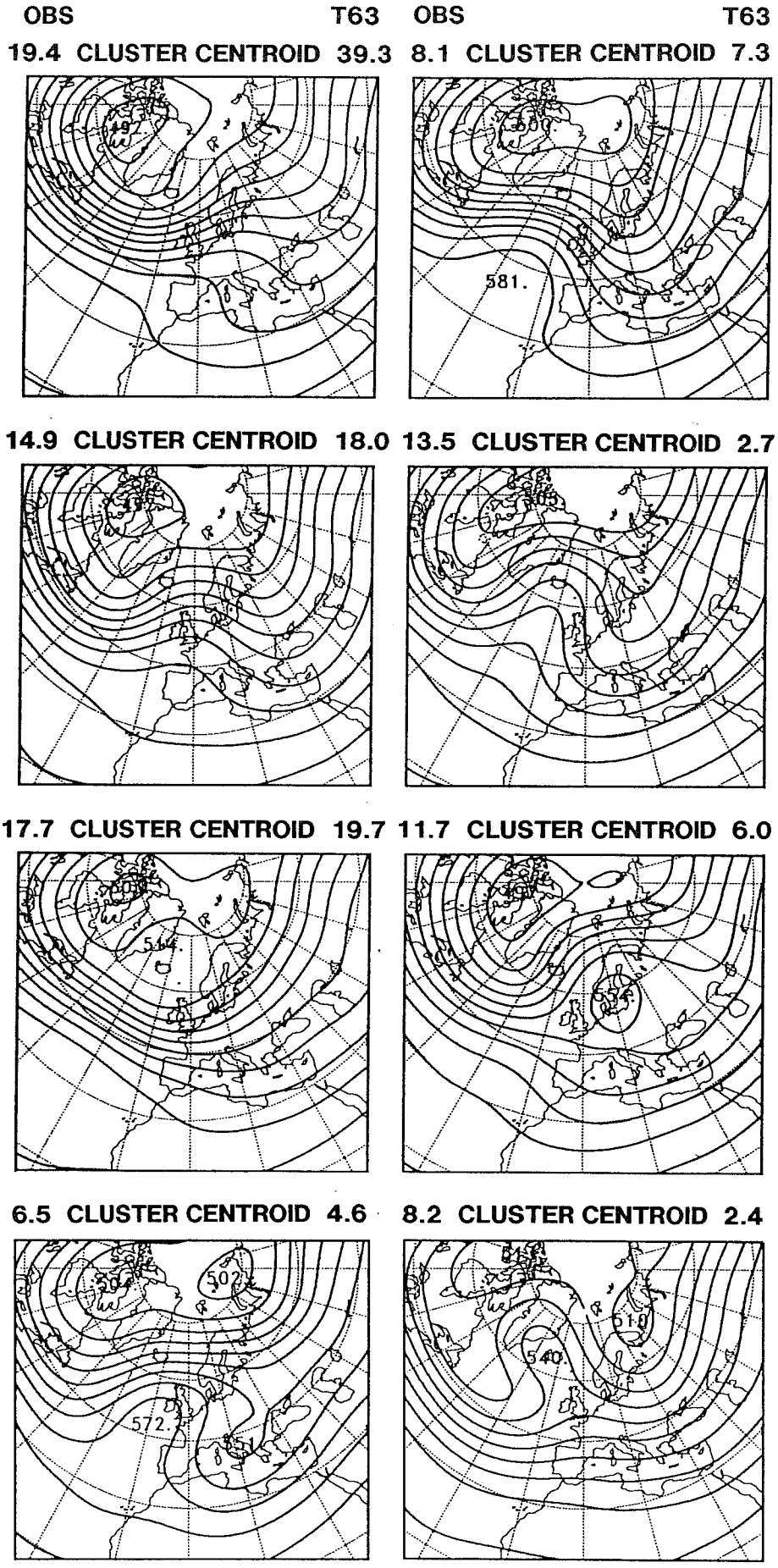


Fig 6 500 hPa height associated with 8 Euro-Atlantic cluster centroids derived from ECMWF analysis archives. The observed frequency is shown for each cluster at top left. Simulated frequencies based on 120-day 'climate' integrations of the EPS model are shown at top right.

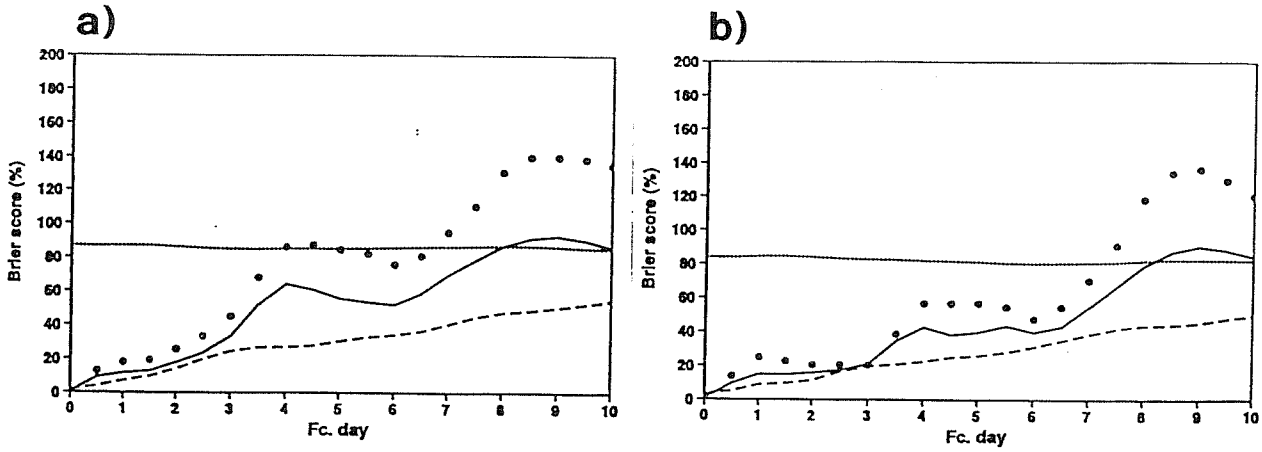


Fig 7 a) Briar score of winter ensembles (solid line), control forecast (large dotted line), climatology (small dotted line), a perfect-model ensemble (one member chosen randomly as verification; dashed line). b) as a) but excluding those ensembles in which the verifying analysis was classified in clusters where the model climatology significantly underestimated the observed frequency.

