Q. J. R. Meteorol. Soc. (1996), **122**, pp. 73–119

The ECMWF Ensemble Prediction System: Methodology and validation

By F. MOLTENI, R. BUIZZA, T. N. PALMER* and T. PETROLIAGIS
European Centre for Medium-Range Weather Forecasts, UK

(Received 1 August 1994; revised 24 May 1995)

SUMMARY

The European Centre for Medium-Range Weather Forecasts (ECMWF) Ensemble Prediction System (EPS) is described. In addition to an unperturbed (control) forecast, each ensemble comprises 32 10-day forecasts starting from initial conditions in which dynamically defined perturbations have been added to the operational analysis. The perturbations are constructed from singular vectors of a time-evolution operator linearized around the short-range-forecast trajectory. These singular vectors approximately determine the most unstable phase-space directions in the early part of the forecast period, and are estimated using a forward and adjoint linear version of the ECMWF numerical weather-prediction model. An appropriate norm is chosen, and 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. A methodology to perform a phase-space rotation of the singular vectors is described, which generates hemispheric-wide perturbations and renormalizes them according to analysis-error estimates from the data-assimilation system.

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; no significant differences are found in some cases. Brier scores for the probability of European flow clusters are presented, which indicate predictive skill up to forecast-day 8 with respect to climatological probabilities. The dependence of these scores on flow-dependent model errors is also discussed. Finally, ensemble-member skill-score distributions are presented, which confirm the overall satisfactory performance of the EPS, particularly in summer and autumn 1993. In winter, cases of poor performance over Europe were associated with the occurrence of a split westerly flow with a blocking high and/or a cut-off low in the verifying analysis.

Two cases are studied in detail, one having large ensemble dispersion, the other corresponding to a more predictable situation. 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.

KEYWORDS: Ensemble prediction Forecasting skill Medium-range forecasts Singular vectors

1. INTRODUCTION

In December 1992, both the US National Meteorological Center (NMC) and the European Centre for Medium-Range Weather Forecasts (ECMWF) began to produce and disseminate medium-range ensemble predictions (Tracton and Kalnay 1993; Palmer *et al.* 1993). This important development is the result of more than two decades of theoretical research and numerical experimentation, following the work of Epstein (1969), Gleeson (1970), Fleming (1971a, b) and Leith (1974) who 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 continuity equation for probability, also known as the Liouville equation (e.g. Gleeson 1970; Ehrendorfer 1994), its practical solution is impossible for nonlinear models with more than a few degrees of freedom. Even restricting 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 forecasting, with a system of nonlinear equations which have no well-defined closure and which cannot be solved for the large models currently used in numerical weather prediction (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 linearized dynamics. In ensemble forecasting, the PDF at initial time is represented through a finite

* Corresponding author: ECMWF, Shinfield Park, Reading, Berkshire RG2 9AX, UK.

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.

The ensemble statistics will approximate the correct PDF if:

(i) the sample of initial states provides a realistic estimate of the probability distribution of analysis errors; and
(ii) the phase-space trajectories computed by the numerical model are good approximations of atmospheric trajectories.

Requirement (ii) is also necessary for 'deterministic' NWP; hence, most of the recent research in ensemble forecasting has focused on point (i). However, as discussed later, 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 (i) 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 orders 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 than, but comparable with, 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 several NWP centres (Murphy 1988, 1990; Tracton *et al.* 1989; Brankovic *et al.* 1990; Déqué and Royer 1992). A 'scaled LAF' in which perturbations are considered equally likely has been formulated by Ebisuzaki and Kalnay (1991).

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 vectors associated with the largest Lyapunov exponents of the NWP model. The NMC ensemble prediction scheme is based on a combination of the breeding and LAF techniques (Tracton and Kalnay 1993), and at least partially satisfies conditions (i) and (ii) above.

However, even assuming an isotropic PDF in phase space for the error at the initial time, the different amplification rates of perturbations along different axes would soon stretch the PDF 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 propagation itself. SV 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), Farrell (1990), Borges and Hartmann (1992), and Molteni and Palmer (1993).

Ensemble forecasting experiments in which unstable SVs computed from a 3-level quasi-geostrophic model were used to construct initial perturbations for a multilevel primitive-equation model were carried out at the ECMWF in the past four years, and have been reported by Mureau *et al.* (1993), and Palmer *et al.* (1993). This approach proved to be more successful than alternative ensemble techniques tested at the ECMWF. However, the inconsistency between the vertical coordinates of the quasi-geostrophic and primitive-equation model created difficulties in the vertical interpolation over high topography. Although this problem did not affect the strongly unstable SVs localized on the western side of the oceans, continental features often had a smaller growth rate in the primitive-equation than in the quasi-geostrophic model (see Mureau *et al.* 1993). Efforts were, therefore, directed towards the computation of SVs in a simplified primitive-equation environment, using an iterative Lanczos algorithm for the solution of the eigenvector problem. Preliminary results on this experimentation were reported by Buizza *et al.* (1993); a more comprehensive description of the structure and dynamical properties of primitive equation SVs can be found in a companion paper by Buizza and Palmer (1995), hereafter referred to as BP.

SV structures are dependent on the choice of inner product. In relating these structures to analysis error we give arguments to suggest that a suitable inner product can be defined in terms of perturbation energy. 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. (Note that 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.) As briefly discussed in section 2, preliminary comparisons suggest that the structures of these forecast sensitivity fields share much of the same dynamical features of the SVs.

From December 1992 to April 1994 the ECMWF produced and disseminated real-time ensemble forecasts on an experimental basis on each Saturday, Sunday and Monday.

From May 1994 ensembles have been run daily. 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 the ECMWF's Member States.

This paper describes aspects of the development of the ECMWF's Ensemble Prediction System (EPS) and the results of the first year of its experimental 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, characterized 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

Several aspects of the SVs approach to ensemble predictions are discussed in this section: definition and properties of the SVs, choice of inner product, relationship of SVs to analysis error, definition of the initial perturbations, statistical properties of ensemble perturbations and ensemble products.

(*a*) *Definition and properties of singular vectors*

The mathematical definition and the dynamical properties of the fastest-growing SVs of baroclinic models have already been extensively documented in previous studies (e.g. Farrell 1990; Molteni and Palmer 1993; BP) and will be only briefly summarized here.

Let $\mathrm{L_P}(t, t_0)$ be the integral propagator of the dynamical equations, linearized about a portion of the nonlinear trajectory of the same dynamical system, so that

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

maps a small perturbation at time t_0, along the trajectory, to a small perturbation at future time t.

Let $\mathrm{L_P^*}$ be the adjoint operator of $\mathrm{L_P}$ with respect to a total energy inner product $\langle \ldots ; \ldots \rangle_e$. Then the total energy of a perturbation at time t is

$$\|x'(t)\|_e^2 = \langle x'(t); x'(t)\rangle_e \tag{2}$$

so that

$$\|x'(t)\|_e^2 = \langle x'(t_0); \mathrm{L_P^*L_P}x'(t_0)\rangle_e. \tag{3}$$

The reason for choosing this inner product is discussed in section 2(b).

In this specific case the propagator $\mathrm{L_P}$ is a T21L19 linear version of the Integrated Forecasting System model developed at the ECMWF and Météo-France (Courtier *et al.* 1991). It is compounded from the action of the linearized versions of the nonlinear normal-mode initialization procedure, the adiabatic part of the model equations and some physical parametrizations, namely 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 maximized for the northern extratropics. The action of the local projection

TABLE 1. CHARACTERISTICS OF THE SINGULAR VECTOR (SV) COMPUTATION

Horizontal resolution	T21
Vertical resolution	L19
Optimization time interval	36 hours
Local projection area	Latitude $\geqslant 30°$N
Iterations	100
No. of acceptable SVs	30:35
Normal-mode initialization	Five gravest modes
Physics	Surface drag and vertical diffusion

operator T (see section 2.4 of BP for more details) allows one to calculate perturbations with maximum amplitude at final time t over a prescribed area; in our case the northern hemisphere from 30°N to the pole. Hence, introducing the symmetric operator T in Eqs. (1) and (2), the energy in the selected region is $\langle x'(t_0);\, \mathrm{K}x'(t_0)\rangle_e$ where

$$\mathrm{K} = \mathrm{L}_P^* \mathrm{T}^2 \mathrm{L}_P. \tag{4}$$

Denote by $v_i(t_0)$ a normalized 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 $v_i(t_0)$, it follows

$$\max_{x'(t_0)\neq 0} \frac{\langle x'(t_0);\, \mathrm{K}x'(t_0)\rangle_e}{\langle x'(t_0);\, x'(t_0)\rangle_e} = \sigma_1^2. \tag{5}$$

The $v_i(t_0)$ and the σ_i are called, respectively, the singular vectors and the singular values of the operator TL_P, while $(t - t_0)$ is called the optimization time interval.

In the Ensemble Prediction System, SVs are computed applying an iterative Lanczos procedure (e.g. Strang 1986) to the linear propagator defined above. Usually 100 iterations of the algorithm are enough to give about 30 SVs with sufficient numerical accuracy. Table 1 summarizes the principal characteristics of the SV computation.

In the studies referenced above, the evolution of the SVs from $v_i(t_0)$ to $v_i(t)$ was found to be particularly non-modal. In particular it was noted:

(i) There was an upscale energy transformation between initial and final time. BP found that at initial time perturbation energy peaked at about wave number 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 wave number 25 for optimization time intervals of about 3 days. At optimization time, SV energy peaked at about wave number 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.

(ii) SV energy peaked mainly in the lower troposphere (approximately the baroclinic steering level) at initial time and at jet-stream level at optimization time. 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 localized 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.

(iii) 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

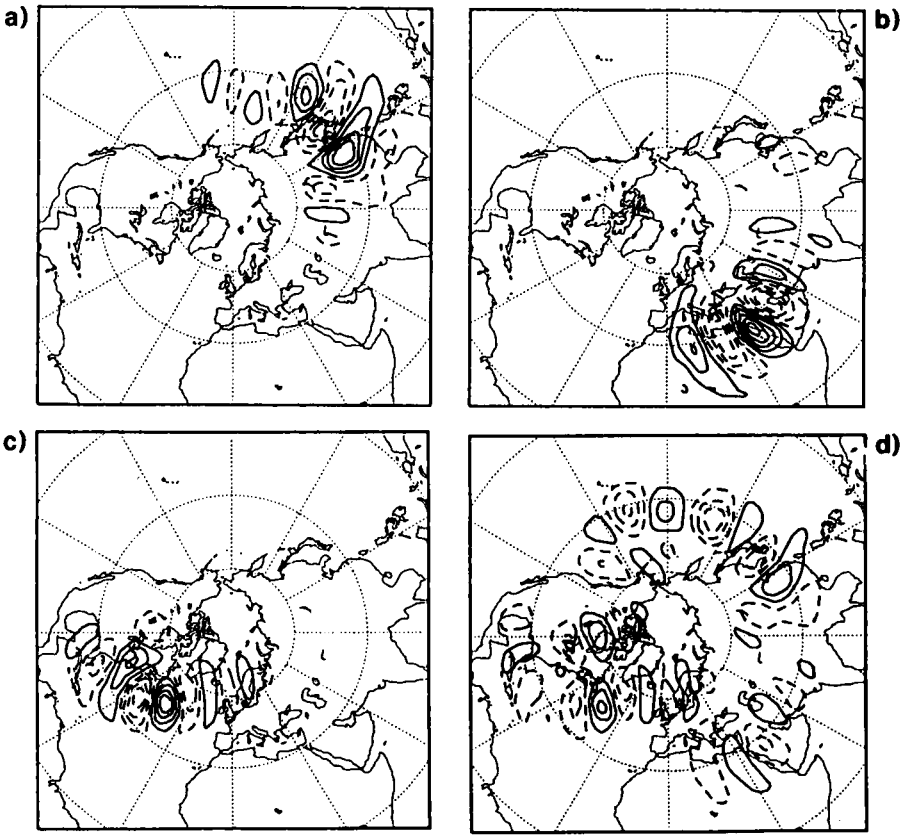

Figure 1. (a), (b) and (c). Examples of temperature patterns at model level 13 ($\approx$ 700 hPa) associated with individual singular vectors localized over the three dominant instability regions in winter; (d) example of a temperature perturbation (at level 13) generated from the rotation and scaling of 16 singular vectors for the same winter case. Contour interval 1 K, with negative values dashed.

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 optimization time. The SV structure 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.

(iv) Singular vectors tended to be localized 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 (examples from one wintertime case are given in Figs. 1(a), (b) and (c)). 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.

Statistics on the growth and distribution of the SVs in different seasons were discussed by BP, and for brevity are not repeated here.

(*b*) *Choice of inner product*

The choice of inner product plays an important role in determining the structure of the singular vectors. To illustrate this, SVs have been calculated for one specific initial date

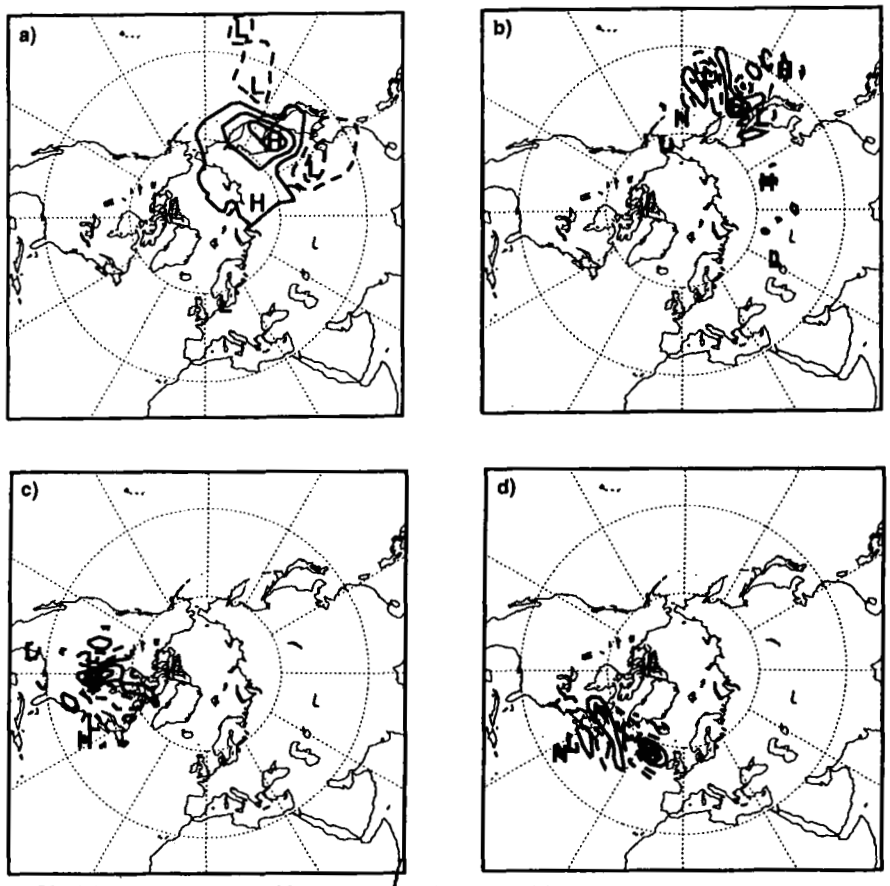

Figure 2. Vorticity structure at model level 11 (approximately 500 hPa) of the most unstable enstrophy-SV at (a) initial and (b) final time, computed for 6 April 1994. (c) and (d) As (a) and (b) but for the most unstable energy-SV. Contour intervals 0.08×10^{-6} s^{-1} at initial time and 1.6×10^{-6} s^{-1} at final time (the SVs have initial unitary total energy norm), with negative values dashed.

(which is believed to be fairly typical) using both energy and enstrophy inner products; for this case study, the tangent model has been integrated at T42 resolution to minimize truncation effects on the energy and enstrophy spectra at synoptic scales.

Figure 2 compares the vorticity structure of the first SV based on these two inner products at model level 11 (approximately 500 hPa), at initial and final time. Apart from the different location, one can see that the first enstrophy-SV has a much larger spatial scale at initial time (Fig. 2(a)) than at final time (Fig. 2(b)), compared with the energy-SV (Figs. 2(c) and (d)). This is easily understood; large-scale initial perturbations have relatively small enstrophy, but can influence the advection of synoptic and sub-synoptic waves in the basic-state trajectory, leading to perturbation growth dominated by small-scale features with very large enstrophy.

