TECHNICAL MEMORANDUM

153

Extended range predictions with ECMWF models. III. Time lagged ensemble forecasting

C. Brankovic, T.N. Palmer, F. Molteni and U. Cubasch

Research Department

June 1989

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

Abstract

A set of 16 ensembles of time-lagged extended-range foreasts have been run at different times in the annual cycle using the T63 version of the European Centre for Medium-Range Weather Forecasts (ECMWF) operational model. Each ensemble was composed of 9 integrations from consecutive 6 hourly analyses.

Theoretical properties of ensemble-mean skill and ensemble spread are studied using a simple model of error growth with a parametrization of the ECMWF model error. The impact of systematic error on the potential improvement in skill of the ensemble-mean forecast is discussed. The presence of model errors considerably reduces the gain from ensemble averaging.

In practice, about a third of the ensemble-mean forecasts, at forecast days 11-20, were more skilful than both persistence and climate, and, in addition, were more skilful than the latest member of the ensemble. At days 21-30, only one of the ensemble-mean forecasts was similarly skilful. Whilst there is an overall hemispheric-scale correlation between ensemble spread and skill, a substantial part of this reflects the impact of the annual cycle on both quantities. In the winter period, however, no clear spread/skill correlation was found.

Within the winter period, there was considerable case-to-case variability in forecast skill. The January 1986 ensemble was the poorest of all the ensembles; the February 1986 ensemble was one of the most skilful. The different character of these two ensembles was shown by considering phase space trajectories of the ensemble forecasts in the plane spanned by the two principal forecast EOFs of 500 mb height. During the first 15 days, the trajectories of the January ensemble forecasts were consistent with each other, but contrary to the observed atmospheric trajectory (which was associated with the onset of blocking over Europe). During the last 15 days, as the January ensemble forecasts migrated from positive to negative PNA index, the trajectories dispersed quite strongly, becoming disordered. By contrast, the trajectories of the February forecasts remained both mutually consistent and in agreement with the real atmosphere's trajectory throughout most of the forecast period.

In order to investigate possible reasons for the failure of the January 1986 ensemble to develop the European block during the first half of the forecast period, two further integrations were made. In the first, the control integration was rerun with a more recent version of the ECMWF model. Development of the block continued to be missed, though the trajectory of this forecast in phase space in the medium range was quite different to any of

the members of the original ensemble. This suggests that a methodology for Monte Carlo forecasting should include perturbations to model formulation as well as perturbations to initial conditions.

Secondly, motivated by known systematic errors in the model's simulation of tropical divergence, and the diagnostic study of Hoskins and Sardeshmukh (1987), an integration was run in which the model's tropical fields were relaxed towards the verifying analysis. In this integration, substantial ridging over the Euro/Atlantic area occurred, and the extratropical skill scores were noticably improved. Phase space trajectories confirm this partial success on a hemispheric scale. However, the intensity of the block was not well captured. It would appear therefore that failure to capture the block is partially associated with the problem of predictability, and partially with the problem of the model systematic errors.

The EOF decomposition of the ensemble forecasts was also used as an objective criterion to test for clustering within the ensemble. According to this analysis, one of the January forecasts (but not the latest) showed quite distinct behaviour, and its extended-range skill was well above the other members of the ensemble.

The cluster analysis was performed on all winter forecasts. When the forecasts were categorized into three clusters (which vary from ensemble to ensemble), it is found that the skill at days 11-20 of one of the categories is always better than the ensemble-mean forecast. However, this more skilful group of forecasts does not always correspond to the more densely populated ensemble. We suggest that this may be associated with a problem of sampling.

Given the results from the present sample of forecasts we believe that prediction beyond the medium range is not currently viable. However, when forecast systematic error, particularly in the tropics, is reduced, and when techniques that can identify the most rapidly growing perturbations have been fully developed, then the cluster analysis suggests that probabilistic forecasting of extratropical time-mean weather may be feasible in the time range of up to two-three weeks into the future.

1. INTRODUCTION

Experimental programmes to study the skill of numerical weather prediction (NWP) models beyond the mean limit of instantaneous deterministic predictability have been under way at a number of major centres over the last few years (see, for example, ECMWF, 1988). The companion paper by Palmer et al. (1989; hereafter referred to as I), shows that, on average, the models lose skill in predicting weather regime transitions by about day 15. On the other hand, individual forecasts exist where skill is maintained longer into the integration. The value of these more skilful forecasts cannot be realized unless good a priori estimate of forecast reliability can be given.