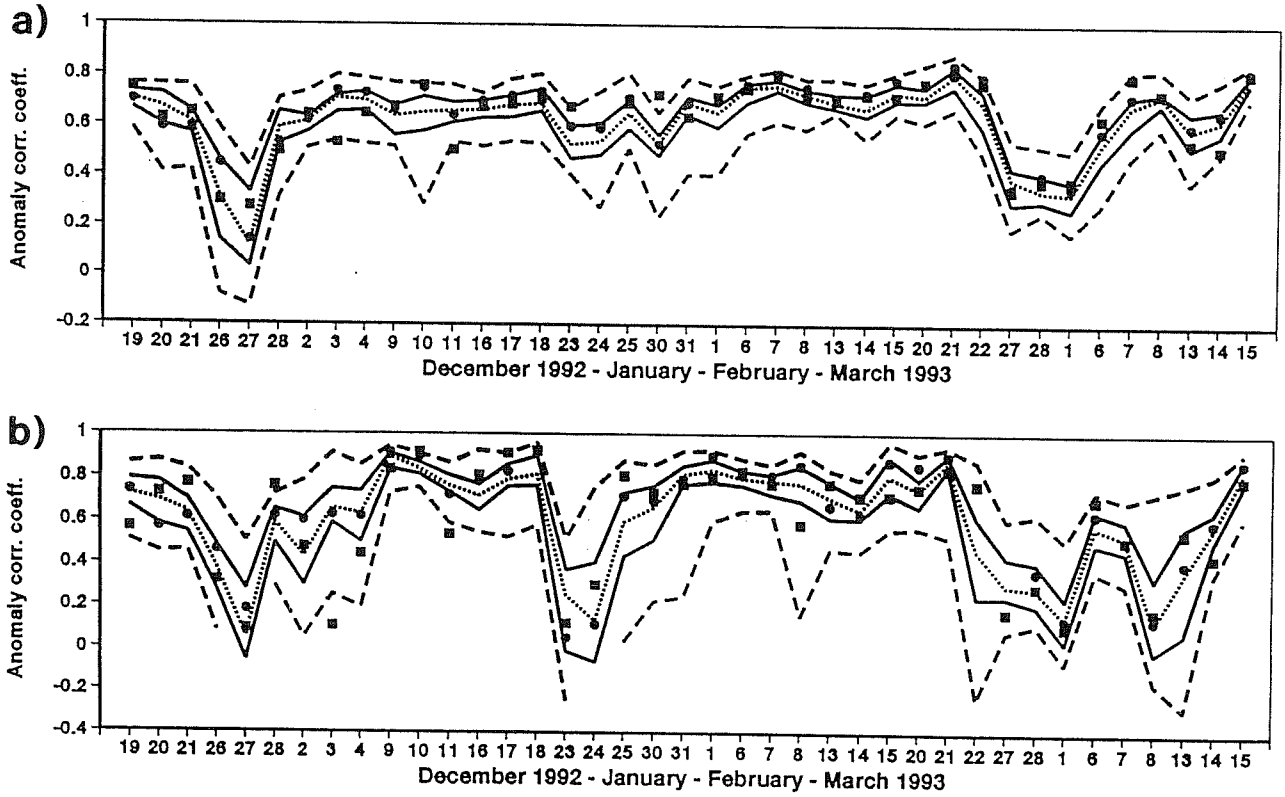


Fig 8 Time series of winter day 7 AC scores. Dashed lines - best and worst members of the ensemble. Solid lines - 25 and 75 percentiles of AC distribution. Dotted line - median of distribution. Solid circle - T63L19 control forecast. Solid square - T213L31 operational forecast. a) NH, b) Europe.

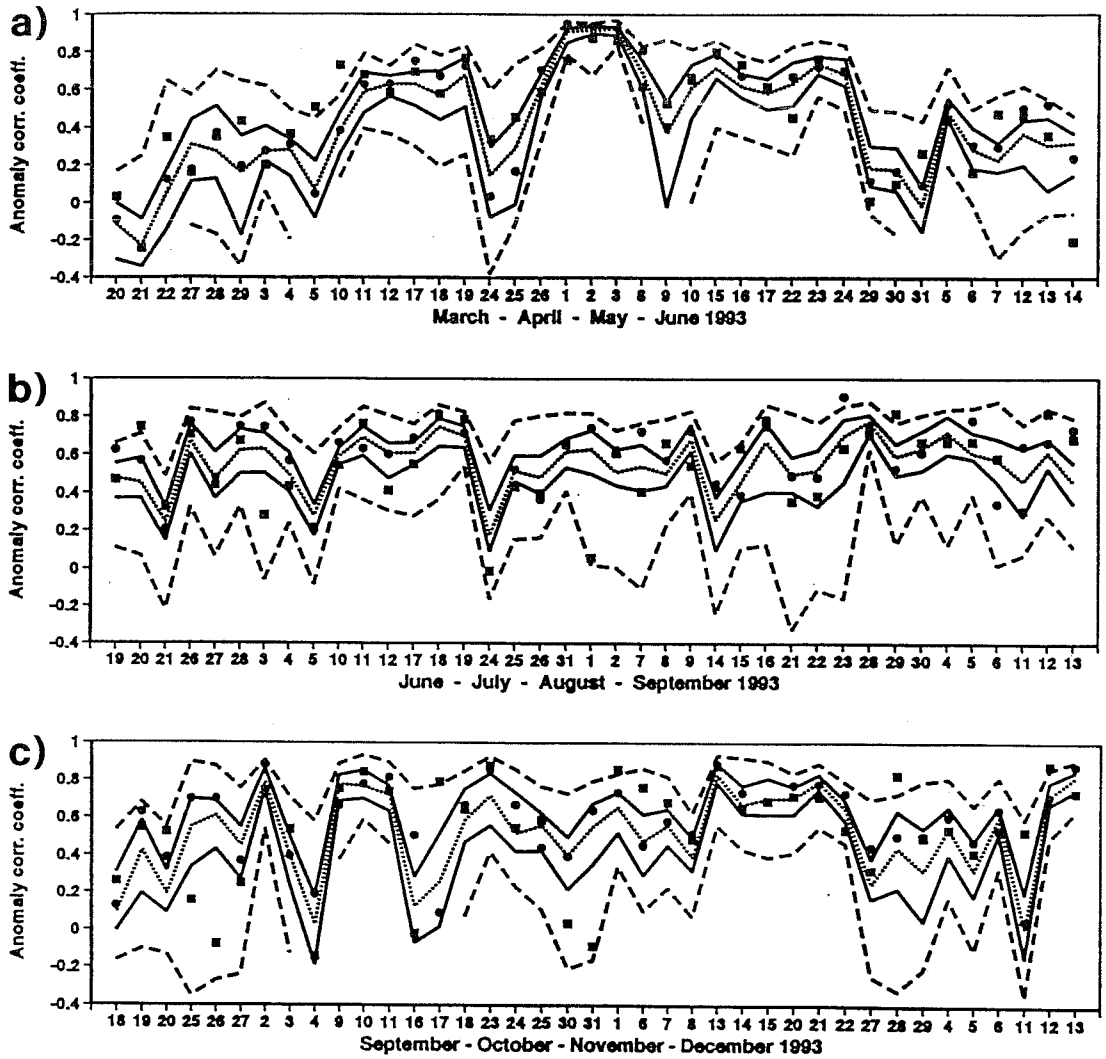


Fig 9 As Fig 8 but for Europe only a) spring, b) summer, c) autumn.