For each choice of inner product, the spectrum of the initial and final distribution of energy and enstrophy, averaged over the first 16 SVs, is plotted in Fig. 3 as a function of the total wave number. Consistent with Fig. 2, the enstrophy spectrum of the enstrophy-SVs (Fig. 3(a)) is red at initial time and blue at final time. Because of the large enstrophy amplification, values at initial time are multiplied by 40 in order to plot them on the same

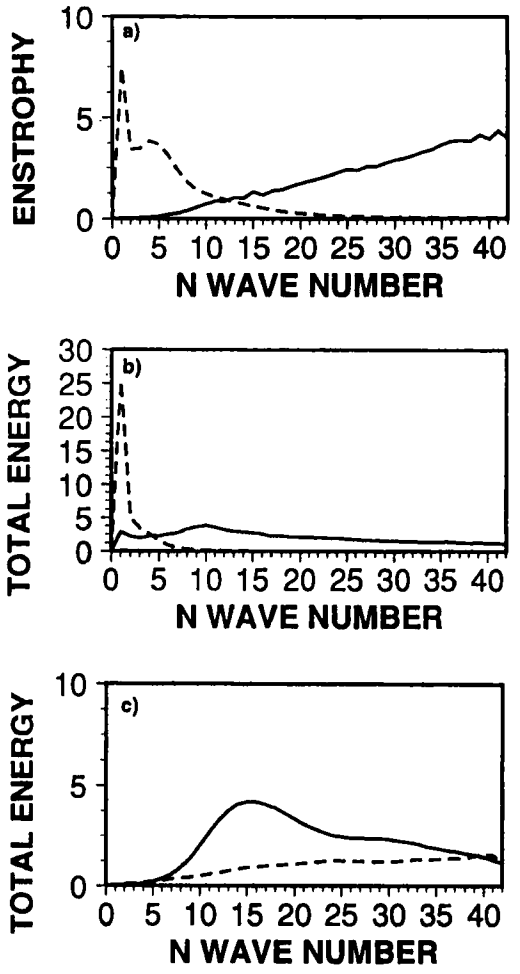

Figure 3. (a) Average enstrophy (s^{-2}) spectrum of the first 16 enstrophy-SVs at initial (dash, $\times 40\ 10^{16}$) and final (solid, $\times 10^{16}$) time, (b) average energy (per unit mass) spectrum ($m^2 s^{-2}$) of the first 16 enstrophy-SVs at initial (dash, $\times 40$) and final (solid, $\times 40$) time, and (c) average energy spectrum of the first 16 energy-SVs at initial (dashed, $\times 40$) and final (solid) time, for 6 April 1994.

scale as the final values. However, when the *energy* spectrum of the enstrophy-SVs is considered, there is still a cascade of energy from the largest to the smallest scales, but the total energy gain is only about a factor of 2 (and hence no re-scaling of initial values is needed in Fig. 3(b)). Conversely, the evolution of the energy-SVs (Fig. 3(c)) leads to a 'reddening' of the energy spectrum, with an increase of energy at all wave numbers, but particularly strong in the synoptic-scale range (as in Fig. 3(a), the initial energy values are multiplied by 40 in the plots).

As an estimation of the components of analysis error which evolve to dominate forecast error, the (initial) SVs of $\mathbf{L}_P$ approximate the eigenvectors of $\mathbf{C}\mathbf{L}_P^*\mathbf{L}_P$ where $\mathbf{C}$ is an (a priori) analysis-error covariance matrix (see, for example, Palmer *et al.* 1993). Clearly the accuracy of this approximation depends on how close $\mathbf{C}$ is to the identity matrix, i.e. the extent to which the analysis-error probability distribution is isotropic in phase space. This is inner product dependent; a distribution that is isotropic in one metric space may

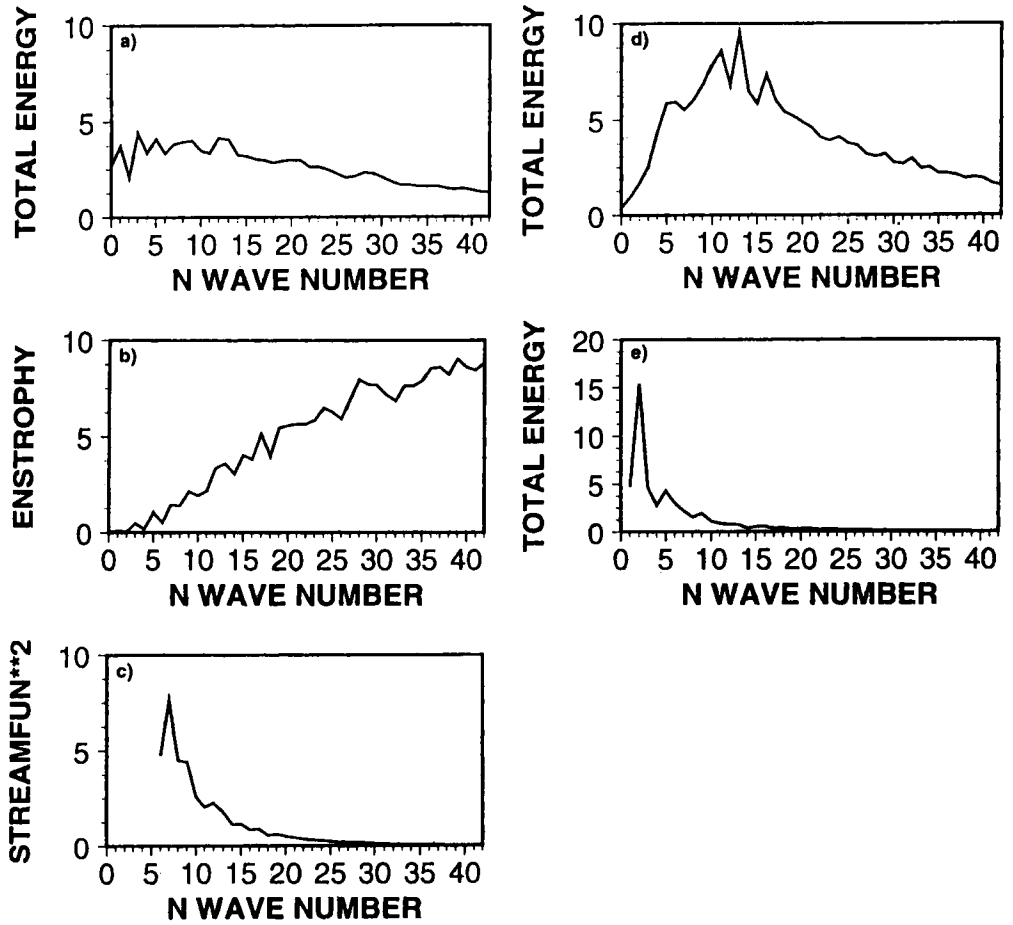

Figure 4. (a) Energy (per unit mass) spectrum (m^2s^{-2}), (b) enstrophy spectrum (s^{-2}) and (c) squared-stream-function spectrum (m^4s^{-2}) of the ECMWF–DWD analysis difference. (d) Energy spectrum (m^2s^{-2}) of the 48 h ECMWF forecast error and (e) energy spectrum (m^2s^{-2}) of the ECMWF analysis, for 3 January 1994. To have comparable amplitudes, the energy, enstrophy and squared stream-function of the ECMWF–DWD analysis difference have been re-scaled by 1, 10^{16} and 10^{-7}, respectively and the energy spectrum of the ECMWF 48 h forecast error and of the analysis by 1 and 10^{-2}, respectively. Where values are not plotted, they greatly exceed the maximum value on the axis.

not be isotropic in another. Let us focus attention on the component of the tangent-model phase space spanned by the (orthogonal) spherical harmonic functions. If the analysis-error probability distribution is isotropic in phase space with a given inner product, then the spectrum of the associated norm of the spherical harmonic projections will be white.

By construction, this spectrum will be white if **C** is used as a metric to define the inner product. In practice, we know little about **C** (especially in the sub-synoptic range). However, some information about the spectrum of possible analysis error can be obtained from the difference between two independent analysis fields. The calculations shown in Figs. 4(a)–(c) are based on the difference between the ECMWF and Deutsche Wetterdienst (DWD) analyses for a specific initial date on which spherical harmonic fields from both analyses were readily available at the ECMWF. Figures 4(a)–(c) show the spectrum of this analysis-difference field in terms of the enstrophy, energy and stream-function variance

(associated with three possible choices of inner product). We can clearly see that the spectrum is whitest in terms of energy, which only decreases by a factor of about 2 (hence norm by $\sqrt{2}$) across the range of scales plotted. By comparison the stream-function and enstrophy spectra are strongly biased red or blue respectively.

It is worth comparing the energy spectrum of this analysis-difference field with that of the corresponding analysis itself, and with the corresponding 2-day forecast error. The analysis-difference spectrum is much whiter than either of these two. (The analysis spectrum was sufficiently red that the energy of the lowest spherical harmonic component exceeded the upper plotting bound.) As would be hoped, the spectrum of the 2-day forecast error has more in common with the SV spectrum at final time that with the spectrum at initial time.

In conclusion, until a more complete description of analysis-error covariance emerges, the most appropriate of the 'simple' inner products to estimate SV structure for the purposes of ensemble prediction appears to be that based on the energy of the perturbation fields.

(*c*) *Relationship of SVs to sensitivity patterns*

The objective of SV analysis is to find perturbations (or phase-space directions) that could generate significant short-range-forecast errors 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 analysis-error structures with the largest amplitude evolve into the largest forecast errors.

One way of tackling this problem is to use the adjoint propagator $\mathbf{L}_P^*$ to map actual short-range-forecast errors back to initial time. Such studies have begun at the ECMWF and preliminary results are documented by Rabier *et al.* (1996). It is straightforward to show that if $\mathbf{E}$ is the day-2 forecast error, then $\mathbf{L}_P^*\mathbf{E}$ is the gradient of the energy norm of $\mathbf{E}$ at the initial time. In other words a perturbation to the analysis $\mathbf{S}_0 = k\mathbf{L}_P^*\mathbf{E}$, where k is an appropriate scaling factor, will optimize the projection of the perturbation at day 2 onto the observed error field. Rabier *et al.* refer to $\mathbf{S}_0$ as a sensitivity pattern.

Rabier *et al.* (1996) have studied the skill of a ten-day forecast where the analysis is perturbed with the $\mathbf{S}_0$ perturbation. They note that the skill can be substantially improved well beyond the 48-hour period used to define the error field and the adjoint propagator. 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.

In order to clarify the relationship between the sensitivity pattern $\mathbf{S}_0$ and the SVs, let us assume that the day-2 error $\mathbf{E}$ is entirely due to the (linear) evolution of an initial error pattern $\mathbf{E}_0$, so that

$$\mathbf{E} = \mathbf{L}_P\mathbf{E}_0. \tag{6}$$

(In practice, $\mathbf{E}_0$ cannot be computed by inverting the above relationship, because $\mathbf{L}_P^{-1}$ is ill-conditioned for those spatial scales dominated by dissipative processes.) Then the sensitivity pattern is given by

$$\mathbf{S}_0 = k\mathbf{L}_P^*\mathbf{E} = k\mathbf{L}_P^*\mathbf{L}_P\mathbf{E}_0. \tag{7}$$

If we expand $\mathbf{E}_0$ onto the orthonormal basis of the SVs v_i of $\mathbf{L}_P$ (which are eigenvectors

of $\mathbf{L}_P^* \mathbf{L}_P$ with eigenvalues σ_i^2), so that

$$\mathbf{E}_0 = \sum_i c_i \nu_i \tag{8}$$

then

$$\mathbf{S}_0 = k \sum_i \sigma_i^2 c_i \nu_i. \tag{9}$$

Therefore, one can see the sensitivity pattern as a filtered version of the (hypothetical) analysis error $\mathbf{E}_0$, where the filtering is performed in the SV space by multiplying each SV projection by a coefficient proportional to the squared singular value. With a rapidly decreasing spectrum of singular values, the sensitivity pattern will be dominated by its projection onto the fastest-growing SVs providing the analysis-error spectrum is reasonably white (c_i = constant, see section 2(b)). Indeed, the horizontal and vertical structures of the sensitivity patterns computed by Rabier *et al.* (1996) bear a strong similarity with the structure of the SVs described here (for example, the energy peak in the lower troposphere, and the strong westward tilt with height through the depth of the troposphere (R. Gelaro (1995), personal communication), despite the fact that T63 resolution was used for the adjoint model in the sensitivity analysis.

Rabier *et al.* (1996) find that, using the energy inner product, the energy associated with the sensitivity pattern lies well within any reasonable local bound of analysis error. However, this result is not independent of inner product. For example, using an enstrophy norm the sensitivity gradient will be dominated by larger scales (consistent with Eq. (9) and Figs. 1 and 2). However, the energy associated with these perturbations can be locally large (consistent with the weak energy growth of enstrophy-SVs, cf. Fig. 3) and can exceed any reasonable a priori estimate of local analysis error.

In summary, there is evidence that the spatial and temporal structure of energy-norm SVs describe a component of analysis error which dominates error evolution in the medium range. On the other hand, individual SVs are too localized to represent hemisphere-wide perturbations for the ensemble system. The means to overcome this problem are discussed in the next section.

(*d*) *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 by Palmer *et al.* (1993). Its aim is to create perturbations which cover most of the northern hemisphere, and have an amplitude comparable (in any region) with the estimates of root-mean-square (r.m.s.) analysis error provided by the optimum-interpolation (OI) data assimilation. The first step is the selection of 16 SVs among the first 30–35 computed by the Lanczos algorithm. This proceeds as follows:

1. The first four SVs are always selected.
2. For each SV, a localization 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, and to 0 elsewhere.

3. An overlap function is defined at each point as the sum of localization 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 (from the 5th onwards) 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 localization 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 local projection operator in March 1993, 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 northern hemisphere *and* in regions with small overlap function. During the boreal winter, most of the SVs computed by the Lanczos algorithm were in the northern hemisphere 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 northern hemisphere *a posteriori* would have been increasingly inefficient from a computational point of view (as shown by Buizza (1994b)); therefore, the localization had to be directly incorporated in the SV computation.

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

$$\mathbf{P}' = \mathbf{VR} \tag{10}$$

and then:

$$\mathbf{P} = \mathbf{P}'\mathbf{D}. \tag{11}$$

Let $p_i' = \{u_i', v_i', T_i'\}$ be one of the orthonormal perturbations defined by Eq. (10), in terms of zonal and meridional wind and temperature respectively. Moreover, let e_u, e_v, e_T be the OI estimates of r.m.s. analysis error for these variables. The continuous function

$$f_i = [\overline{(u_i'/e_u)^8 + (v_i'/e_v)^8 + (T_i'/e_T)^8}]^{1/8} \tag{12}$$

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 as to minimize the cost function

$$CF = \sum_{i=1}^{N} f_i^2. \tag{13}$$

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 minimization iteratively by constructing **R** as the product of a series of 2 × 2 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 nonlinear nature of the cost function may generate more than one minimum in phase space.

Once the rotation has been performed, the perturbations are re-scaled in order to have a realistic local amplitude. The non-null elements of the diagonal matrix **D** in Eq. (11) are given by:

$$d_{ii} = \alpha / f_i \tag{14}$$

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(d). This can be compared with the localized SV structures shown in panels (a) to (c).

(*e*) *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 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.

Figure 5 shows the r.m.s. amplitude of the zonal wind component at model level 11 (approximately 500 hPa) and of temperature at model level 13 (approximately 700 hPa) for the ensemble perturbations in winter (panels (a) and (d)), summer ((b) and (e)) and autumn ((c) and (f)). (The equivalent fields for spring are shown in Figs. 6(b) and (f).) For winter, the r.m.s. amplitude is maximized at approximately the regions where the dominant SVs occur (see Fig. 1), i.e. over the western Pacific and Atlantic basins, and over subtropical north Africa.

The impact of the application of the local projection operator can be inferred by comparing Figs. 5(a) and (d) with r.m.s. amplitude distributions for other seasons. As discussed above, this operator is designed to find SVs in which energy growth is maximized in the extratropical northern hemisphere. Consequently, the local projection severely damps the r.m.s. amplitude of the subtropical north African perturbations, while the perturbation amplitude in the storm-track regions is almost unaffected. In common with the distribution of dominant SVs through the annual cycle (Fig. 5 of BP), the r.m.s. amplitude of perturbations is more zonally asymmetric in winter compared with other seasons, and more uniform in summer.

In Fig. 6 we show, for the spring season, the zonal wind and temperature 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. The wind amplitudes shown here (less than 1 m s^{-1}) are everywhere smaller than the OI error estimates: in practice, the OI estimate of temperature error puts the strongest constraint on the perturbation amplitude. The r.m.s. temperature component of the perturbations is largest around the 700 hPa level in most areas, with maxima between 1 and 1.5 K. These values are indeed comparable with 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.