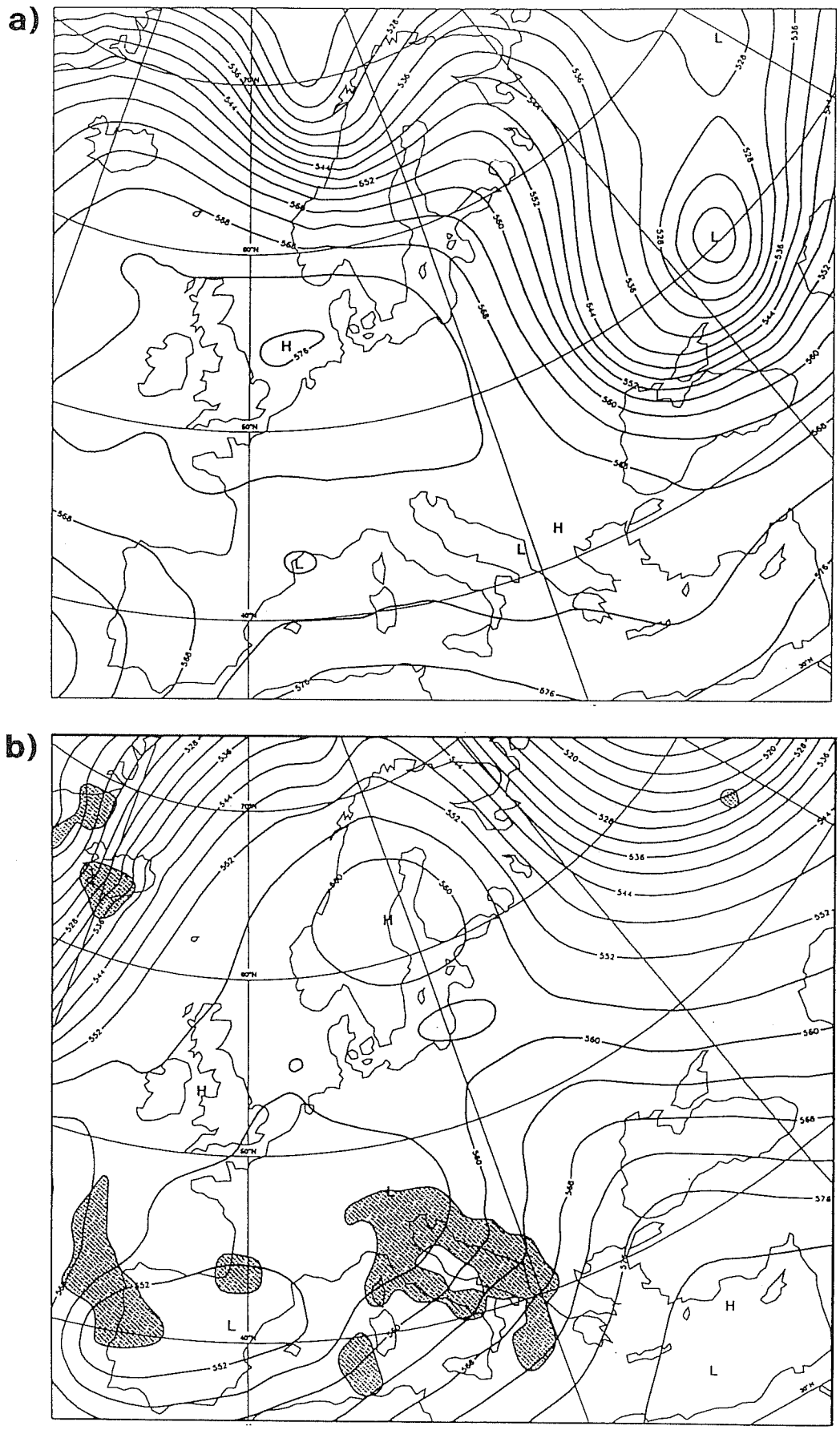


Fig 10 500 hPa height from ECMWF analyses for 12Z a) 30 October, b) 6 November. Superimposed on b) is shown region where 24 hour rainfall centred on 6 November exceeded 10 mm.

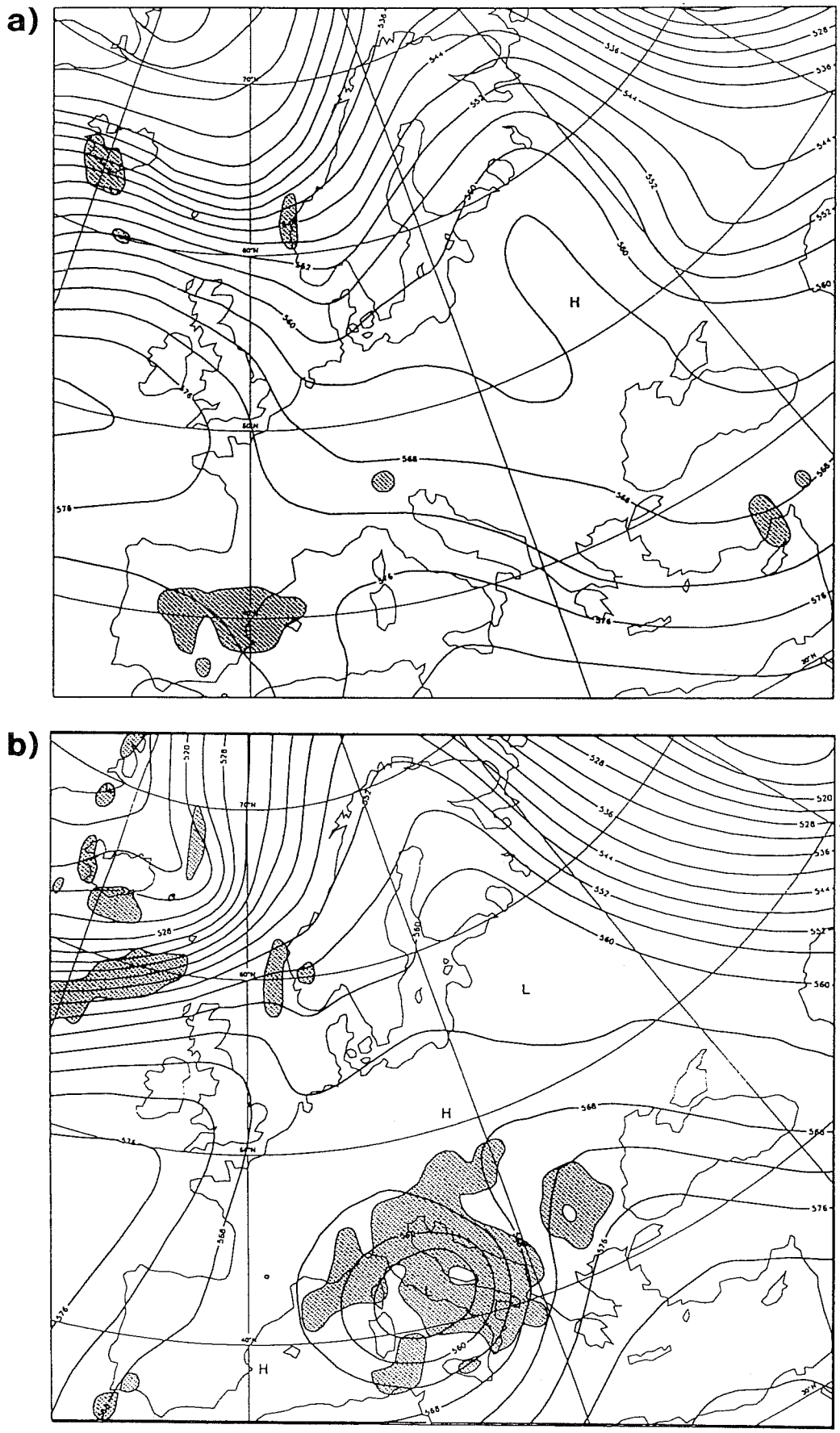
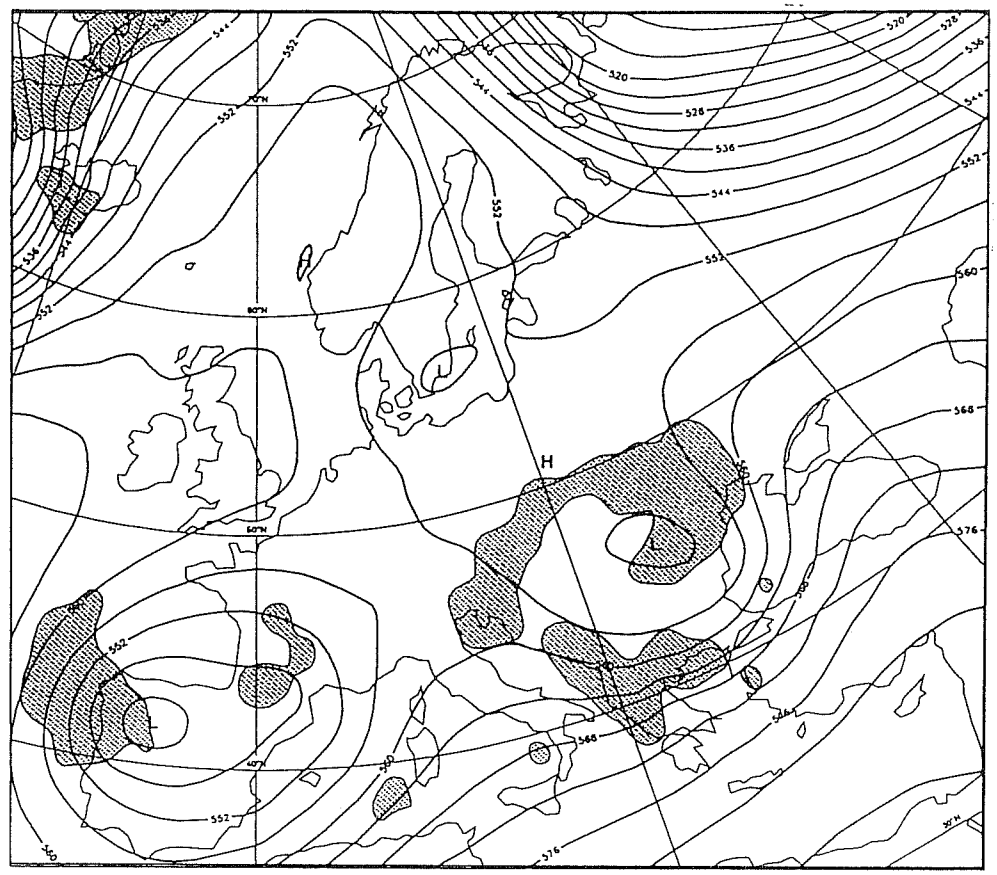