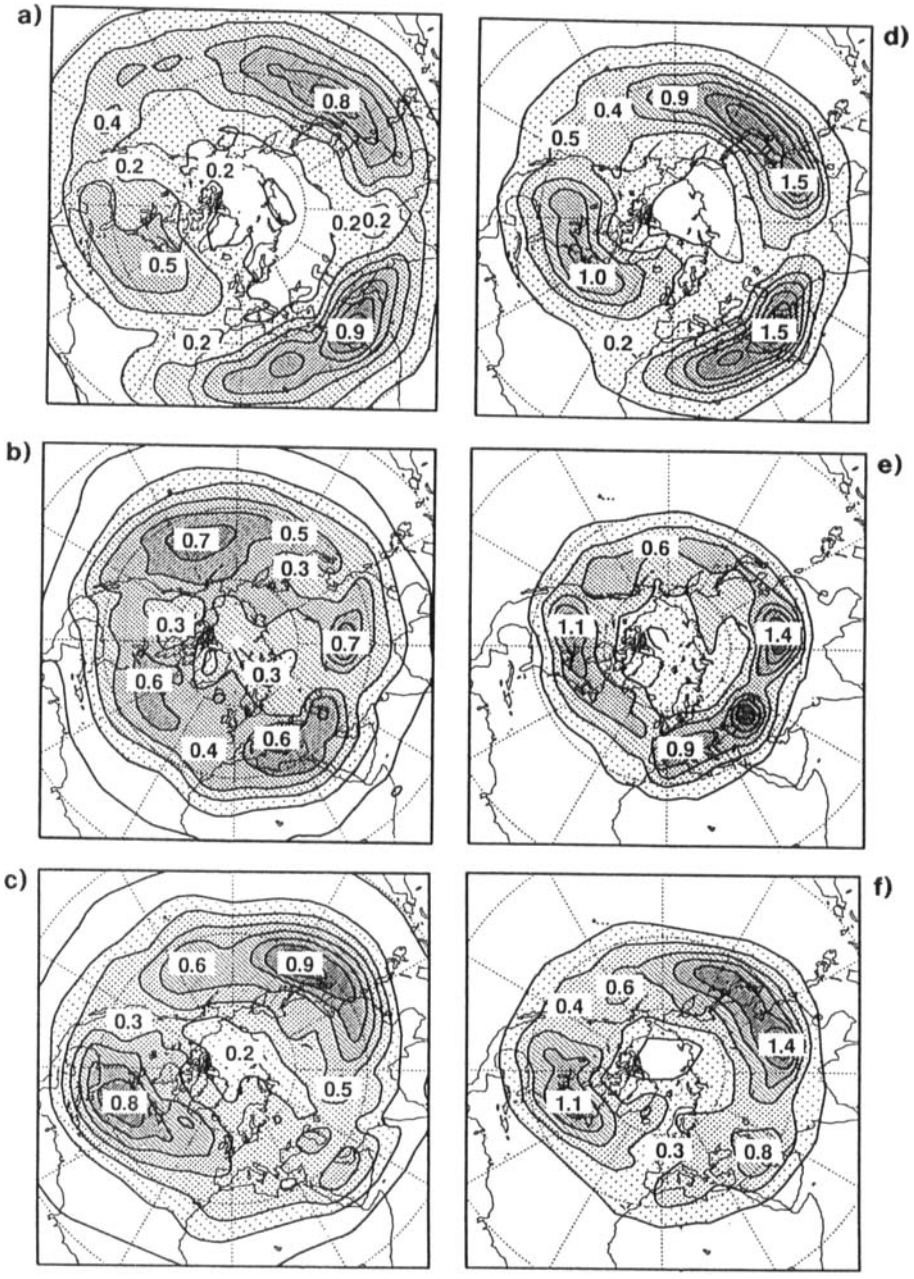

Figure 5. Left column: r.m.s. amplitude of the u-wind component of the initial perturbations at model level 11 ($\approx$ 500 hPa) in (a) winter 1992/93, (b) summer 1993 and (c) autumn 1993 (contour interval 0.1 m s^{-1}). Right column: r.m.s. amplitude of the temperature component of the initial perturbations at model level 13 ($\approx$ 700 hPa) in (d) winter 1992/93, (e) summer 1993 and (f) autumn 1993 (contour interval 0.2 K).

(*f*) *Ensemble products*

In this subsection we briefly describe ensemble products which are routinely disseminated to the 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

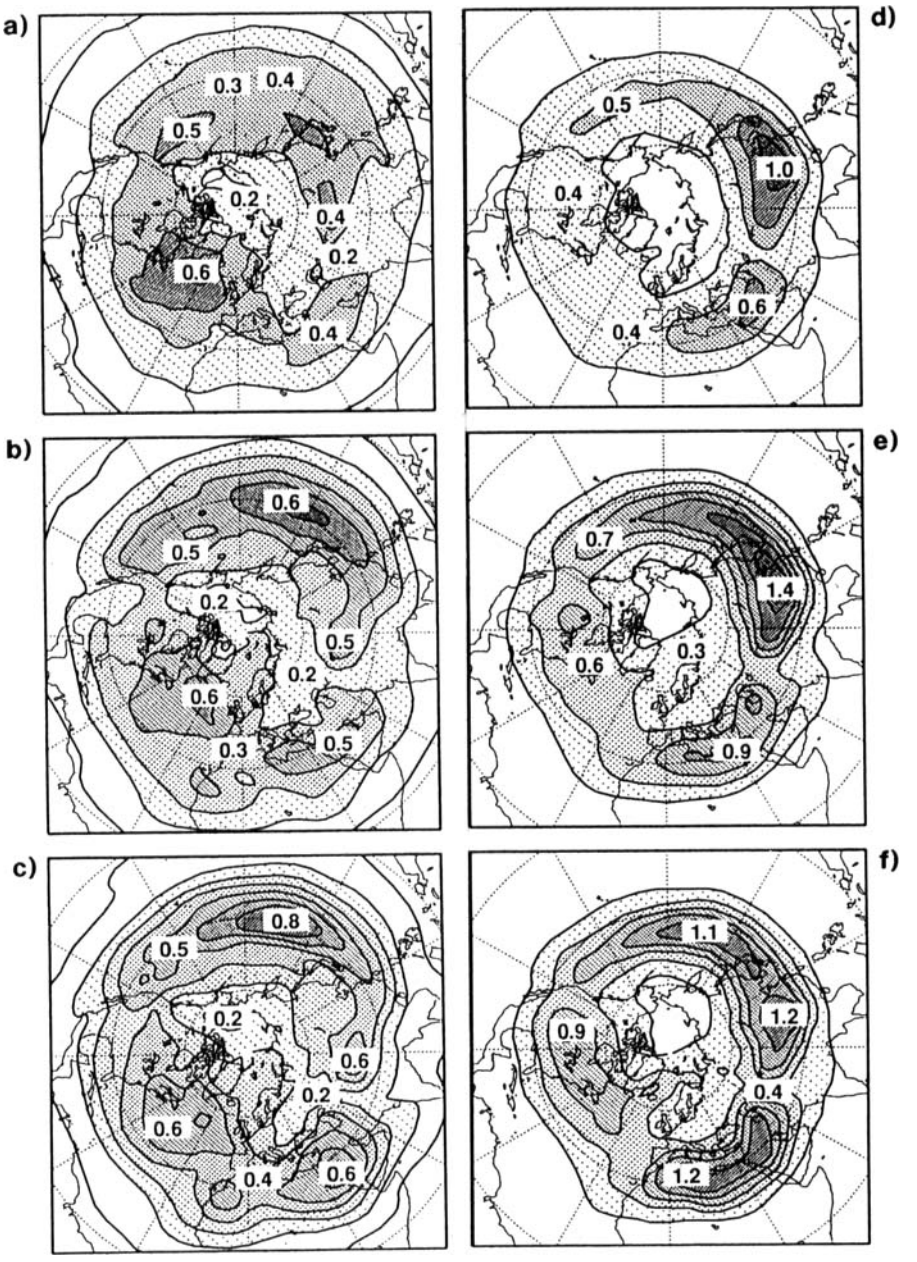

Figure 6. Left column: r.m.s. amplitude of the u-wind component of the initial perturbations for spring 1993 at (a) model level 9 ($\approx$ 300 hPa), (b) model level 11 ($\approx$ 500 hPa) and (c) model level 13 ($\approx$ 700 hPa) (contour interval 0.1 m s^{-1}). Right column: r.m.s. amplitude of the temperature component of the initial perturbations for spring 1993 at (d) model level 9, (e) model level 11 and (f) model level 13 (contour interval 0.2 K).

the human eye to assimilate qualitatively information from each individual forecast. The set of 500 hPa height forecast maps over Europe can be plotted on a single sheet of paper of A4 size or similar (e.g. Fig. 17 in section 4(a)). Although the size of an individual map is clearly minimal, it conveys the principal feature 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. It is expected that 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 used by Brankovic *et al.* (1990) and Palmer *et al.* (1993), we have used Ward's hierarchical clustering algorithm (e.g. Anderberg 1973), 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 analysed). For operational implementation it was felt that information for more than one forecast time should be provided, while 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 r.m.s. 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 set at 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 varies 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. An example of clusters obtained in a case of blocking onset (the ensemble started on 25 January 1993) is given in Fig. 7. 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

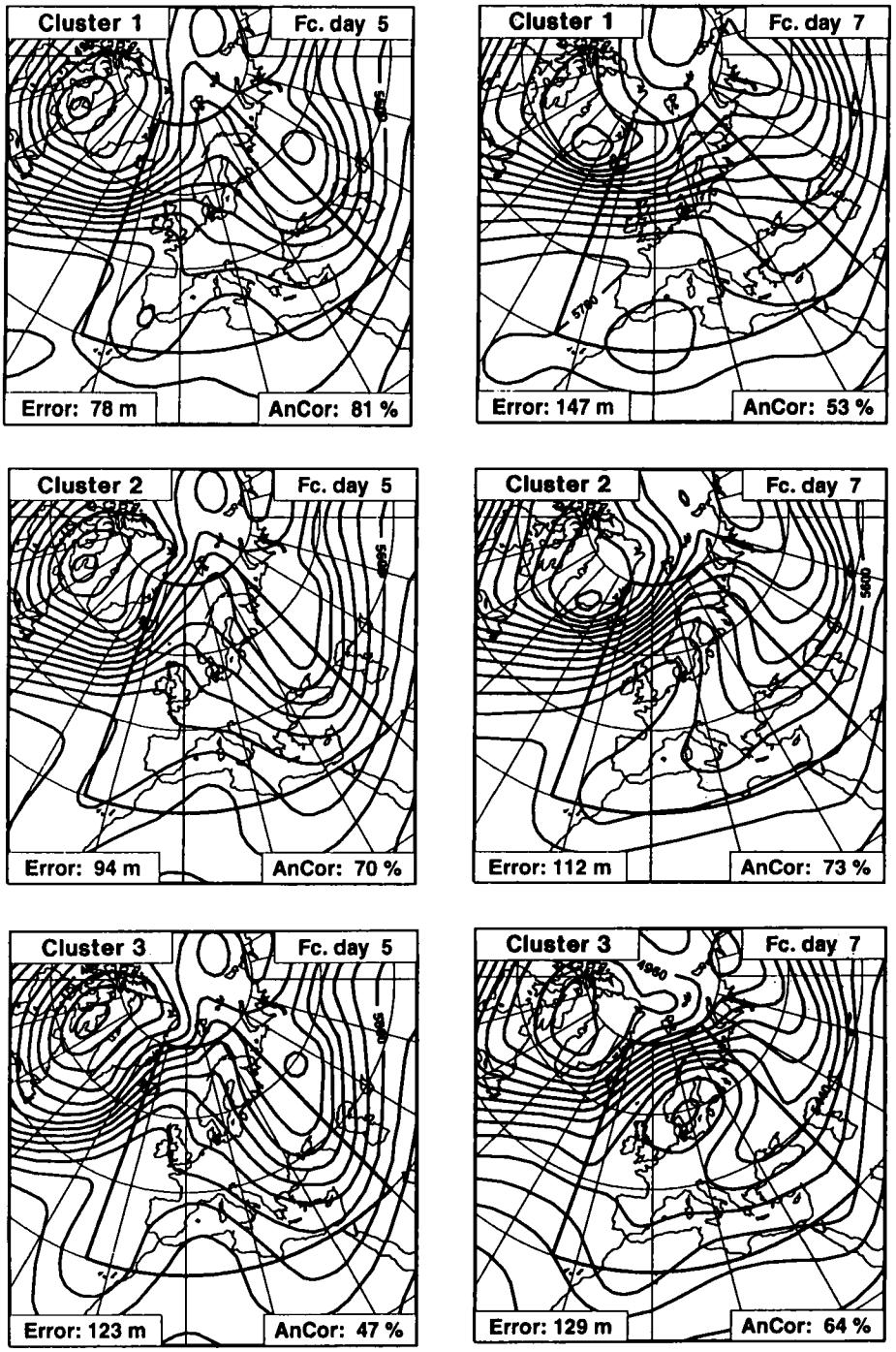

Figure 7. Clusters of day 5-to-7 500 hPa height over Europe from the ensemble forecast started on 25 January 1993. Top: forecast-day 5 and 7, cluster 1 (15 members); centre: forecast-day 5 and 7, cluster 2 (9 members); bottom: forecast-day 5 and 7, cluster 3 (9 members).

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. It should be noted that, strictly speaking, the control should be given a higher weight then the ensemble members. The weight will decrease with forecast lead time. However, at the time of writing this weighting has not been applied operationally.

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 inconsistences 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 later in Fig. 24.

(iv) *Probability maps.* Two-dimensional fields representing probabilities of rainfall, 10 m wind speed and 850 hPa temperature anomalies for specific forecast days are also post-processed 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^{-1}, > 5 mm day^{-1}, > 10 mm day^{-1} and > 20 mm day^{-1}. The wind speed categories are: > 10 m s^{-1} and > 20 m s^{-1}. 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 section 4 (see Figs. 19 and 23).

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 the skill of any individual forecast (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, even for large spread, the control forecast could be skilful if, by chance, it happened to be close to the verification trajectory. To account for sampling problems of this kind, we will compare actual results with those from a hypothetical perfect-model ensemble when discussing relationships between ensemble spread and control-forecast skill.

Figures 8(a) and (b) show scatter diagrams of skill and spread for the northern hemisphere in winter and summer respectively. (Diagrams for spring and autumn are not shown since the seasonal trend in both spread and skill indices may introduce a trivial correlation between them.) Here the ensemble spread is taken as the 75th percentile of the distribution of the r.m.s. 500 hPa height difference between the perturbed ensemble members and the control. The day-7 r.m.s. error of the control is taken as the skill value. 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 2×2 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. (Spread/skill relationships have also been studied using anomaly correlation as the measure of distance between either

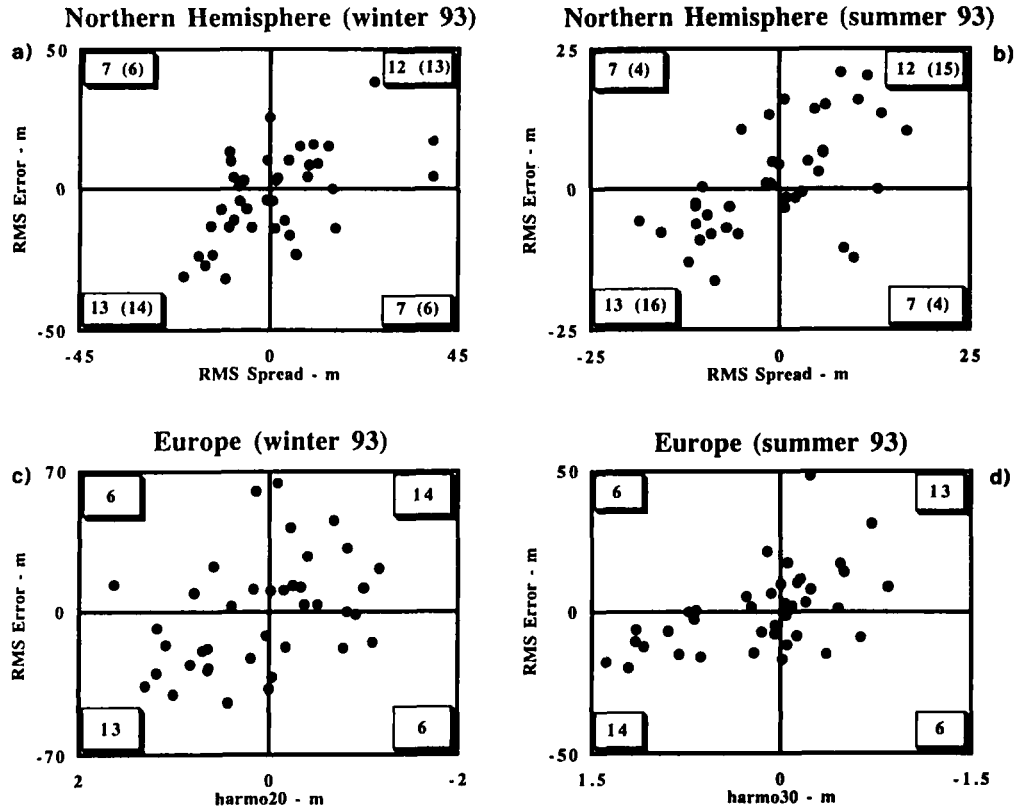

Figure 8. (a) Scatter diagram between day-7 northern hemisphere r.m.s. error of control forecast (ordinate) versus day-7 northern hemisphere r.m.s. spread between ensemble members and control forecast (abscissa) for winter 1992/93. (b) As in (a), but for summer 1993. (c) Scatter diagram between day 5-to-7 European r.m.s. error of control forecast (ordinate) versus the consistency index defined as the forecast day on which a chosen probability contour of the 500 hPa height plumes vanishes (abscissa) (see text for further details) for winter 1992/93. (d) As in (c) but for summer 1993. Data are plotted as deviations from the median value.

control or verifying analysis, but are not discussed here for brevity. Results do not depend strongly on the measure used.)