Fig 11 500 hPa height from operational forecasts verifying on 12Z 6 November. a) day 7, b) day 6, c) day 5. Superimposed is shown region where operational 24 hr forecast rainfall centred on 12Z 6 November exceeded 10 mm.

c)



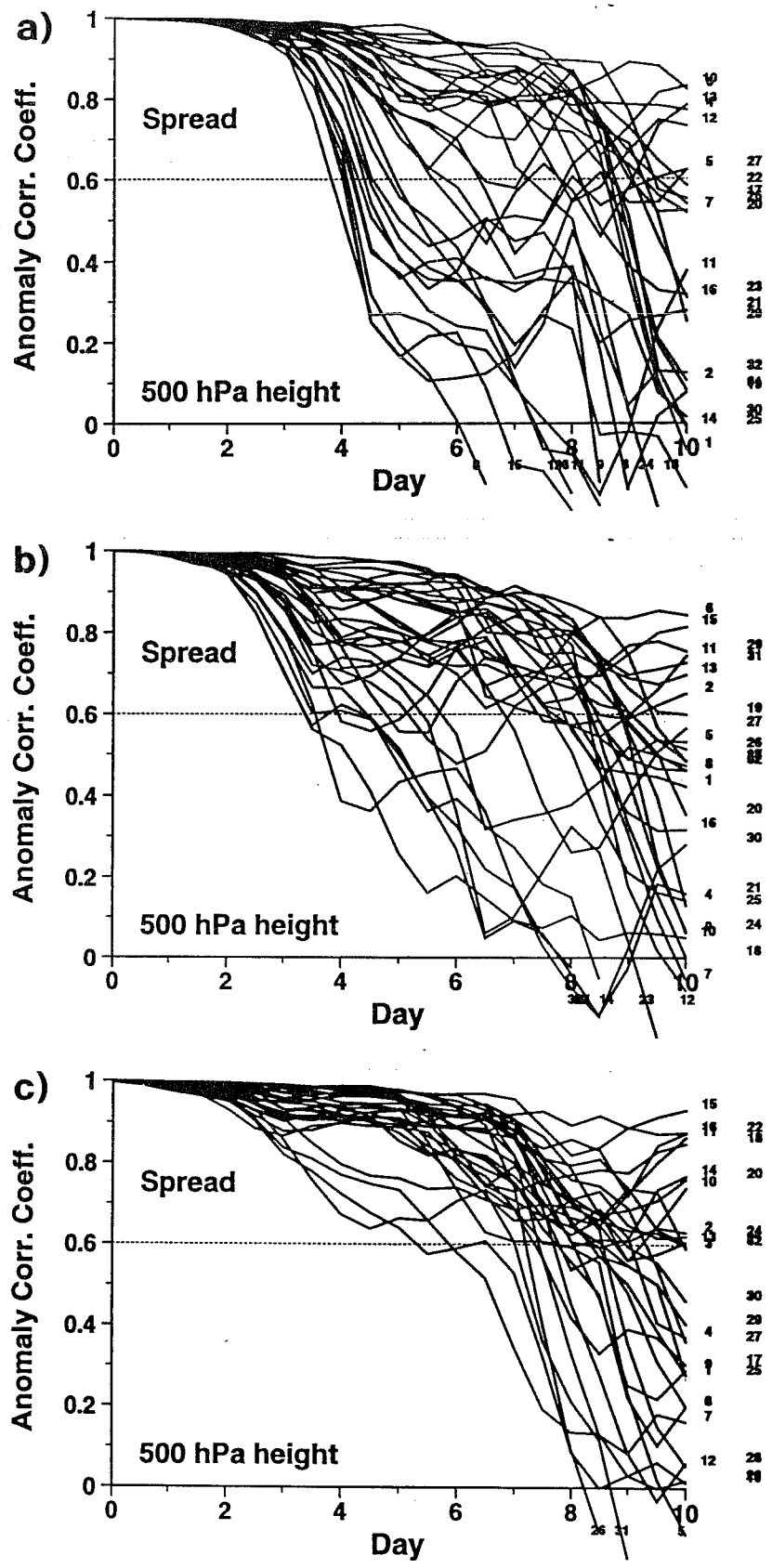


Fig 12 Anomaly correlation of 500 hPa height over Europe between individual ensemble members and control forecast as a function of forecast time; Panel a) ensemble from 30 October, b) ensemble from 31 October and c) ensemble from 1 November.