For both seasons the diagonal entries are notably more populated than the off-diagonal entries. Whilst the off-diagonal entries are not negligible, it was 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 realizations. The contingency table for this perfect-model verification is given in parentheses in Figs. 8(a) and (b).

The comparison of the actual contingency tables with the perfect-model simulations gives a better indication of the strength of the spread–skill relationship than any 'absolute' index. Because of the use of medians to define the class boundaries, the difference in frequency must be equal for all the four quadrants of the table. The difference in winter (1 unit) has only a 25% chance to be significant according to a chi-square test, while the difference in summer (3 units) is statistically significant.

The linear correlation coefficients for the scatter diagrams in winter and summer are 0.56 and 0.59 respectively. These values compare well with perfect-model estimates of a

medium-range asymptotic limit of 0.6 reported by Barker (1991). The spread/skill relations appear superior to those obtained by other means for this time range (e.g. the day-7 northern hemisphere values between 0.3 and 0.4 reported by Molteni and Palmer (1991)), although comparable results obtained with regression estimates of anomaly correlation at earlier forecast times have been reported by Wobus and Kalnay (1994, personal communication).

It is interesting to note directly from the scatter diagrams that in general the dispersion of points along the y-axis (that is the variability in the error of the control forecast) is smaller for ensembles with small spread than for ensembles with large spread. This was anticipated by the remarks at the beginning of this section. We have studied sub-regions of the northern hemisphere using these measures of spread and skill. Results are not shown here for brevity, but we found comparable contingency relations to those in Figs. 8(a) and (b).

A second method of analysing the relation between forecast skill and ensemble dispersion is illustrated in Figs. 8(c) and (d), which show scatter diagrams for the European region (30°N–75°N, 20°W–45°E) based on forecast probability plumes for 500 hPa geopotential height at grid points distributed uniformly throughout the region. Again, results for winter and summer are shown. The spread is estimated from the forecast time a chosen probability contour first vanishes (for example, from the plume shown in Fig. 24(b), the 30% contour first vanishes at about day 9). The mean forecast time averaged over all the grid points is then taken as a measure of ensemble consistency (the smaller the spread the longer the consistency time index). This consistency index has been investigated as a predictor of the medium-range skill of the control forecast, measured by the r.m.s. 500 hPa height error averaged over days 5 to 7.

As may be expected, the correlation between these indices is highest when the consistency index is representative of the spread in the day 5-to-7 range. To obtain this, the probability contour chosen to estimate ensemble dispersion must vary with season; for winter we have used the 20% contour, for summer the 30% contour. Ensemble spread is generally weaker in summer, so that the 30% contour generally extends into the medium range and is, therefore, a reasonable predictor of medium-range skill. By contrast, in winter the 30% contour may already vanish in the short range and, therefore, would not be a good indicator of medium-range skill.

As in the previous diagrams, 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; however, in this case a perfect-model contingency table could not be easily estimated because of the rather complex nature of the predictor. Similarly to the northern hemisphere spread/skill statistics, the contingency tables are dominated by diagonal entries. It can be seen that a clearer relation between plume dispersion and forecast skill occurs in summer than in winter.

(*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 section 2(b), 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 from day to day makes an objective verification of the cluster probabilities impossible with standard skill tests.

For this purpose, as by 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 frequences) 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) and Ferranti *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 eight clusters computed from the analysis sample, according to their similarity with observed fields belonging to the clusters.

Figure 9 shows the eight 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 each of the three cluster sets (12, 8 or 4 clusters) 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 eight 'climatological' clusters, from forecast time $t = 0$ to $t = 10$ days at 12-hour intervals. We can, therefore, define 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

$$P_a(j_a, t) = 1 \quad P_a(j \neq j_a, t) = 0. \tag{15}$$

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 computed as

$$B_e(t) = \overline{\sum_{j=1}^{M} [P_e(j, t) - P_a(j, t)]^2} \tag{16}$$

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 Eq. (16).

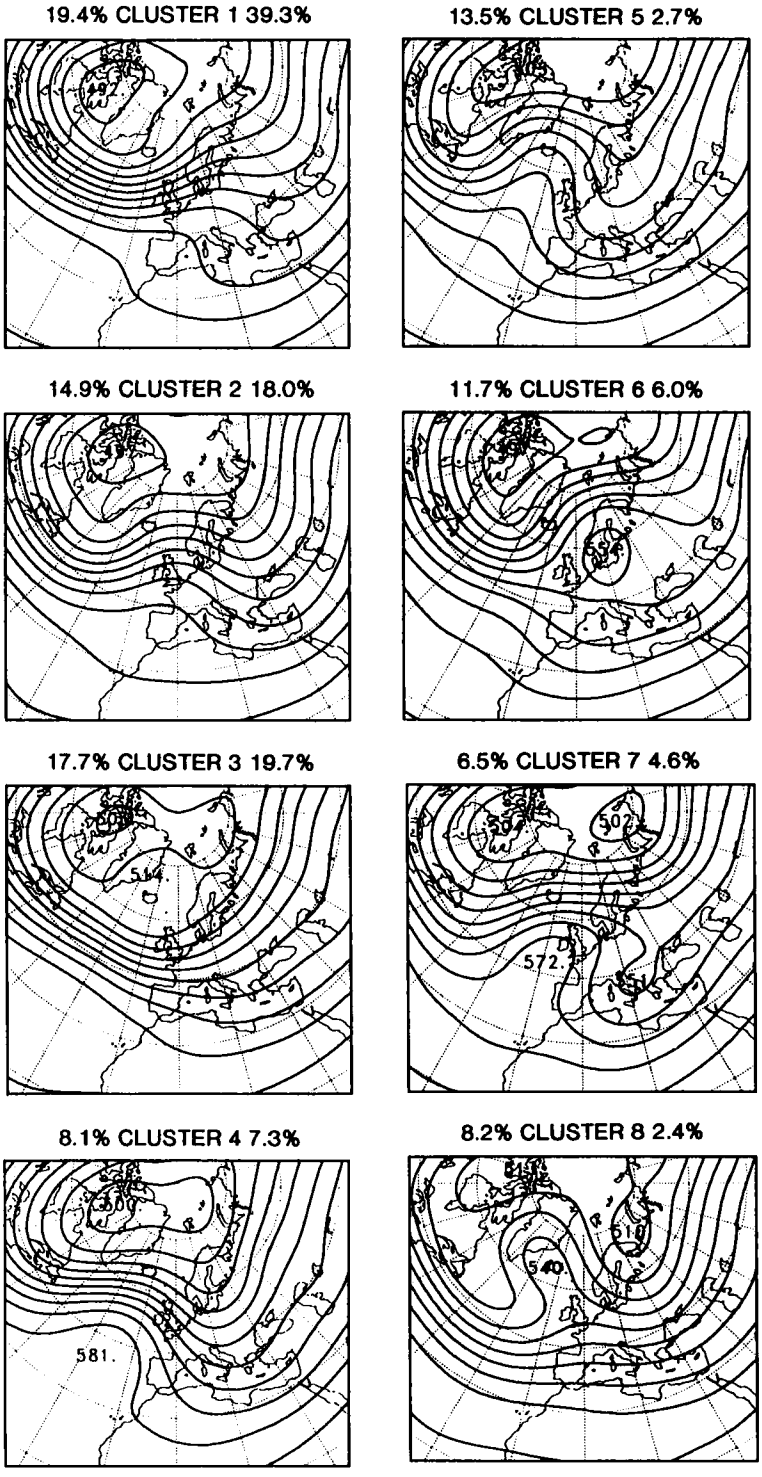

Figure 9. 500 hPa height associated with eight Euro–Atlantic cluster centroids derived from the ECMWF analysis archives. The observed frequency is shown for each cluster at top left. Simulated frequencies based on 120-day 'climate' integrations of the Ensemble Prediction System model are shown at top right.

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

$$B_{cl} = 1 - \sum_{j=1}^{M} P_{cl}^2(j) \tag{17}$$

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

Using the definition of P_a, one can rewrite Eq. (16) as:

$$B_e(t) = 1 + \overline{\sum_{j=1}^{M} P_e^2(j, t)} - 2\overline{P_e(j_a, t)} \tag{18}$$

where $\overline{P_e(j_a, t)}$ is the average probability of the verifying cluster. The second term on the right-hand side of Eq. (18) 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

$$B_c(t) = 2[1 - \overline{P_c(j_a, t)}]. \tag{19}$$

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

$$B_{ep}(t) = 1 - \overline{\sum_{j=1}^{M} P_e^2(j, t)}. \tag{20}$$

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 $2B_{cl}$ (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. 10(a) for the eight-cluster verifications. Firstly, we notice that the ensemble score B_e crosses the climatological score at forecast-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 Eq. (19)) that the average probability of a correct cluster prediction by the control forecast does not decrease monotonically with forecast time. Such behaviour can be understood by considering 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:

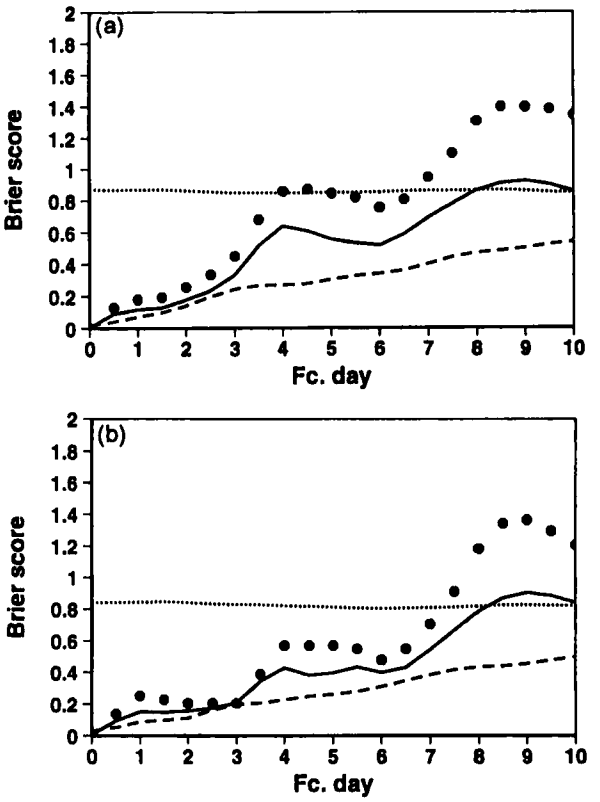

Figure 10. (a) Brier score of winter ensembles (solid line), control forecast (large dotted line), climatology (small dotted line), a perfect-model ensemble (one member chosen randomly as verification; dashed line). (b) As (a) but excluding those ensembles in which the verifying analysis was classified in clusters where the model climatology significantly underestimated the observed frequency.

(i) it makes the correct transition at the correct time;
(ii) it makes the correct transition but at a different forecast time; and
(iii) it persists in the initial cluster throughout the forecast range, or 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 evident in scores like r.m.s. error (this behaviour may occur, for example, when a transition over Europe is triggered by a perturbation which propagates across the Atlantic at a different speed in the forecast and in the analysis). In our set of forecasts, 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 the third type of behaviour, which implies a monotonic growth in the Brier score.

If the residence time in the clusters has an exponential distribution, as suggested by a number of studies on atmospheric regimes, a plateau in the Brier score of the control forecast should be expected in the time range in which errors in transition times predominate over other types of forecast behaviour. In a relatively small sample of forecasts, this plateau may be turned into a non-monotonic curve because of differences between the number of transitions occurring at a particular forecast time and the number expected from the theoretical distribution.

This argument, however, cannot be applied to the ensemble; although the transition times of individual members may differ from the analysis, if the PDF of the initial error is properly sampled the ensemble probability of transition should peak at the correct time (unless systematic deficiencies are present in the model variability). Indeed, the Brier score B_{ep} of the 'perfect ensemble' is always strictly monotonic in Fig. 10(a).

Let us now return to the comparison between B_e, B_{ep} and B_c as illustrated in Fig. 10(a). Ideally, the ensemble score should be as close as possible to B_{ep} and substantially lower than the control score 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 (or evolves into) one of those clusters whose frequency is severely underestimated in the long-term climatology of the model. We have, therefore, re-computed the Brier scores excluding those ensembles in which the verifying analysis was classified in clusters 5 to 8 of the eight-cluster partition (Fig. 9) for more than half the time beyond forecast-day 3. The average Brier scores for the remaining 22 ensembles are shown in Fig. 10(b).

While the B_{ep} score is only marginally changed, consistently with a perfect-model assumption, the scores 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 Brier score indicates that the ensemble probability-distribution estimate is skilful in the medium range at least up to about forecast-day 8. The ensemble performance (as measured by this score) is significantly affected by model error; however, there are indications that deficiencies in the initial perturbations (e.g. in their resolution, Hartmann *et al.* 1995) are also contributing to insufficient spread in the earlier part of the forecast.

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

We will now turn to a more conventional kind of validation by showing time series of the ensemble distributions of conventional skill scores. Anomaly correlation (AC) and r.m.s. error of 500 hPa height have been computed over various areas for all the individual forecasts in each ensemble as well as for the T63L19 control and the T213L31 operational forecast. Here, scores are illustrated for the whole northern hemisphere (north of 20°N) and for Europe (as defined in section 2(d)(ii)). Figures 11 and 12 show for each ensemble the two extreme scores, the 25% and 75% percentiles and the median. In addition, the score of the T63 control and the score of the operational forecast are indicated.

The diagram for day-7 AC over the northern hemisphere in winter is shown in Fig. 11(a). 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 the two unperturbed forecasts was lower than the AC of the best member of the ensemble (which was usually around 0.8), and in most cases 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 r.m.s. error (not shown) support the results above. However, r.m.s. errors tend to give a more favourable view of the performance of the EPS

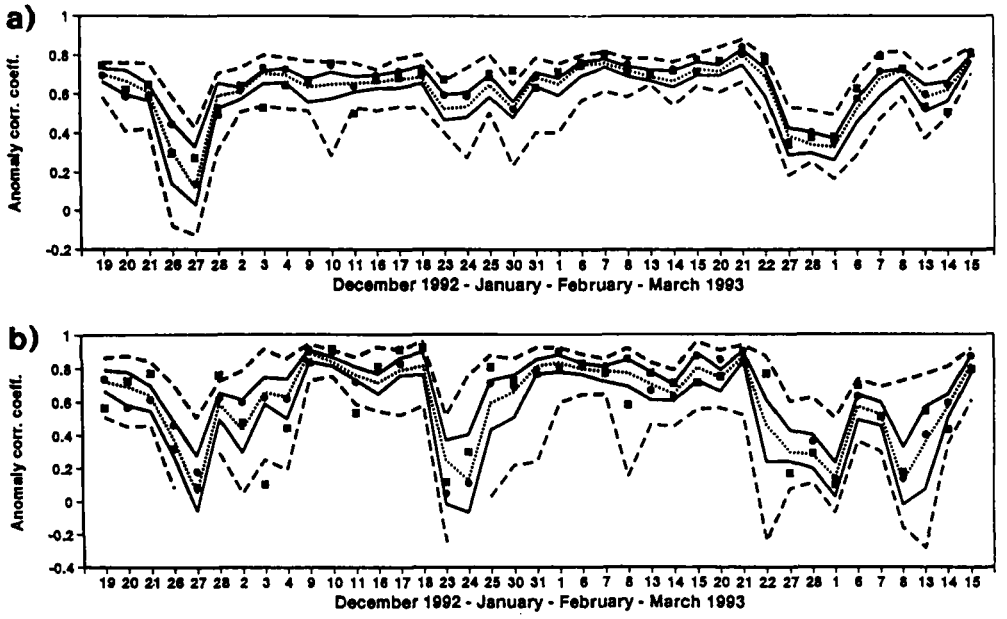

Figure 11. Time series of day-7 anomaly correlation (AC) scores in winter 1992/93 for (a) northern hemisphere and (b) Europe. Dashed lines: best and worst members of the ensemble. Solid lines: 25th and 75th percentiles of AC distribution. Dotted line: median of the distribution. Solid circle: T63L19 control forecast. Solid square: T213L31 operational forecast.

with respect to the operational forecast; this reflects the tendency of the T213L31 model to produce more intense features which, if out of phase with the verifying analysis, have a stronger negative impact on the r.m.s. error than on the AC.

Figure 11(b) shows the wintertime AC statistics for Europe, again at forecast-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 forecast-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 likelihood that model error may have contributed to the failure of some of the ensembles.

The three panels of Fig. 12 show the day-7 AC distribution over Europe in the other seasons. EPS scores were least satisfactory in spring; one can see in Fig. 12(a) 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

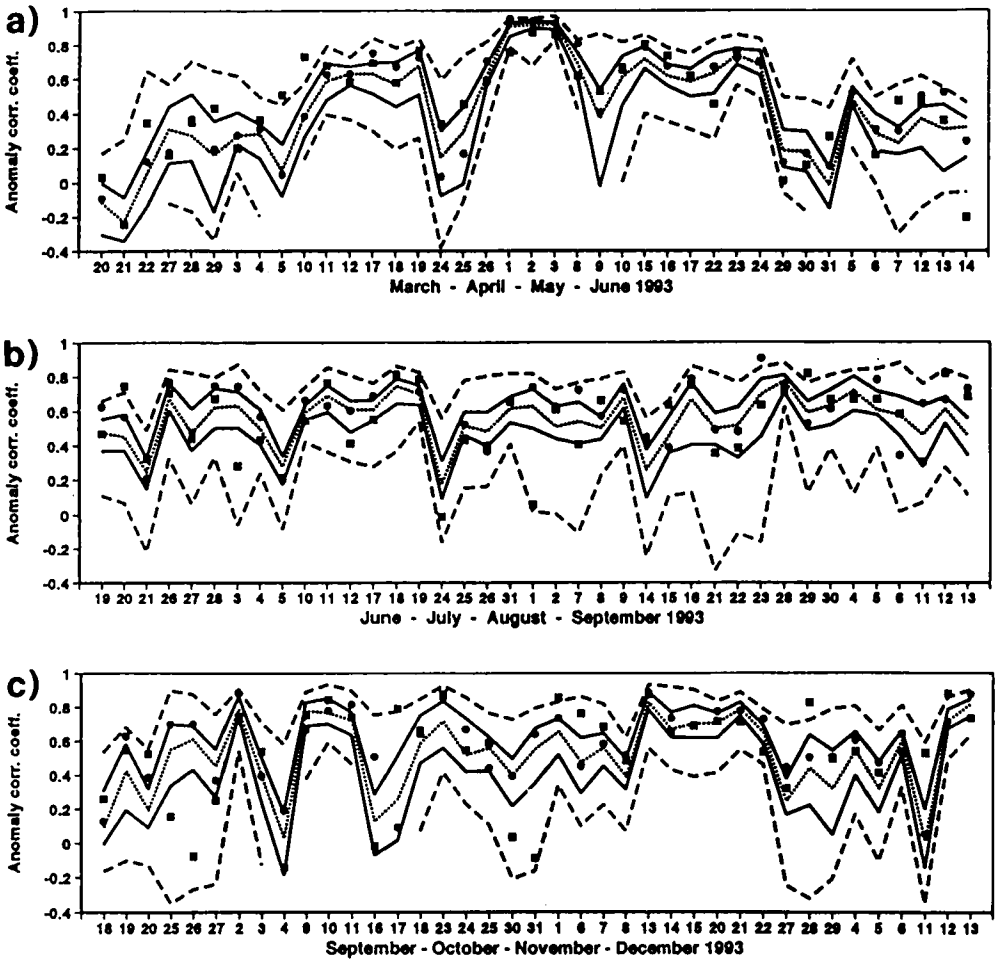

Figure 12. As Fig. 11 but for Europe only; (a) spring, (b) summer, and (c) autumn 1993.

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.

We can estimate the frequency of major ensemble failures by computing the percentage of cases where the AC of the most skilful ensemble member was less than some prescribed threshold AC_T. Results are given in Table 2 for day 7 over Europe, based on $AC_T = 0.6$. This threshold isolates cases where the verifying analysis trajectory lay significantly 'outside' the ensemble distribution. For comparison, the percentage of corresponding operational forecasts with $AC < 0.6$ is shown. As anticipated above, the most numerous ensemble failures occurred in spring when, in more than a quarter of the ensembles, the best member at day 7 had an $AC < 0.6$. However, this period was also 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 score distribution can also be used to assess whether the ensembles provided additional information with respect to the unperturbed control forecasts. One way to quantify this is to look at the percentages of perturbed forecasts which had better scores than the control T63. These percentages, based on AC statistics, have been computed for all seasons

TABLE 2. PERCENTAGE OF ENSEMBLE FORECASTS WHERE THE ANOMALY CORRELATION OF THE MOST SKILFUL MEMBER FOR DAY 7 OVER EUROPE WAS LESS THAN 0.6

	Ensemble	Operational forecast
Winter	8	38
Spring	28	56
Summer	5	49
Autumn	5	44

For comparison, the percentage of operational forecasts whose skill was less than 0.6 is also shown.

TABLE 3. AVERAGE PERCENTAGE OF PERTURBED ENSEMBLE MEMBERS WITH HIGHER ANOMALY CORRELATION THAN THE CONTROL FORECAST

	Northern hemisphere			Europe		
	Forecast day			Forecast day		
	5	7	10	5	7	10
Winter	22	32	37	36	40	42
Spring	21	30	45	30	39	51
Summer	17	29	42	33	37	43
Autumn	19	33	46	30	35	46

and several verification areas at forecast-day 5, 7 and 10; the values for the northern hemisphere (NH) and Europe are listed in Table 3. For the NH, about 20% perturbed forecasts are more skilful than the control at day 5, about 30% at day 7 and 40–45% at day 10. For the smaller European area, the variability in scores between individual ensemble members is more pronounced, especially in the earlier part of the forecast; so, the percentages rise to 30–35% at day 5, 35–40% at day 7 and 40–50% at day 10. (The European values are fairly representative of the percentages obtained for other 'limited', i.e. non-hemispheric, areas). Over both the NH and Europe, the winter percentages are the highest at day 5, but from day 5 to day 10 they grow less than in the other seasons.

A similar comparison performed between ensemble members and the operational T213L31 forecast reveals wider variations in the percentages from area to area and from season to season. In comparison with the figures quoted above, we generally find a smaller number of perturbed ensemble members which are more skilful than the operational forecast at day 5, but the advantage of T213L31 over the lower-resolution T63L19 control is usually lost by day 7, in agreement with experience at other NWP centres (e.g. Tracton and Kalnay 1993).

In judging whether these percentages can be considered satisfactory, one should compare them with estimates from perfect-model experiments. Unfortunately, for these particular statistics such estimates are not readily available. However, some lower and upper bounds can be computed on the basis of simple statistical assumptions. Firstly, we know that the control becomes equivalent to any other ensemble member when predictability due to initial conditions is lost; therefore one should expect to find half the perturbed forecasts better than the control when the forecast time approaches the limit of deterministic predictability (about 15 days according to many theoretical and numerical estimates, e.g. Lorenz 1969).

On the other hand, let us consider the situation at the beginning of the forecast. If the initial perturbations have an amplitude close to the expected norm of the analysis error, the probability that a perturbed initial condition is closer to the true atmospheric state than the control initial condition is very low in a multi-dimensional space. As demonstrated in the appendix, for a perturbed forecast to have a smaller initial error than the control, the projection of the (unperturbed) analysis error onto a given perturbation must be negative and (in absolute value) greater than half the perturbation amplitude. It follows, that if the analysis-error norm and the perturbation norm were exactly equal, then no more than three orthogonal perturbations could satisfy this condition. Assuming that the analysis error is unbiased and has an isotropic, multi-normal distribution in the sub-space of our 16 orthogonal perturbations, one can estimate that on average just one perturbation (with either positive or negative sign) will lead to a smaller initial error. The fact that, already at day 5, the percentages in Table 3 are much higher than their expected values at the initial time (about 3%) is an indication of the ability of our perturbations to capture the directions associated with the fastest error growth in phase space, even in the nonlinear phase.

(*d*) *Skill of the ensemble mean*

We conclude this section on objective validation by comparing the average skill of the ensemble mean with the skill of the control forecast as a function of forecast time.

In the 1970s and 1980s, improving the skill of deterministic forecasts by means of ensemble averaging was considered the primary objective of ensemble forecasting (e.g. Leith 1974; Hoffmann and Kalnay 1983). This is certainly an important goal in extended-range predictions, when one looks beyond the limit of deterministic predictability. However, in the medium range, information on the probabilistic distribution of the atmospheric state provided by just its first-order moment is of limited operational usefulness. However, the verification of the skill of the ensemble mean (compared with the mean skill and spread of ensemble members) may still be useful as a way of testing the properties of the ensemble distribution against reference theoretical models.

For the four seasons of: winter 1992/93, spring, summer and autumn 1993,

Fig. 13 compares the average values (as a function of forecast time) of the r.m.s. error of the control forecast and of the r.m.s. error of the ensemble mean. R.m.s. values are used here for an easier comparison with perfect-model estimates. An improvement in the skill of the ensemble mean with respect to the control forecast becomes evident after about day 4 in spring and autumn, and after about day 5 in summer.

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 describe 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 also illustrate the range of products disseminated to the National Meteorological Services of the ECMWF Member States.

(*a*) *30 October–6 November 1993*

In discussing results from the case study for 30 October–6 November 1993 we take into account the fact that three consecutive ensemble predictions are made each week.

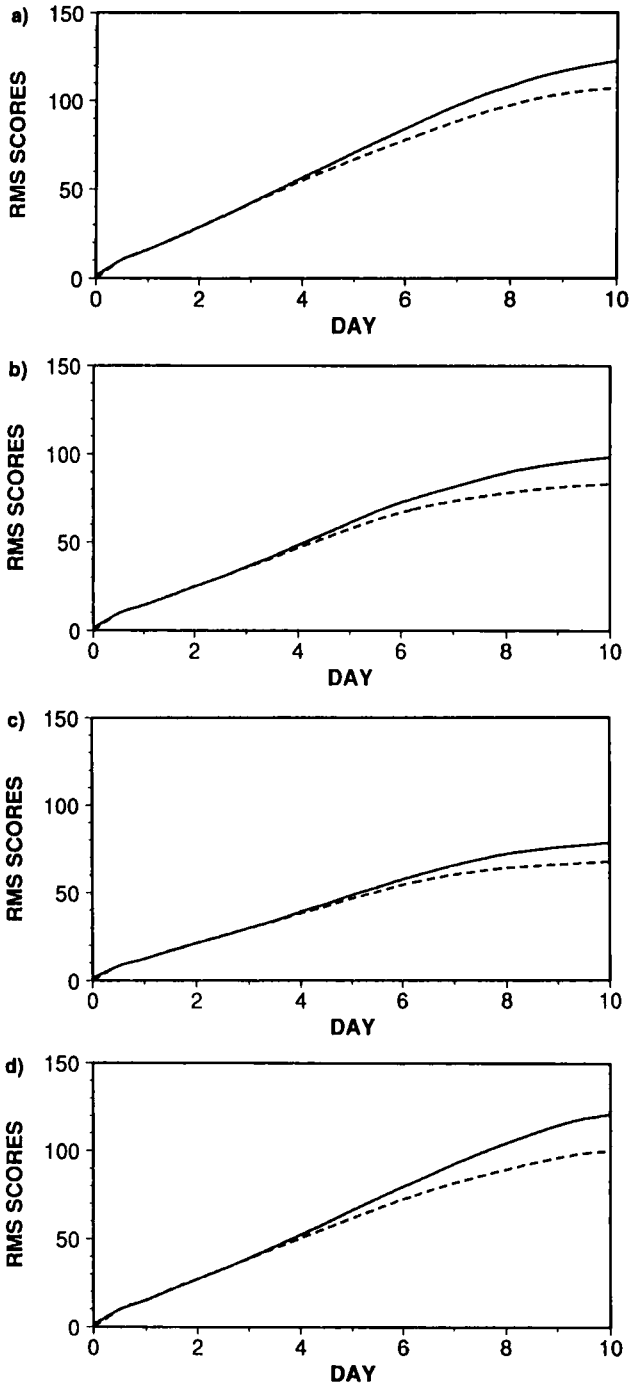

Figure 13. Seasonal average northern hemisphere r.m.s. error of the control forecast (solid line) and of the ensemble mean (dashed line) for (a) winter 1992/93, (b) spring 1993, (c) summer 1993 and (d) autumn 1993.

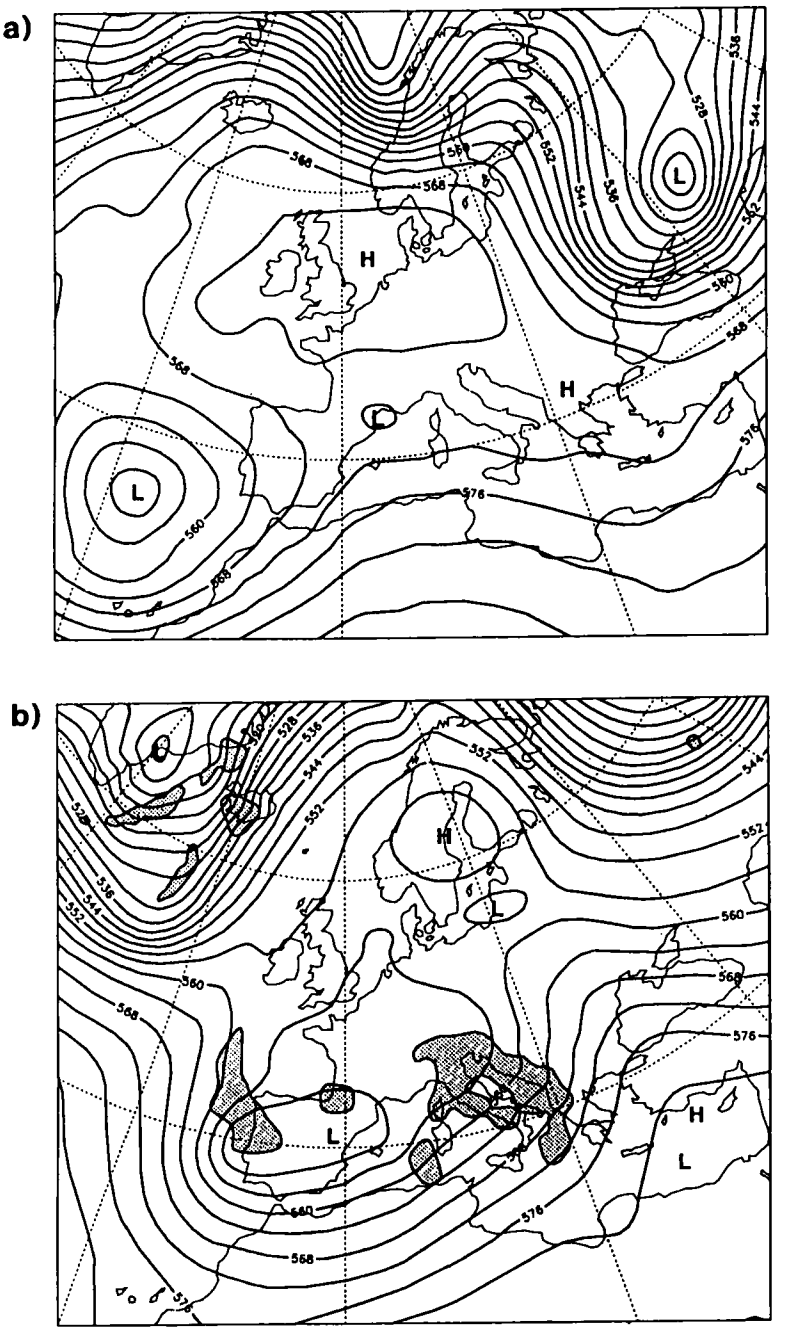

Figure 14. 500 hPa height from ECMWF analyses for 12 GMT (a) 30 October and (b) 6 November 1993. Superimposed on (b) (shaded) is shown the region where 24-hour rainfall centred on 6 November exceeded 10 mm.