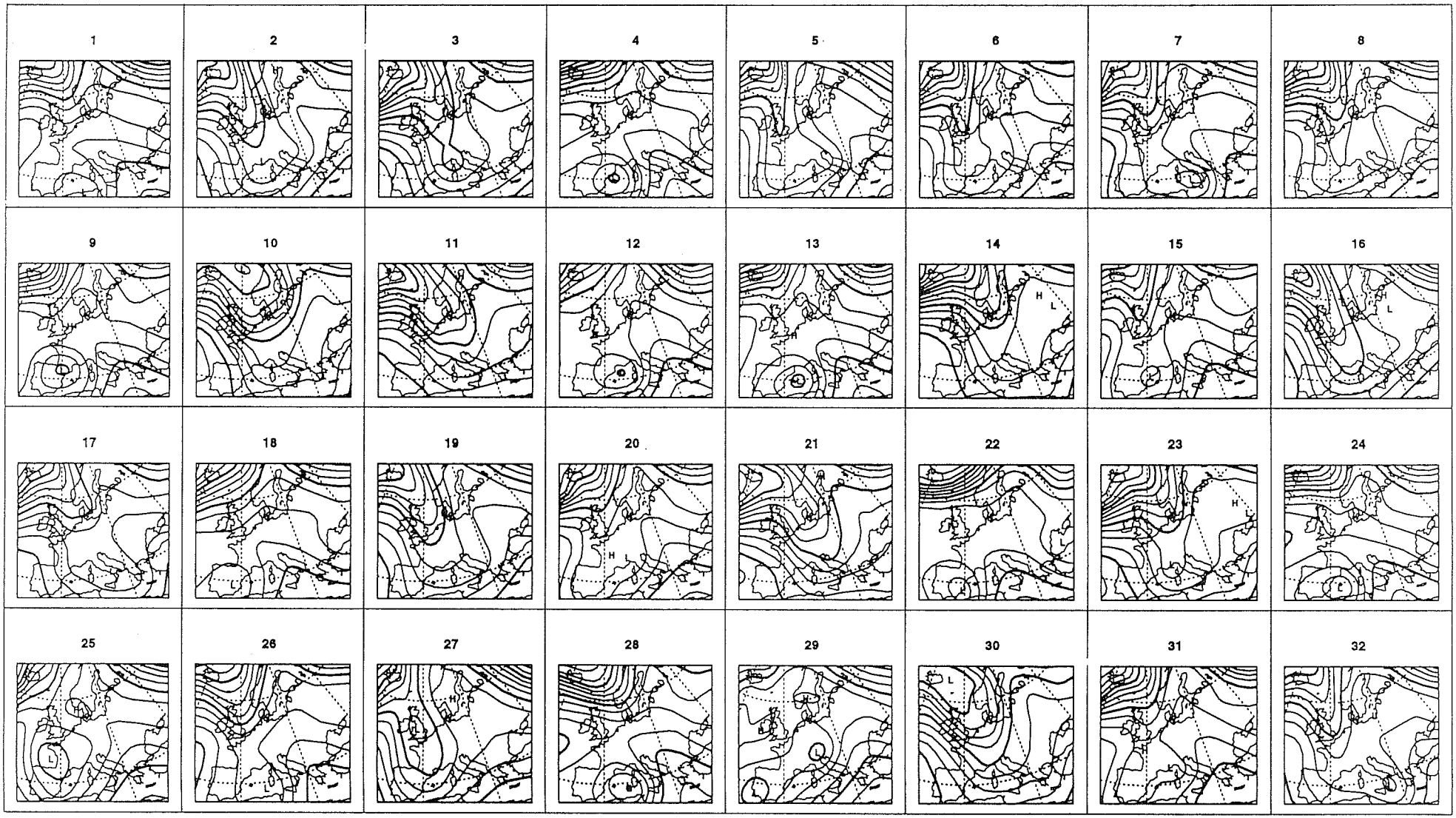


Fig 13 Individual forecast 500 hPa fields ("stamp maps") for day 6 ensemble valid for 6 November.

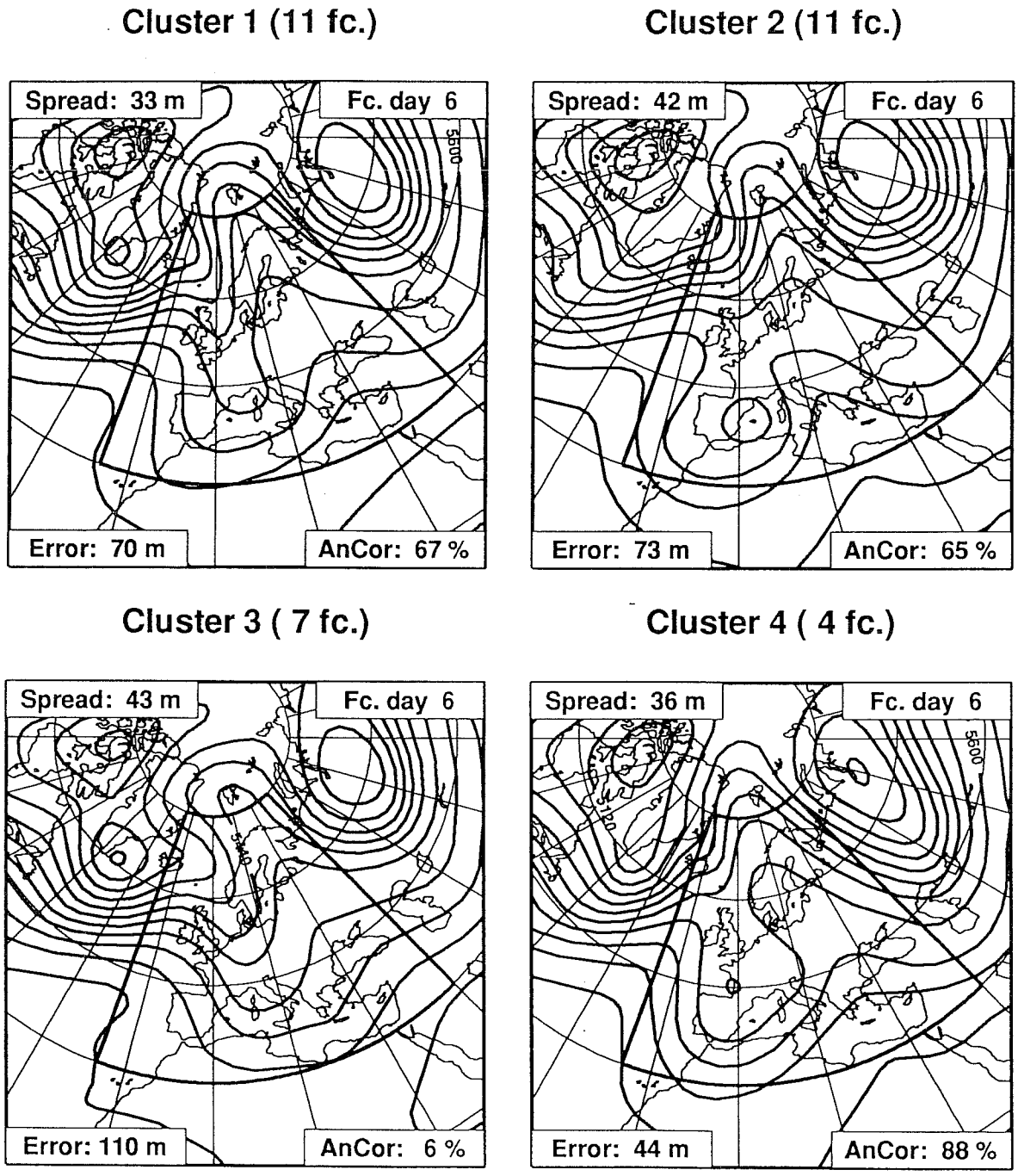


Fig 14 Clusters of 500 hPa height field for day 6 ensemble valid for 6 November.