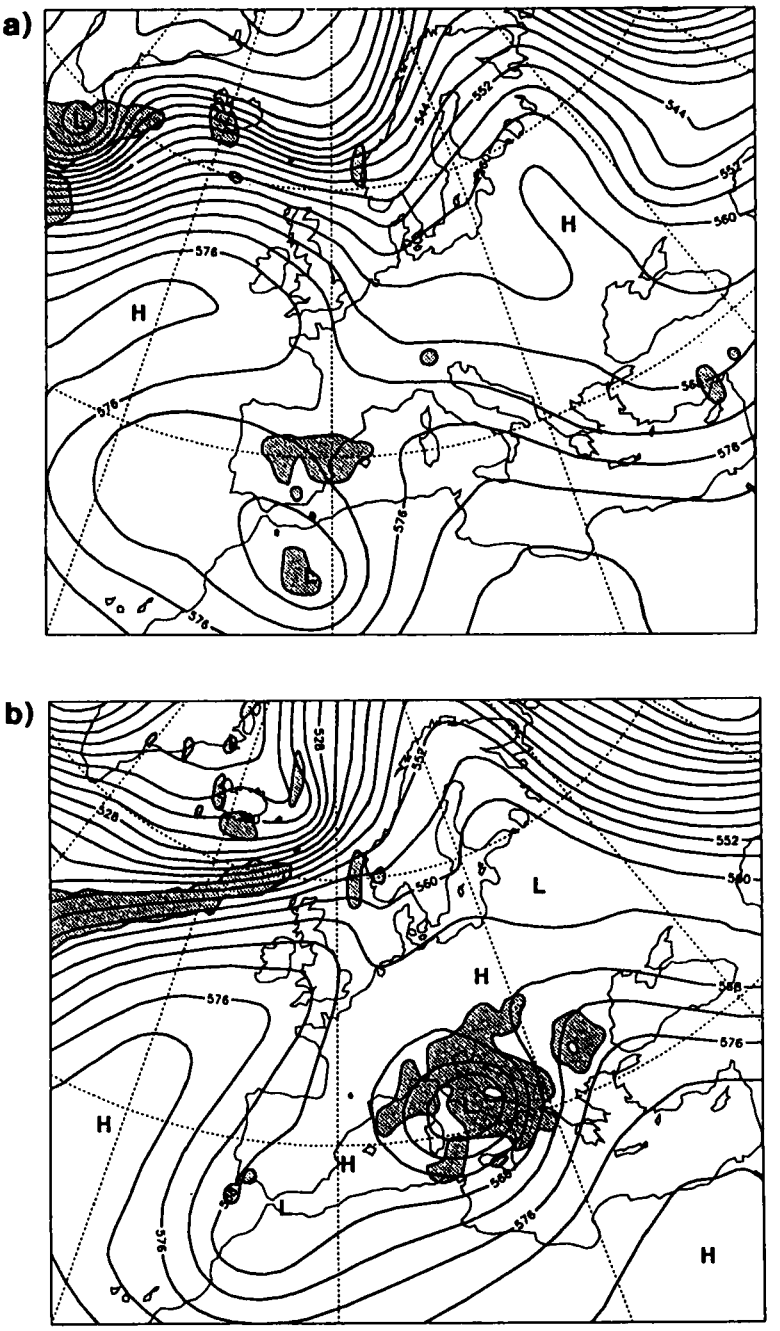

Figure 15. 500 hPa height from operational forecasts verifying on 12 GMT 6 November 1993; (a) day 7, (b) day 6 and (c) day 5. Superimposed is shown the region where operational 24-hour forecast rainfall centred on 12 GMT 6 November exceeded 10 mm.

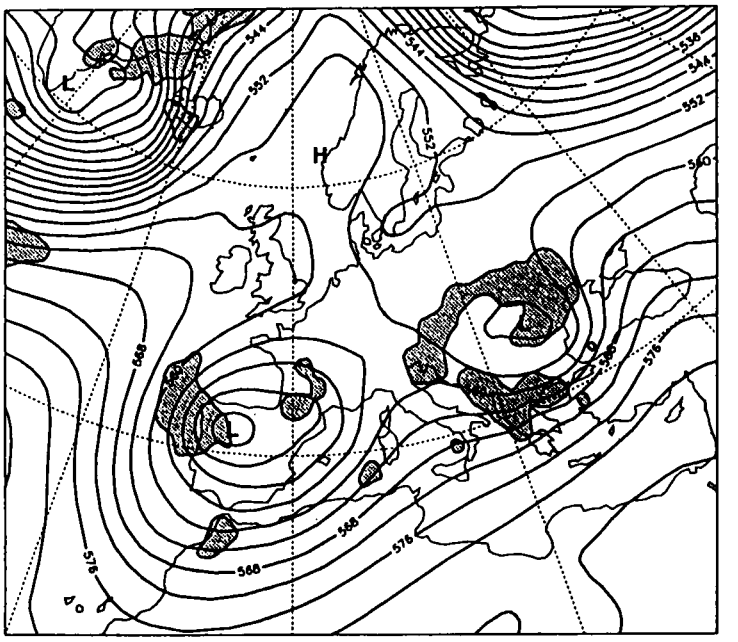

Figure 15. Continued.

Initial conditions ($t_0 = 30$ October 1993) are shown for the first of these forecasts, and verification is made for $t_v = 6$ November, corresponding to forecast-day 7 of the first ensemble, day 6 of the second ensemble, and day 5 of the third ensemble. 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).

Figure 14 shows the analysed 500 hPa height fields. At t_0 (Fig. 14(a)), a meridionally oriented dipole is already well established over western Europe and the eastern Atlantic. This pattern evolves slowly through to t_v (Fig. 14(b)) with low geopotential-height values dominating most parts of Europe, including two minima over the Iberian peninsula and over northern Italy. The region of high geopotential height has moved northward, with a maximum over Scandinàvia. The region where rainfall exceeded 10 mm day^{-1} at t_v is shown superimposed on the height field in Fig. 14(b). (This verification is actually taken from a 24-hour operational forecast from 5 November. We have checked, where observed data are available, that this estimate is accurate.) The largest rainfall rates are associated with the low centred over northern Italy.

Day 7, 6 and 5 operational forecast height and precipitation fields verifying at t_v are shown in Fig. 15. Over northern Europe the flow is too zonal at both days 7 and 6, resulting in significant precipitation over the west coast of Norway. Over southern Europe the position of a cut-off low is forecast for the Strait of Gibraltar at day 7, Italy at day 6, and northern Spain at day 5 (with a smaller low over Rumania). As a result, there is considerable inconsistency in the three rainfall predictions over southern Europe.

Figure 16 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. There is an overall tendency towards increased consistency between the 30 October ensemble and the 1 November ensemble. For the first ensemble, 16 out of the 32 members have an anomaly correlation of less than 0.6 with

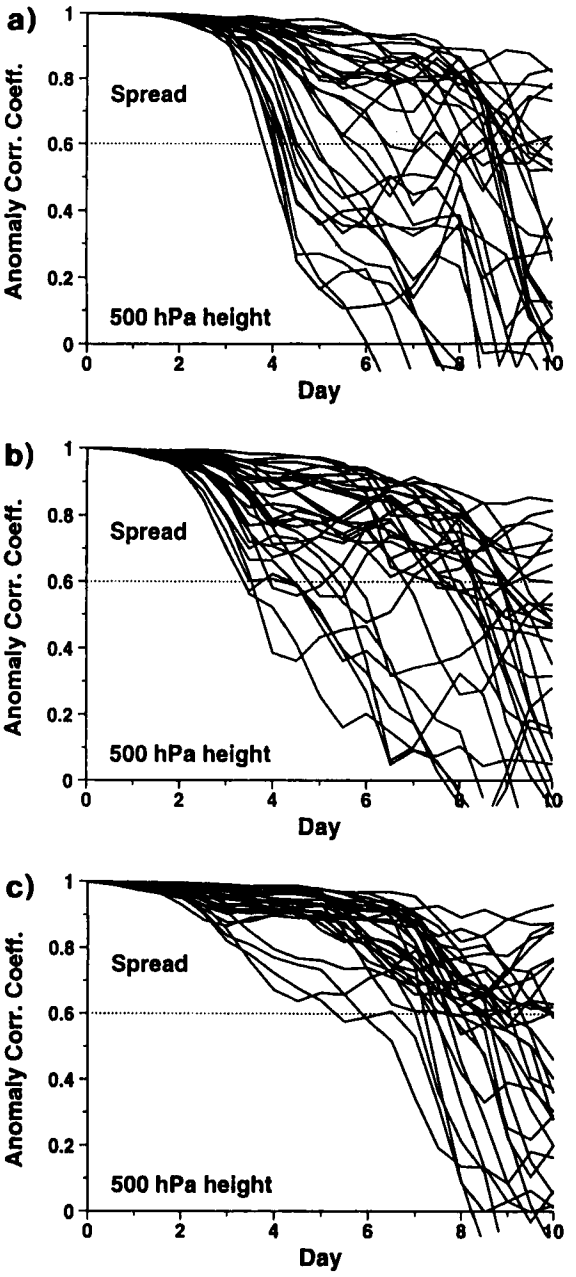

Figure 16. Anomaly correlation of 500 hPa height over Europe between individual ensemble members and control forecast as a function of forecast time; Panel (a) ensemble from 30 October, (b) ensemble from 31 October and (c) ensemble from 1 November 1993.

the unperturbed control forecast at day 7. For the second and third ensembles (again at day 7) this number reduces to 8 and 2 out of 32 respectively. Hence not only is the day-5 prediction for t_v more reliable than the day-7 or day-6 forecasts because of shorter lead time, but also the atmosphere is evolving towards an intrinsically more predictable phase from 30 October onwards.

Figure 17 shows an example of 500 hPa height stamp maps at t_v described in section 2(e)(i), for the day-6 ensemble forecast from 31 October. A forecaster studying this set of fields would have 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).

Figure 18 shows the results of an objective clustering of the day-6 forecasts, using the method described in section 2(e)(ii) (though, for illustrative purposes, the clustering is only applied to the fields in Fig. 17). For this ensemble, four clusters were selected. We show the r.m.s. and AC 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. Both these clusters are more skilful than the operational forecast, which has an anomaly correlation of 51%. 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.

Figure 19 shows probability maps for rainfall exceeding 10 mm day^{-1} at t_v (i.e. between 6 and 7 November), computed from the three ensemble forecasts as discussed in section 2(e)(iv). 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. 14(b)), the ensemble probability increases monotonically, whilst it decreases for regions in northern Europe, again in agreement with verification data.

Overall, this is a case where forecaster confidence would have been relatively low because of the large ensemble dispersion. Nevertheless, useful probabilistic information was provided by the three consecutive ensembles.

(*b*) *13–20 November 1993*

The analysed 500 hPa height for the initial date of the second case study (t_0 = 13 November 1993, two weeks after the first case study) is shown in Fig. 20(a), while Fig. 20(b) shows the analysis for 20 November, which has been chosen as the verification time t_v. Superimposed on the height contours for t_v are regions where the 850 hPa temperature anomaly either exceeded 4 K, or was less than −4 K. At t_0 (Fig. 20(a)), the flow is zonal across the Atlantic and western Europe, with a reversed-gradient height dipole at about 40°E. At t_v (Fig. 20(b)), this dipole has retrogressed and 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, and 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 dispersion of the ensemble started on t_0 is shown in Fig. 21. 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_v have an anomaly correlation of at least 0.6 with the control forecast.

The four clusters computed from the day-7 ensemble forecast verifying on t_v are shown in Fig. 22. The centroids of the three most populated clusters (including 13, 8 and 7 members respectively) all reproduce the blocking dipole, although they differ in the intensity and the position of the high and low centres. The centroid of the least populated cluster (with 5 members) has a strong high over Scandinavia, but the flow over southern Europe is far too zonal. As shown by the scores at the bottom of the panels in Fig. 22, in

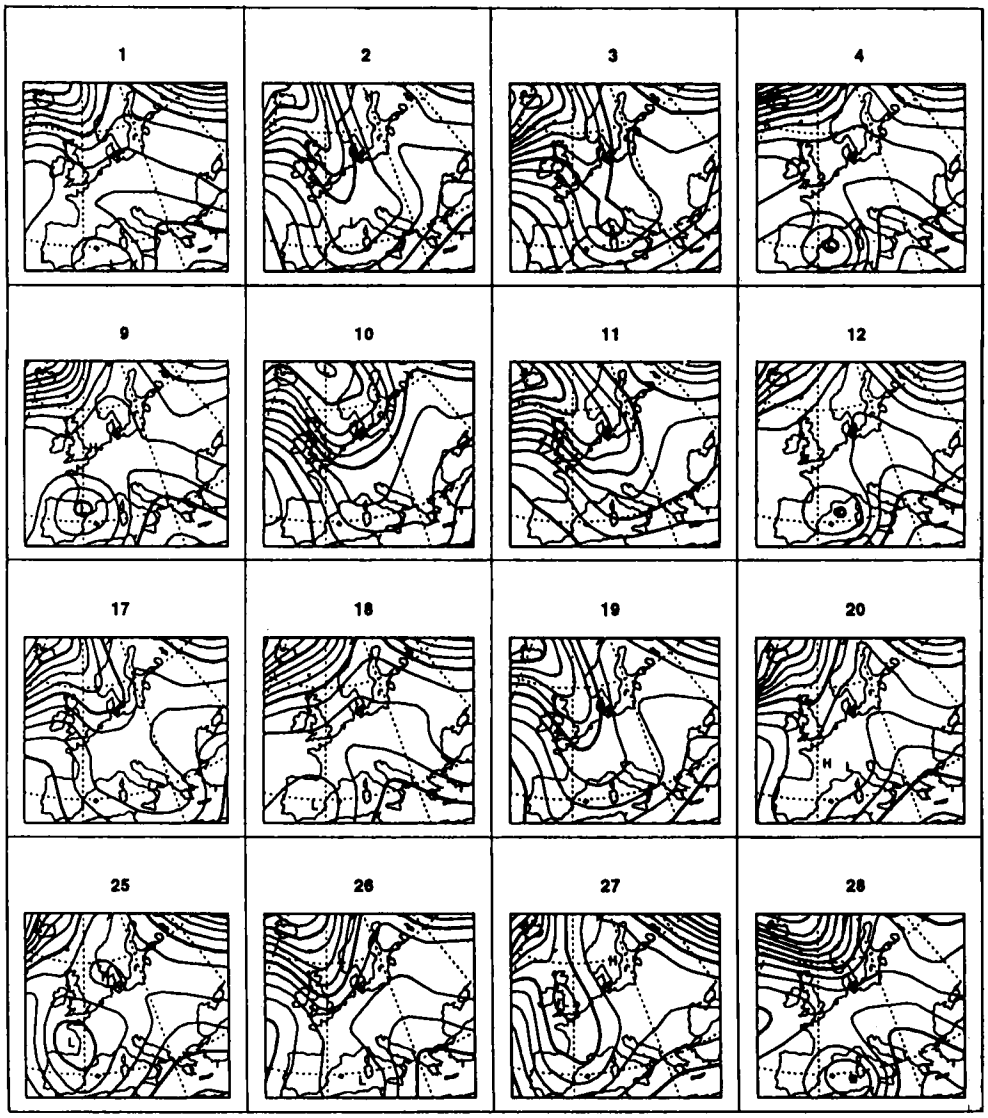

Figure 17. Individual forecast 500 hPa fields ('stamp maps') for day-6 ensemble forecast valid for 6 November 1993.

this case the cluster population is well correlated with the skill of the predicted centroids. It is interesting to note that, over western Europe, the circulation 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 predicted with fairly high consistency by the ensemble forecast.

In Fig. 23 we show the probability that the temperature anomaly either was greater than 4 K, or was less than −4 K. 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 (near the edge of the observed warm anomaly) there is a small probability of both cold and warm temperature anomalies.

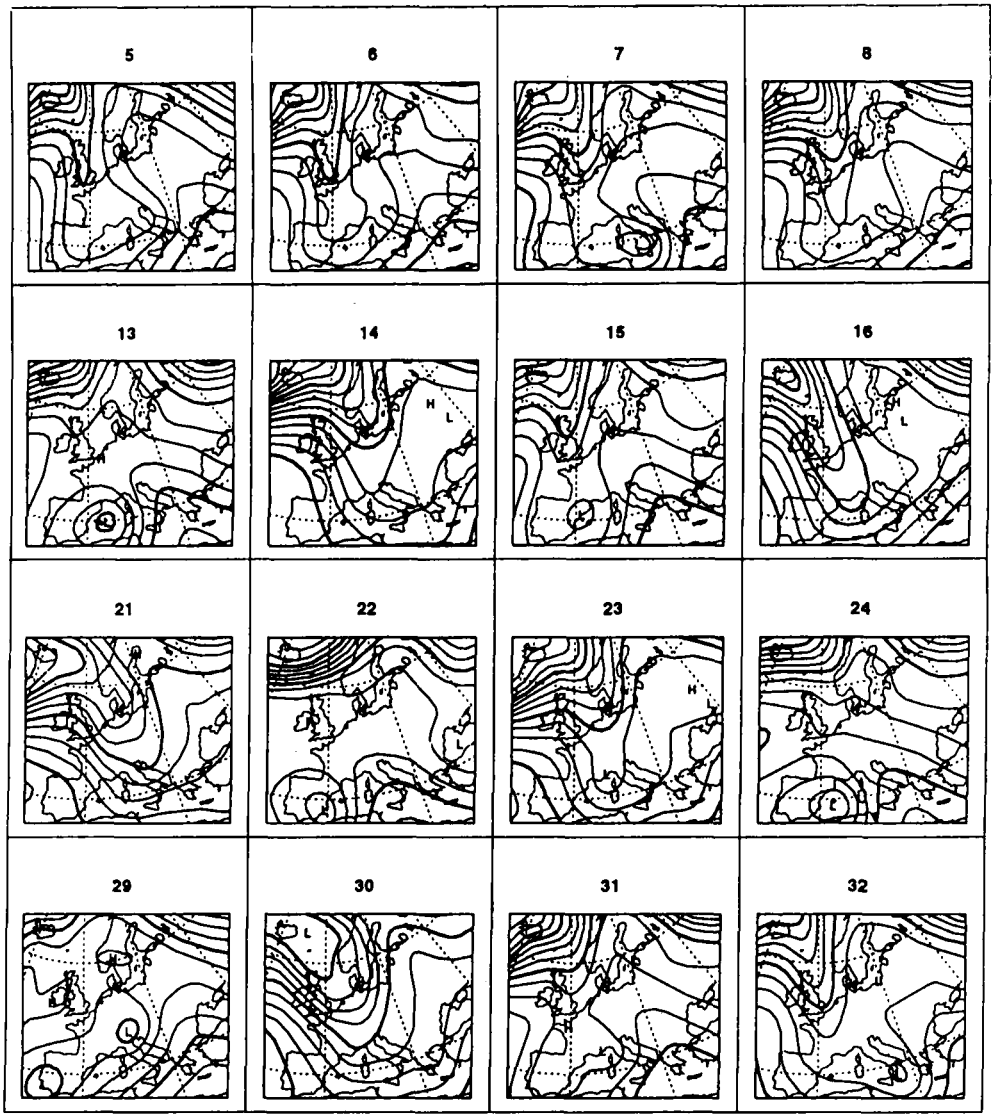

Figure 17. Continued.

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. 24(a) we show 850 hPa temperature probability plumes for the ensemble from 13 November for three locations: Longyearbyen (Spitzbergen), Paris, and Funchal (Madeira). For comparison, the control forecast and verifying analysis are also shown.

The three plumes indicate varying local predictability. For example the 30% contour disappears at day 5 in the first plume, day 9 in the second plume, and continues to the end of the forecast range in 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.

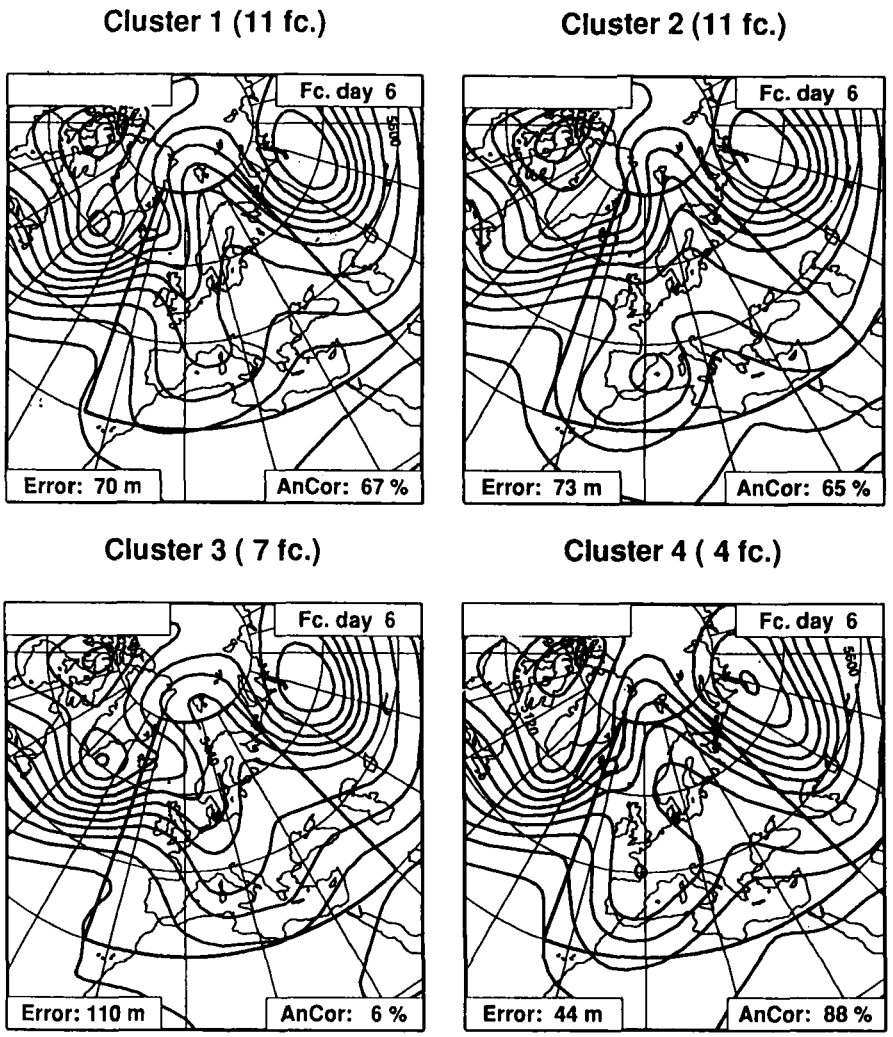

Figure 18. Clusters of 500 hPa height field from day-6 ensemble forecast valid for 6 November 1993.

5. Conclusions

Since the initial state of the atmosphere is known only approximately, a complete weather prediction should 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 become nonlinear.

Initial error can project onto many possible phase-space directions; therefore, some sampling procedure is necessary for the choice of initial perturbations for the ensemble forecast. We use the leading singular vectors of the linear propagator calculated from a primitive-equation model, linearized about a short-range forecast trajectory. These singular vectors, described in more detail in the companion paper by Buizza and Palmer (1995), identify the directions in phase space associated with maximum perturbation growth during the early parts of the forecast period. Arguments are given to suggest that these singular vectors may describe potentially important analysis-error structures using an energy-based inner product.

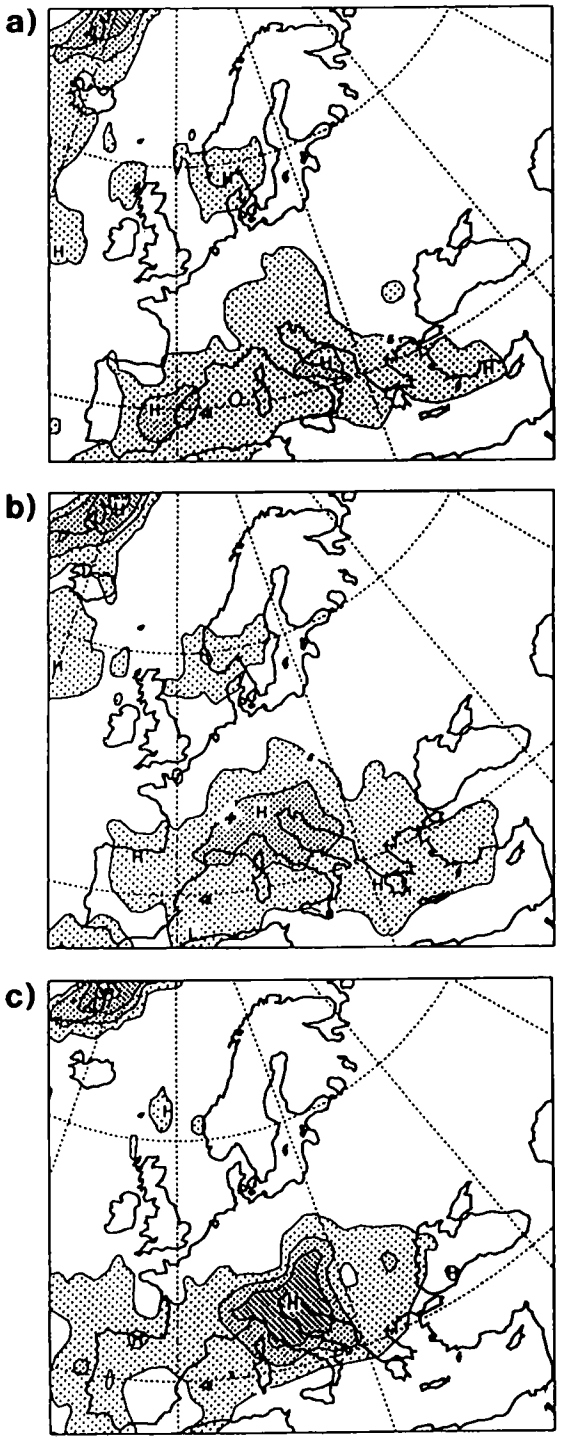

Figure 19. Maps of probabilities that total precipitation exceeds 10 mm day^{-1} between 12 GMT 5 November and 12 GMT 6 November 1993 from the ensemble originated from (a) 30 October, (b) 31 October and (c) 1 November. Contours 5, 35, 65 and 95%.

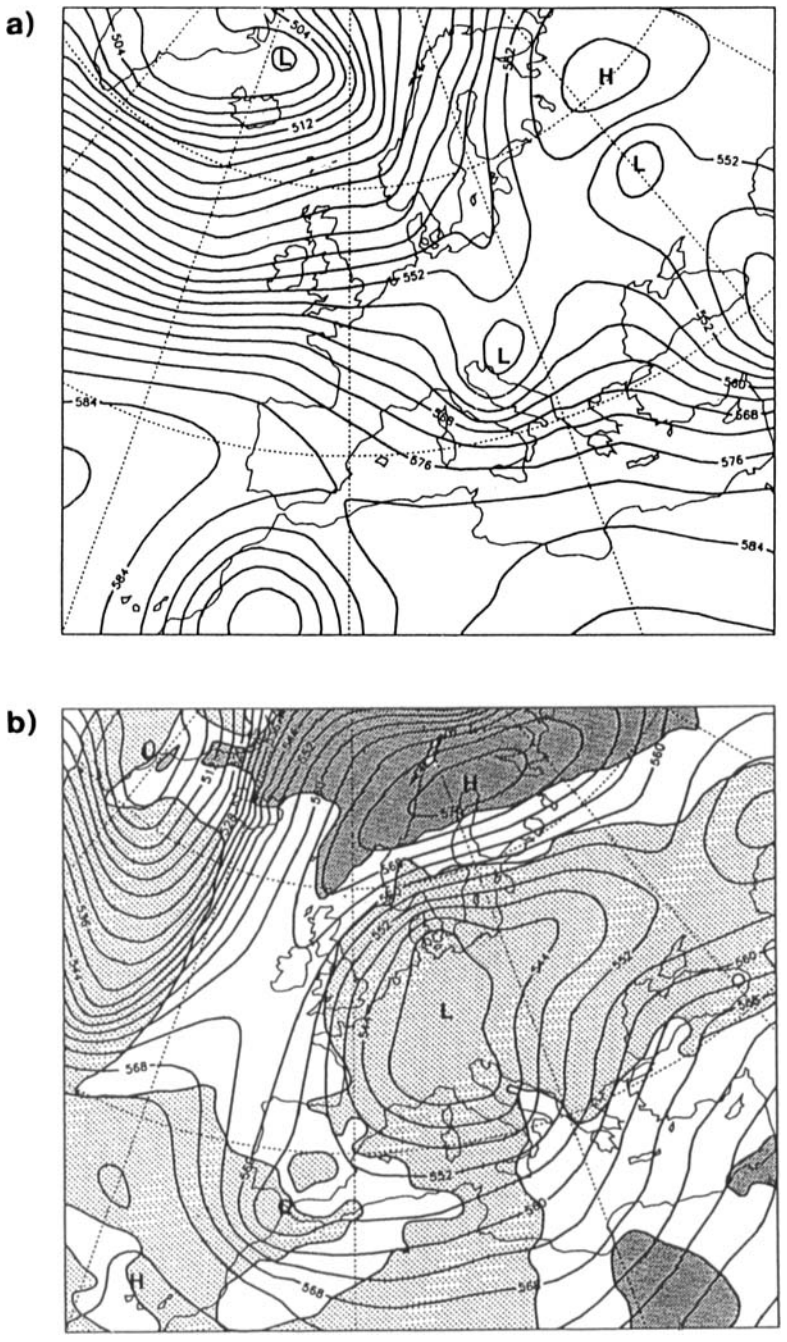

Figure 20. Analysed 500 hPa height fields for (a) 13 November and (b) 20 November 1993. Superimposed on (b) are regions where analysed temperature anomaly of 850 hPa was greater than 4 K (heavy shading) or less than −4 K (light shading).

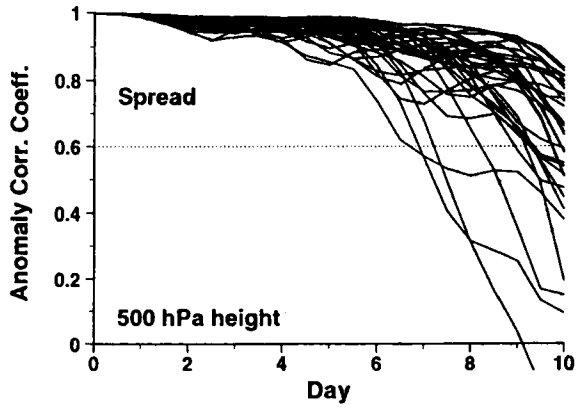

Figure 21. Anomaly correlation of 500 hPa height over Europe between individual ensemble members and control forecast as a function of forecast time. Ensemble from 13 November 1993.

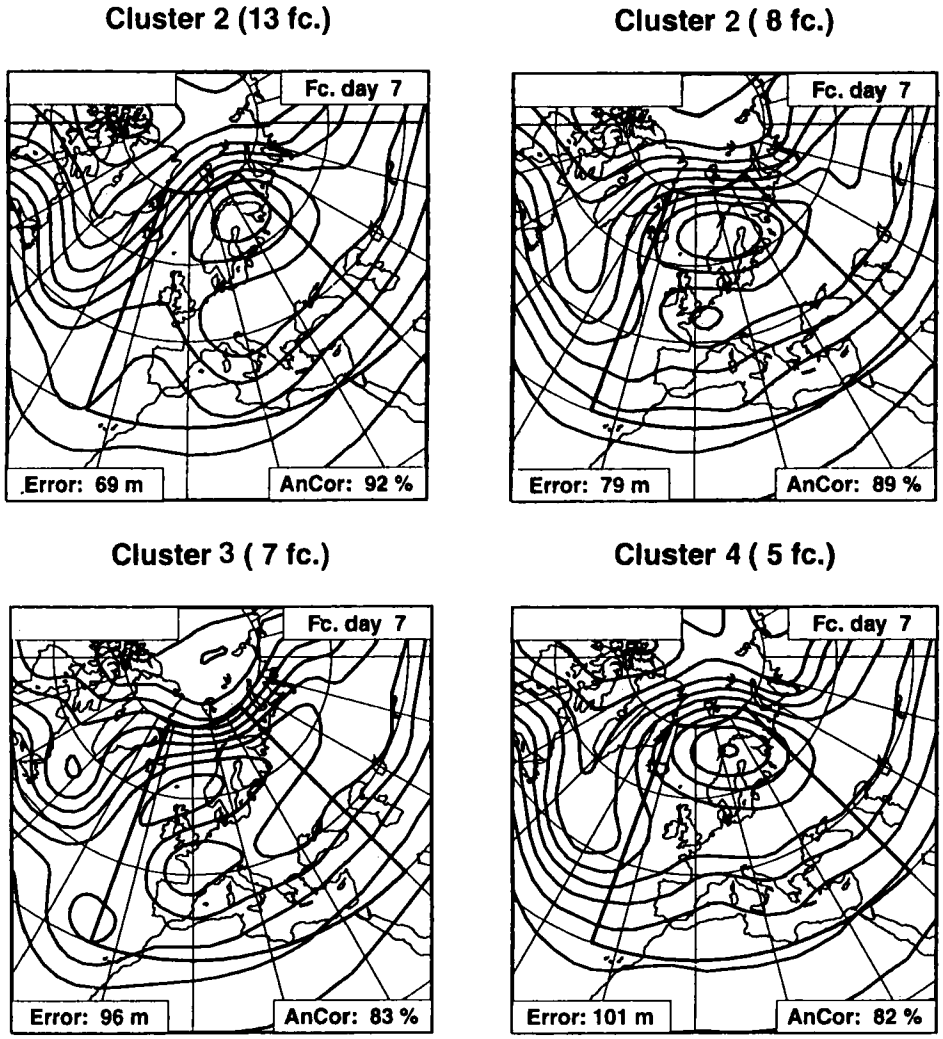

Figure 22. Clusters of 500 hPa height field from day-7 ensemble forecast valid for 20 November 1993.