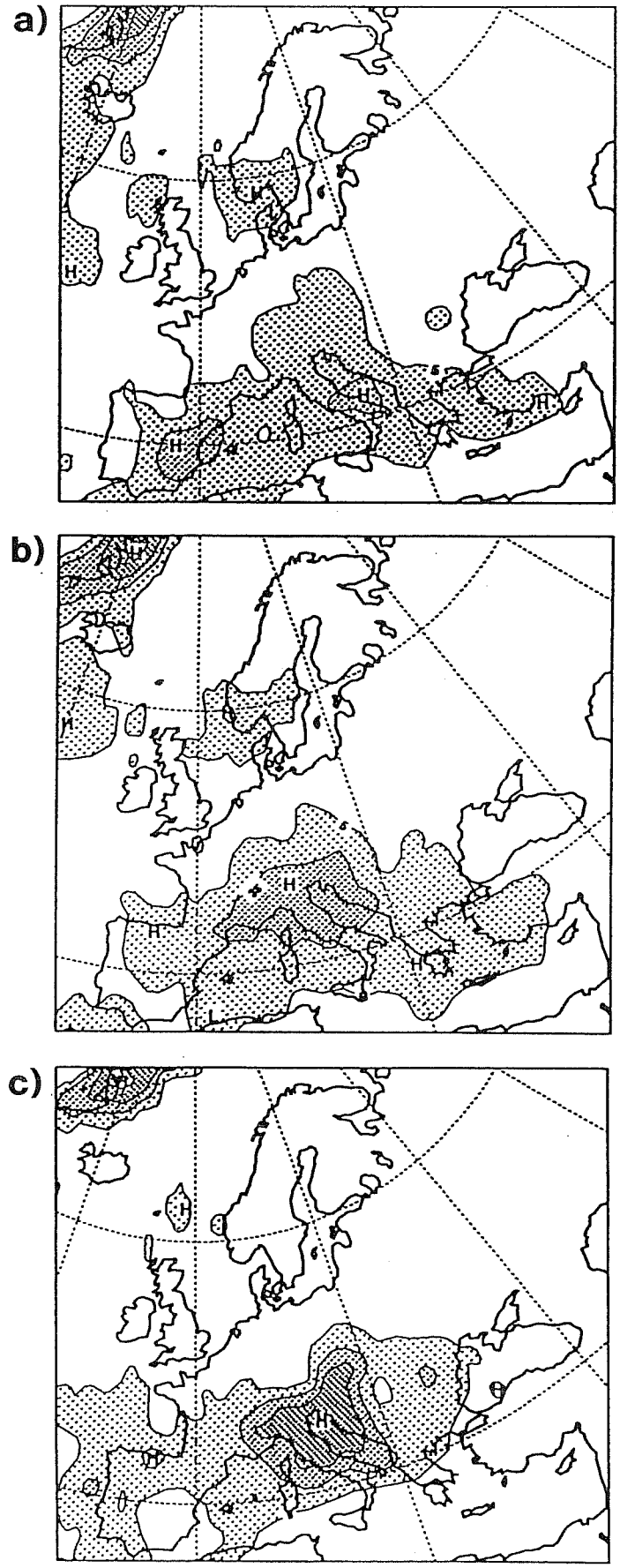


Fig 15 Maps of probabilities that total precipitation exceeds 10 mm/day between 5 November 12UTC and 6 November 12UTC, from the ensemble originated from a) 30 October, b) 31 October and c) 1 November. Contours 5, 35, 65, 95%.

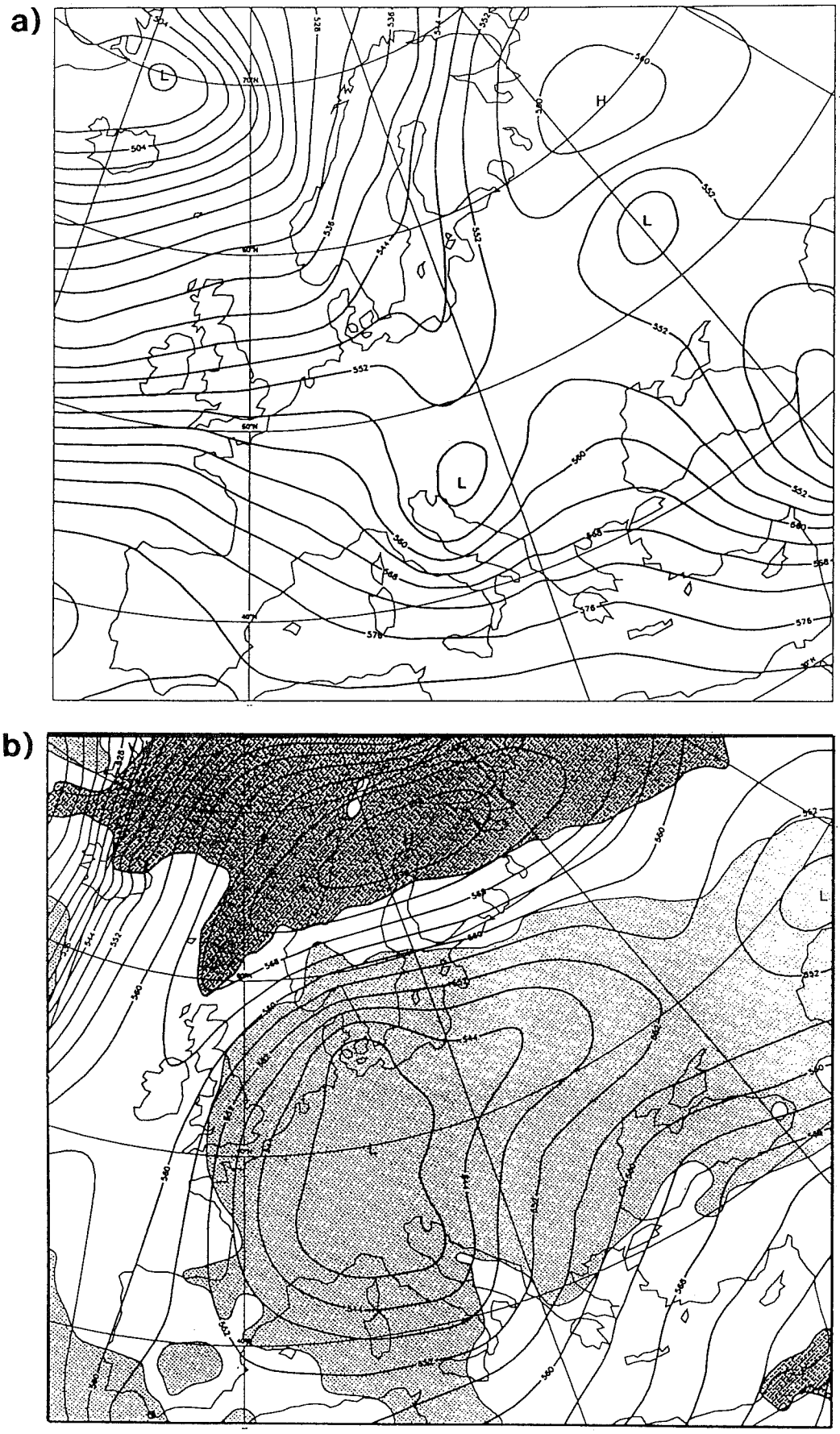


Fig 16 Analysed 500 hPa height fields for a) 13 November b) 20 November. Superimposed on b) are regions (shaded) where analysed temperature anomaly of 850 hPa was greater than 4K (heavy shading) or less than -4K (light shading).