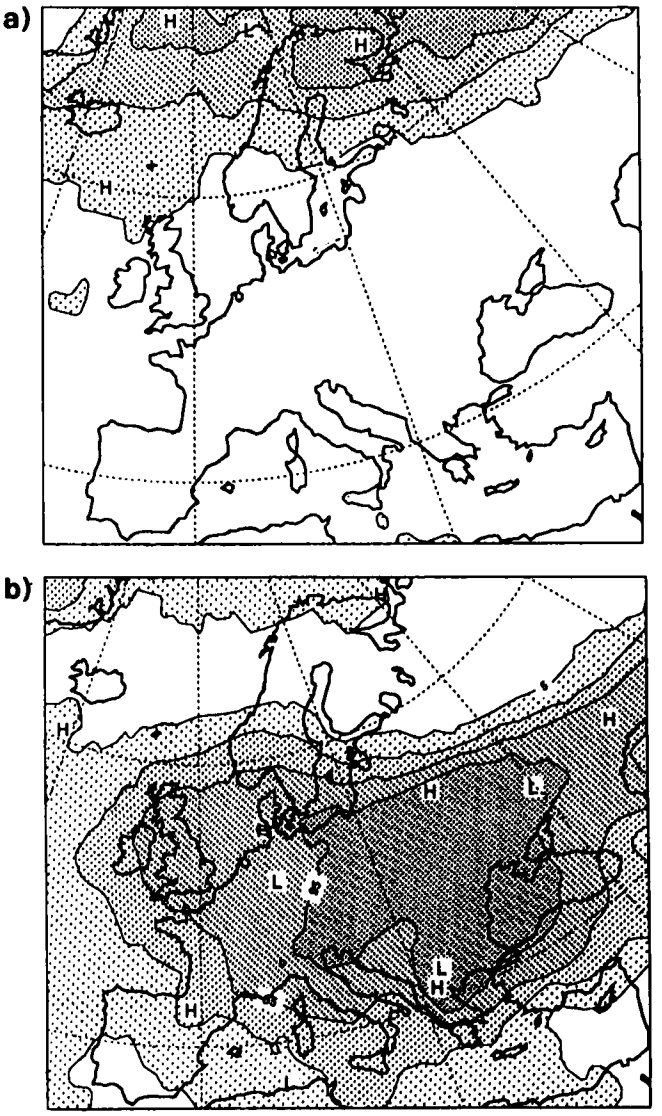

Figure 23. Maps of probabilities that 850 hPa forecast temperature anomaly for 20 November 1993 (a) exceeds 4 K and (b) is less than −4 K, from the ensemble initialized on 13 November 1993. Contours 5, 35, 65 and 95%.

Given their localized nature, individual singular vectors cannot be taken as likely representations of analysis-error fields. To construct realistic perturbations from linear combination of singular vectors, it was necessary firstly to find a set which covered a reasonable proportion of the northern hemisphere, secondly to perform a phase-space rotation in such a way as to delocalize the perturbations in physical space, and finally to determine the amplitude of the rotated perturbations from the operational analysis-error estimates.

Thirty-three-member ensemble forecasts using the T63L19 version of the operational model have been made routinely since December 1992. Ensembles from the first year of experimental trials have been validated in terms of contingency tables of spread/skill relationships and Brier scores of cluster probabilities. Distributions of ensemble-member scores have also been studied.

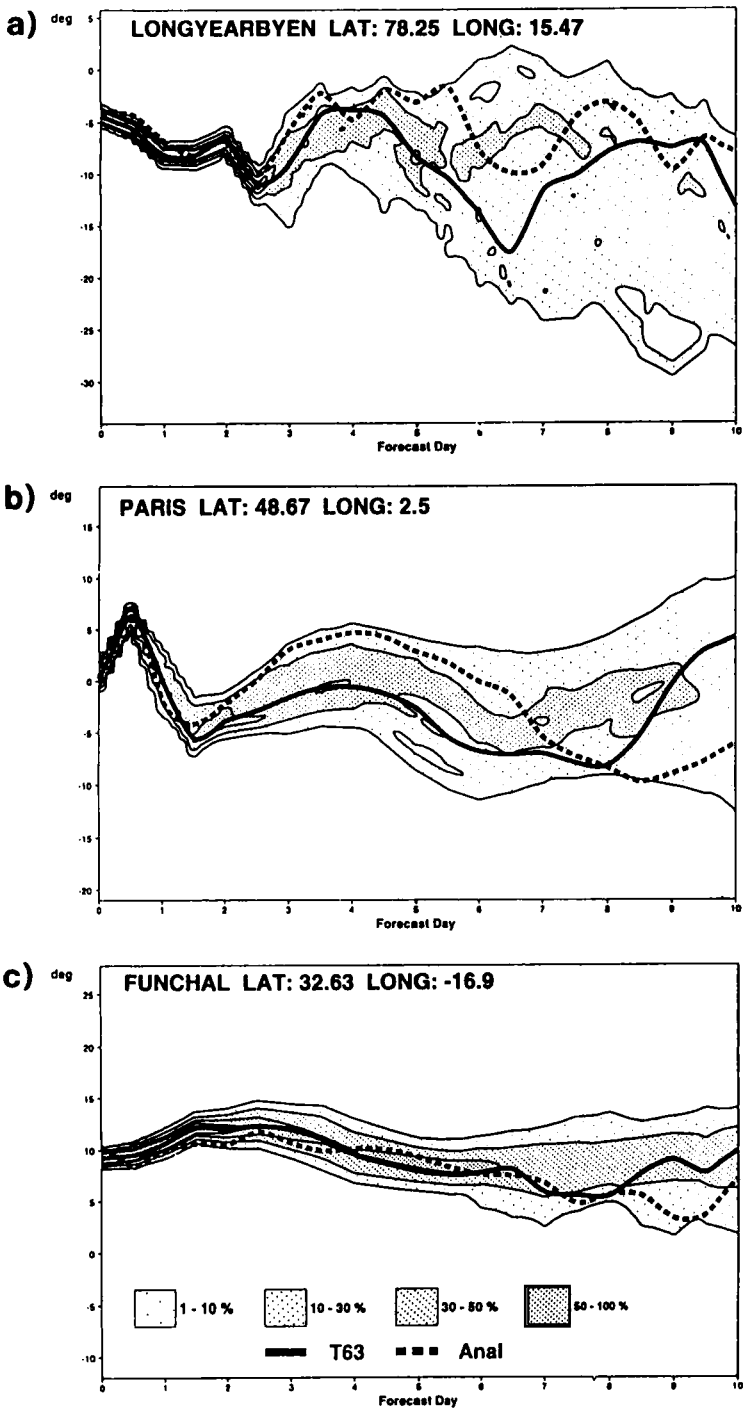

Figure 24. 850 hPa forecast temperature probability 'plumes' for the ensemble forecast from 13 November 1993, (a) Longyearbyen (Spitzbergen), (b) Paris and (c) Funchal (Madeira). Probabilities are based on 1 K intervals (see section 2(e)(iii) for details). The control forecast (solid) and verifying analysis (dashed) are also shown.

Results show a good relationship between ensemble skill and consistency, which was close to that obtained in a perfect-model environment. Probabilities of occurrence of Euro–Atlantic flow patterns were skilful up to forecast-day 8; verifications of individual ensemble members indicated that in the medium range about one third of the perturbed forecasts were closer to the verifying analysis than either the T63L19 or the T213L31 unperturbed forecasts. On the other hand, on a number of occasions the dispersion of the ensemble trajectories was insufficient to provide solutions close to the actual atmospheric state. We found evidence that model error contributed adversely at certain times. In particular, underprediction of ensemble spread during cases where strong ridges and/or a cut-off low developed in the real atmosphere was shown to be consistent with the underestimation of the frequency of these flow types in the model climatology.

A deficiency of the T63L19 model which is probably relevant for this type of behaviour is a gradual loss of eddy kinetic energy during the course of the integration. It is known that the operational T213L31 model does not suffer from a similar kind of problem. Ensemble experiments with a higher-resolution model (T106) will be performed in the near future; in addition to better energetics, an increased resolution would certainly be beneficial as far as probabilistic estimates of weather parameters like rainfall and near-surface temperature are concerned.

However, the advantages of a higher resolution of the forecast model must be weighed against the benefits of using the same amount of computer resources to increase the ensemble size. As shown by BP, there are many more unstable directions than are presently sampled, although associated with smaller growth rates. Diagnostic studies are currently under way at the ECMWF in order to estimate the number of singular vectors needed to explain most of the variance of short-range forecast errors.

These studies will also address this question of the optimal resolution for the computation of the singular vectors (e.g. Hartmann *et al.* 1995). Both the sensitivity patterns of Rabier *et al.* (1996) and the 'bred modes' of Toth and Kalnay (1993) show smaller-scale structures than our present T21 singular vectors. Routine computation of singular vectors at T42 is fairly expensive, but is feasible with the present ECMWF computer. In fact, the ECMWF EPS has been running with initial perturbations generated using T42 SVs since 14 March 1995, optimized over a 48 h time interval (the optimization time interval was increased from 36 h to 48 h on 23 August 1994 to have SVs optimized over the same time interval used to compute the sensitivity field), and with a smaller initial amplitude (i.e. the constant factor in Eq. (14) has been reduced to $\alpha = \sqrt{1.5}$, due to the larger growth of the T42 SVs compared with the T21 SVs).

After more than two decades of theoretical and experimental work in the area of probabilistic weather prediction by dynamical methods, operational ensemble forecasting is now becoming a reality. 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 the two case studies above. Initial reactions 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.

Acknowledgements

The implementation of the Ensemble Prediction System has been the result of the work of many ECMWF staff and consultants in both the Research and the Operations Department during the past few years. Our appreciation goes to all of them, and in particular to R. Mureau and J. Tribbia, who have contributed to the recent developments of the system

both scientifically and technically. We thank R. Gelaro for the results shown in Fig. 4. We also thank A. Hollingsworth, A. Simmons and P. Courtier for helpful comments on an earlier version of the manuscript. The Lanczos scheme used here for the computation of singular vectors was written by Prof. B. N. Parlett, and the code was kindly supplied to us (in pre-release form) by the Numerical Algorithm Group.

APPENDIX

Initial error of control and perturbed forecasts

Let us consider the following vectors in the phase space of a NWP model:
$\mathbf{X}_t$: true state of the atmosphere;
$\mathbf{X}_0$: initial state of the control forecast (= operational analysis);
$\delta\mathbf{X}_i$: ith perturbation to the operational analysis;
$\mathbf{E}_0 = \mathbf{X}_0 - \mathbf{X}_t$: operational analysis error.
The squared norm of the initial error of the ith perturbed forecast is given by:

$$\|(\mathbf{X}_0 + \delta\mathbf{X}_i) - \mathbf{X}_t\|^2 = \|\mathbf{E}_0 + \delta\mathbf{X}_i\|^2 = \|\mathbf{E}_0\|^2 + \|\delta\mathbf{X}_i\|^2 + 2\langle\mathbf{E}_0;\, \delta\mathbf{X}_i\rangle \tag{A.1}$$

where $\langle\ .\ .\ ;\ .\ .\ \rangle$ represents the chosen inner product in phase space; it will be smaller than $\|\mathbf{E}_0\|^2$ if:

$$\langle\mathbf{E}_0;\, \delta\mathbf{X}_i\rangle < -\frac{1}{2}\|\delta\mathbf{X}_i\|^2. \tag{A.2}$$

Let us assume that the N perturbations are orthogonal and have unit norm. Clearly the relationships above are still valid if $\mathbf{E}_0$ represents the projection of the analysis error onto the N-dimensional sub-space spanned by the perturbations (rather than the full field). Then we can write

$$\|\mathbf{E}_0\|^2 = \sum_{i=1}^{N}\langle\mathbf{E}_0;\, \delta\mathbf{X}_i\rangle^2. \tag{A.3}$$

If $\mathbf{E}_0$ has exactly the same norm as the perturbations (= 1 according to our normalization), then the condition (A.2) cannot be satisfied by more than three perturbations, because in this case the projection on each perturbation which satisfies (A.2) accounts for more than 1/4 of the squared norm of the analysis error.

More generally, we can assume that $\mathbf{E}_0$ is unbiased and has an isotropic, multi-normal PDF in the perturbation sub-space; then, the PDF for each individual projection

$$x_i = \langle\mathbf{E}_0;\, \delta\mathbf{X}_i\rangle$$

is given by:

$$\mathrm{PDF}(x_i) = \frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{x_i^2}{2\sigma^2}\right) \tag{A.4}$$

where $\sigma^2 = \overline{\|\mathbf{E}_0\|^2}/N$.

The probability $P_\pm$ that the ith perturbation (with either positive or negative sign, as in the EPS) generates an initial state which is closer to the true atmospheric state than the operational analysis is given by the integral of (A.4) for $|x_i| > 1/2$. From this value, the probability $P(n)$ that n out of N positive and N negative perturbations will satisfy this condition can be computed as:

$$P(n) = \frac{N!}{n!(N-n)!}P_\pm^n(1 - P_\pm)^{N-n}. \tag{A.5}$$

If $\overline{\|\mathbf{E}_0\|^2} = 1$ (that is, if the *average* analysis-error amplitude is equal to the perturbation amplitude), and $N = 16$ as in the EPS, one obtains the following values:

$$P_{\pm} = 0.046$$

$$P(0) = 0.474$$

$$P(1) = 0.362$$

$$P(2) = 0.130$$

$$P(3) = 0.029$$

$$P(4) = 0.004.$$

References

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

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

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

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

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

Brankovic, C., Palmer, T. N. and Ferranti, L. 1994 Predictability of seasonal atmospheric variations. *J. Climate*, **7**, 218–237

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

1994b Localization of optimal perturbations using a projection operator. *Q. J. R. Meteorol. Soc.*, **120**, 1647–1682

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

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

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

Déqué, M. and Royer, J. F. 1992 The skill of extended-range extra-tropical winter dynamical forecasts. *J. Climate*, **5**, 1346–1356

Ebisuzaki and Kalnay, E. 1991 Research activities in atmospheric and oceanic modelling. WMO Report n. 15

Ehrendorfer, M. 1994 The Liouville equation and its potential usefulness for the prediction of forecast skills. Part I: Theory. *Mon. Weather Rev.*, **122**, 703–713

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

Farrell, B. F. 1990 Small error dynamics and the predictability of atmospheric flows. *J. Atmos. Sci.*, **47**, 2409–2416

Ferranti, L., Molteni, F., Brankovic, C. and Palmer, T. N. 1994 Diagnosis of extra-tropical variability in seasonal integrations of the ECMWF model. *J. Climate*, **7**, 849–868

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

1971b On stochastic dynamic prediction. II. Predictability and utility. *Mon. Weather Rev.*, **99**, 927–938

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

Hartmann, D. L., Buizza, R. and Palmer, T. N. 1995 Singular vectors: the effect of spatial scale on linear growth of disturbances. *J. Atmos. Sci.*, **52**, 3885–3894

Hoffman, R. N. and Kalnay, E. 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
Lacarra, J. F. and Talagrand, O.	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. Weather Rev.*, **102,** 409–418
Lorenz, E. N.	1965	A study of the predictability of a 28-variable atmospheric model. *Tellus*, **17,** 321–333
	1969	The predictability of a flow which possesses many scales of motion. *Tellus*, **21,** 289–307
	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 Palmer, T. N.	1991	A real-time scheme for the prediction of forecast skill. *Mon. Weather Rev.*, **119,** 1088–1097
	1993	Predictability and finite-time instability of the northern winter circulation. *Q. J. R. Meteorol. Soc.*, **119,** 269–298
Mureau, F., Molteni, F. and Palmer, T. N.	1993	Ensemble prediction using dynamically conditioned perturbations. *Q. J. R. Meteorol. 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., Molteni, F., Mureau, R., Buizza, R., Chapelet, P. and Tribbia, J.	1993	'Ensemble prediction'. ECMWF seminar proceedings 'Validation of models over Europe: Vol 1'. ECMWF, Shinfield Park, Reading, UK
Rabier, F., Klintier, E., Courtier, P. and Hollingsworth, A.	1996	Sensitivity of forecast errors to initial conditions. *Q. J. R. Meteorol. Soc.*, **122,** 121–150
Strang, G.	1986	*Introduction to applied mathematics*. Wellesley-Cambridge press
Thépaut, J.-N., Hoffman, R. N. and Courtier, P.	1993	Interactions of dynamics and observations in a four-dimensional variational assimilation. *Mon. Weather Rev.*, **121,** 3393–3414
Toth, Z. and Kalnay, E.	1993	Ensemble forecasting at NMC: the generation of perturbations. *Bull. Am. Meteorol. Soc.*, **74,** 2317–2330
Tracton, M. S., and Kalnay, E.	1993	Operational ensemble prediction at the National Meteorological Center: Practical Aspects. *Weather and Forecasting*, **8,** 379–398
Tracton, M. S., Mo, K., Chen, W., Kalnay, E., Kistler, R. and White, G.	1989	Dynamical extended range forecasting (DERF) at the National Meteorological Center. *Mon. Weather Rev.*, **117,** 2230–2247