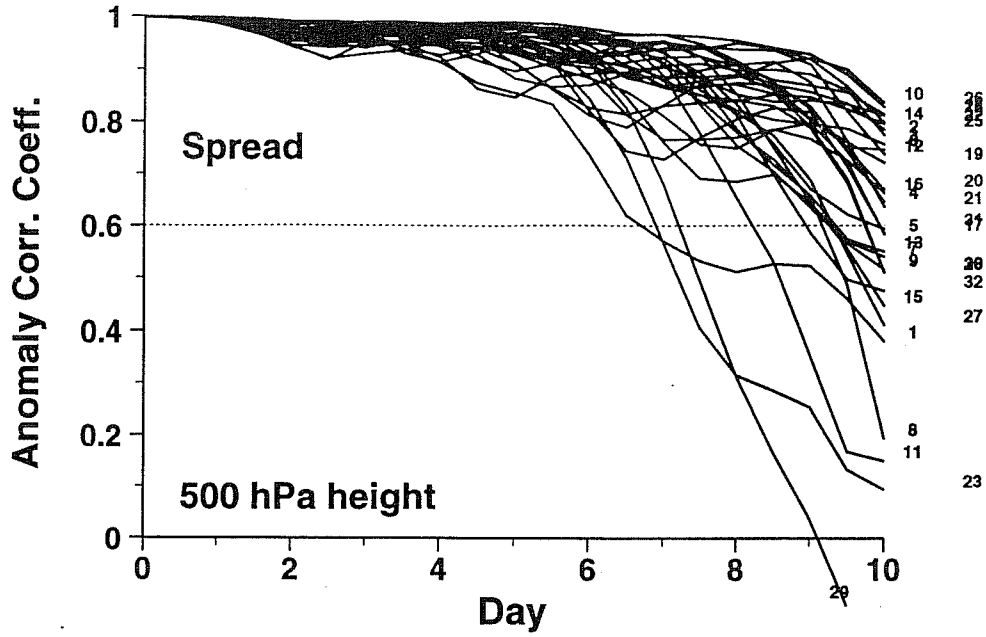
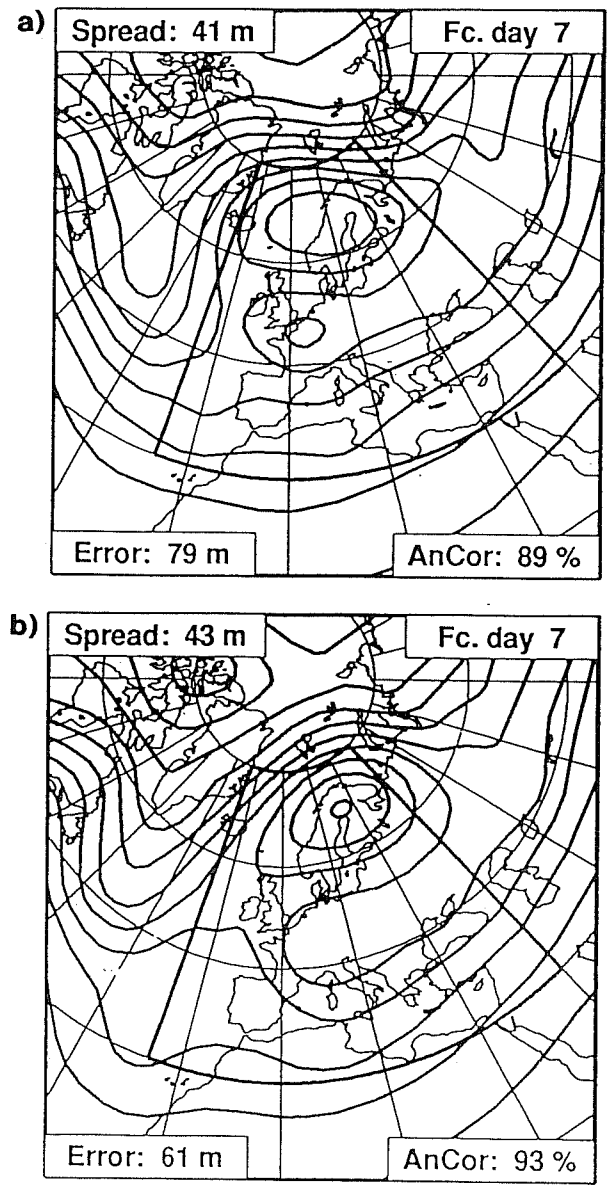


Fig 17 Anomaly correlation of 500 hPa height over Europe between individual ensemble members and control forecast as a function of forecast time. Ensemble from 13 November.



Cluster 1 (8fc.)

Cluster 2 (8 fc.)

Fig 18 The dominant 2 clusters of 500 hPa height field for day 7 ensemble valid for 20 November.

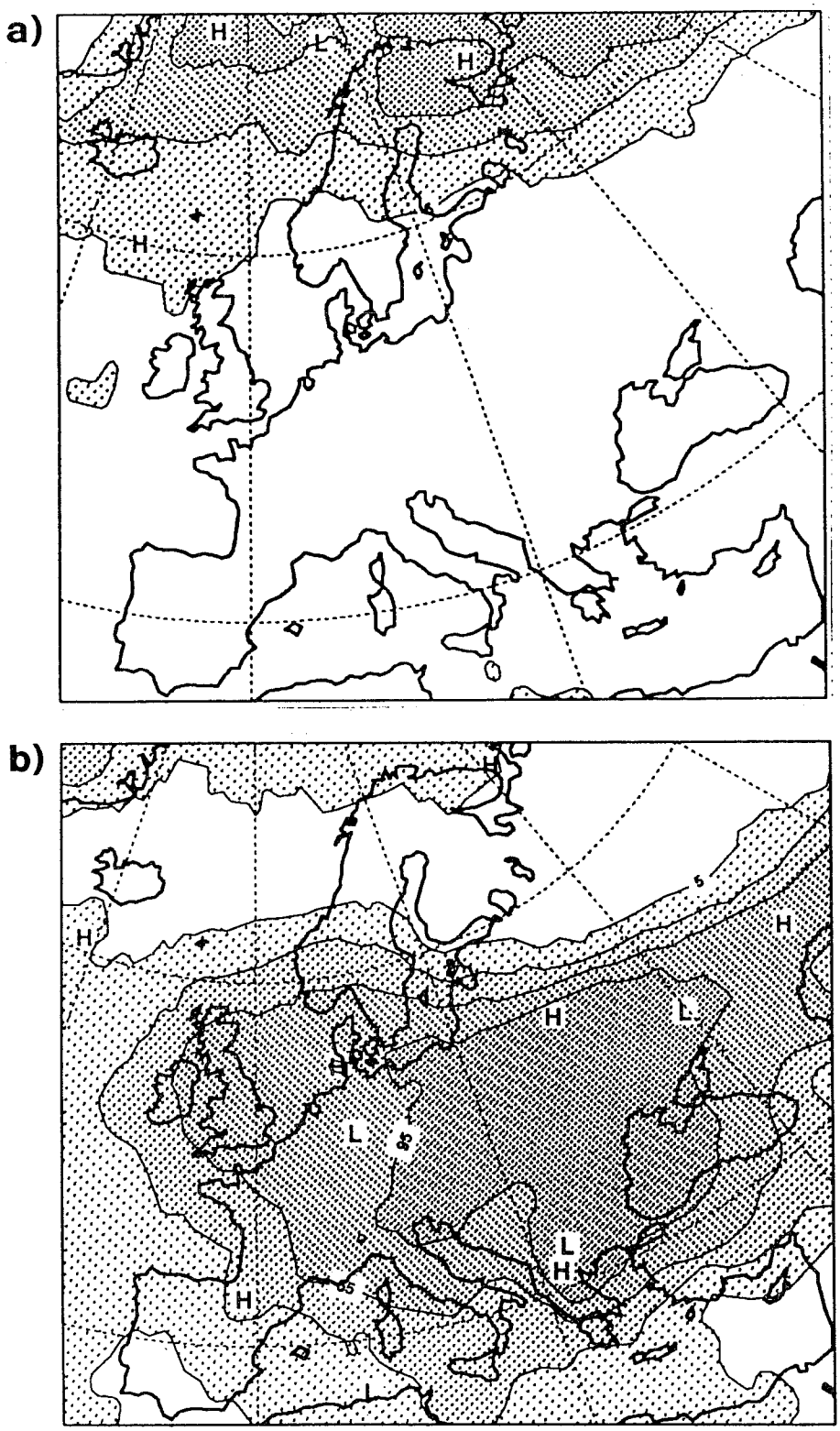


Fig 19 Maps of probabilities that 850 hPa forecast temperature anomaly for 20 November a) exceeds 4K, b) is less than -4K, from the ensemble initialised on 13 November. Contours 5, 35, 65, 95%.

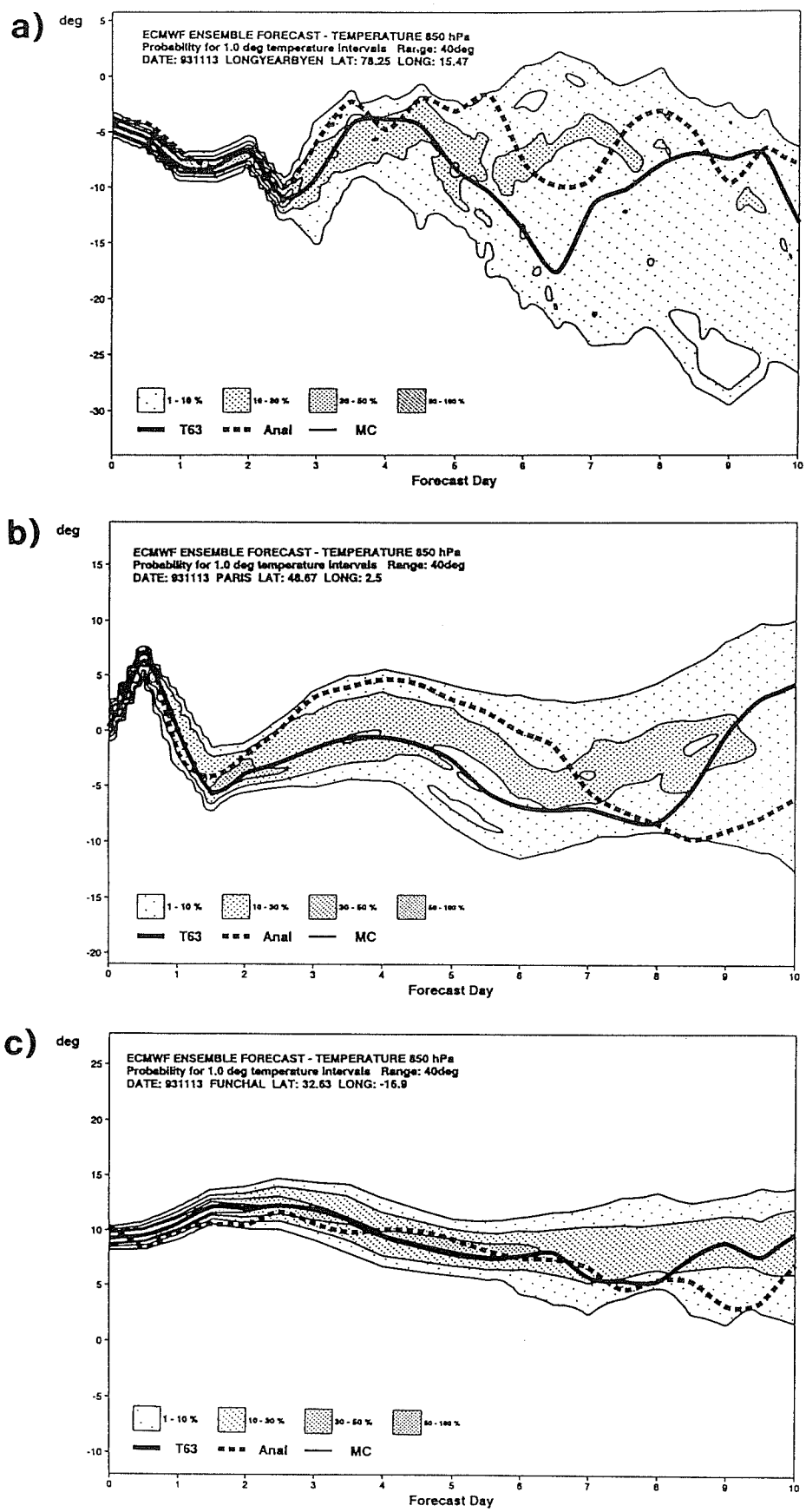


Fig 20 850 mb forecast temperature probability "plumes" for the ensemble forecast from 13 November 1993 a) Longyearbyen (Spitzbergen), b) Paris, c) Funchal (Portugal). Probabilities are based on 1 K intervals (see section 2e(iii) for details).

References

Abarbanel, H D I, R Brown and M B Kennel, 1991: Variation of Lyapunov exponents on a strange attractor. J Nonlinear Sci, 1, 175-199.

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

Borges, M D and D L Hartmann, 1992: Barotropic instability and optimal perturbations of observed non-zonal flows. J Atmos Sci, 49, 335-354.

Brankovic, C, T N Palmer and L Ferranti, 1994: Predictability of seasonal atmospheric variations. J Clim, 7, 218-237.

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, 1994a: Sensitivity of optimal unstable structures. Q J R Meteor Soc, 120, 429-451.

Buizza, R, 1994b: Location of optimal perturbations using a projection operator. Q J R Meteorol Soc. In print.

Buizza, R, J Tribbia, F Molteni and T N Palmer, 1993: Computation of optimal unstable structures for a numerical weather prediction model. Tellus, 45A, 388-407.

Buizza, R and T N Palmer, 1994: The singular-vector structure of the atmospheric general circulation. J Atmos Sci, submitted.

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

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

Hoffman, R N and E Kalnay, 1983: Lagged average forecasting, an alternative to Monte Carlo forecasting. Tellus, 35A, 100-118.

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

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.

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

McIntyre, M E, 1988: Numerical Weather Prediction: A vision of the future. Weather, 43, 294-298.

Molteni, F and T N Palmer, 1991: A real-time scheme for the prediction of forecast skill. Mon Wea Rev, 119, 1088-1097.

Molteni, F and T N Palmer, 1993: Predictability and finite-time instability of the northern winter circulation. Q J R Meteor Soc, 119, 269-298.

Mureau, R, F Molteni and T N Palmer, 1993: Ensemble prediction using dynamically-conditioned perturbations. Q J R Meteor Soc, 119, 299-323.

Murphy, J M, 1988: The impact of ensemble forecasts on predictability. Q J R Meteorol Soc, 114, 463-493.

Palmer, T N, F Molteni, R Mureau, R Buizza, P Chapelet and J Tribbia, 1993: Ensemble prediction. ECMWF seminar proceedings "Validation of models over Europe: Vol 1", ECMWF, Shinfield Park, UK, 285 pp.

Rabier, R, P Courtier, M Herveou, B Strauss and A Persson, 1993: Sensitivity of forecast error to initial conditions using the adjoint model. ECMWF Research Department Technical Memorandum No. 197.

Rabier, R, E Klinker, P Courtier, B Strauss and M Herveou, 1994: Sensitivity of forecast error to initial conditions using the adjoint technique. Extended abstract for 10th AMS conference on Numerical Weather Prediction.

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

Thépaut, J-N, R N Hoffman and P Courtier, 1993: Interactions of dynamics and observations in a four-dimensional variational assimilation. Mon Wea Rev, 121, 3393-3414.

Toth, Z and E Kalnay, 1993: Ensemble forecasting at NMC: the generation of perturbations. Bull Am Met Soc, 74, 2317-2330.

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