The prediction of forecast skill for extended, medium and short range forecasting is a topic of much interest at present. Possible techniques for estimating forecast reliability include the use of stochastic-dynamic models (Epstein, 1969), Monte Carlo forecasting (Leith, 1974), and statistical prediction techniques based on data from time-series of deterministic integrations (Palmer and Tibaldi, 1988).

Problems of formulation, closure, and sheer computational cost will require solution before the stochastic dynamic technique can be considered viable, at least beyond the short range. On the other hand, statistical techniques do not appear to offer a satisfactory solution for the extended-range problem, not least because of the very limited archive sample with which to derive, for example, regression coefficients.

The notion of Monte Carlo forecasting originated as an alternative to the stochastic-dynamic technique, and was defined by a finite sample or ensemble of integrations of a deterministic model, the initial conditions of each individual integration being obtained by randomly perturbing the spectral coefficients (or grid point values) of an operational analysis. As such, much of the perturbation amplitude may be lost through the initialisation procedure, especially in the height field (Baumhefner, 1988), even when the wind and mass fields of the perturbations are in geostrophic balance (Hollingsworth, 1980).

However, a simple and convenient way (in an operational environment) of effectively obtaining an ensemble of balanced initial states for a Monte Carlo ensemble, is through the time-lagged technique advocated by Hoffman and Kalnay (1983). In the current operational analysis/forecast system at the European Centre for Medium-Range Weather Forecasts (ECMWF), where analyses are produced at 6 hourly intervals, an n-member ensemble at time t=0 would be composed of the operational analysis at t=0, the 6hr forecast initialised from the analysis at t=-6 hr, and so on to the 6(n-1)hr forecast from the analysis at t=-6(n-1)hr (see Fig. 1). The effective perturbations at t=0 are considered to be the forecast errors at t=0. Preliminary tests of extended-range ensemble forecasting using this

technique were performed at ECMWF with low-resolution version of the operational model (Molteni et al., 1988).

Whilst this technique does not suffer from the problem of unbalanced perturbations, it clearly represents a compromise in other respects. In data void regions, for example, where the analysis is taken from the previous 6hr forecast (the 'first guess' field), the effective perturbation will be zero. Moreover, the larger the value of n, the more unlikely that the effective perturbation is representative of analysis uncertainty at t=0. Within a given data assimilation system, this effectively limits the size of the ensemble, and gives rise to a nonisotropic distribution of effective perturbations about the t=0 analysis which may favour the latest members of the ensemble.

For extended range forecasts, however, these potential drawbacks are probably not serious enough to outweigh the advantage of the technique in terms of its relative ease of implementation in an operational environment. Beyond the first few days, initial analysis errors propagate downstream and project onto the dominant modes of instability of the flow field (e.g. Palmer, 1988). Ultimately, the structure of forecast errors may depend more on the geographical distribution of these modes of instability than on the distribution of the initial analysis errors. Furthermore, as will be shown explicitly in this study, whilst in the first ten days of the forecast period, forecasts from two adjacent analyses will tend to resemble each other more than two forecasts from non-adjacent analyses, this relationship breaks down later in the forecast period. This suggests that beyond about the first ten days of the forecast period, members of the ensemble can indeed be treated as, a priori, equally likely.

In the present paper, we study the problem of predicting extended-range forecast skill using the time lagged technique over the entire annual cycle. Following earlier studies (Leith, 1974; Hoffman and Kalnay, 1983) we recognise that at least 10 integrations are necessary to form a reasonable sized ensemble. This would cause practical computing problems if it was necessary to integrate at the currently operational T106 resolution of the ECMWF model. However, results from the companion paper of Tibaldi et al. (1989; hereafter referred to as II) have indicated that the extended-range performance of the ECMWF model is not significantly worse at T63 resolution than at T106 resolution. The time-lagged ensembles described in this paper have therefore been integrated at T63 resolution.

In section 2 we describe the forecast ensembles that have been made. In section 3 we discuss the skill of ensemble-mean forecasts that can be deduced within a simple theoretical framework using a parametrization of known model systematic errors. In section 4, we

describe the ensemble-mean skill of the integrations, and examine whether the hemispheric scale ensemble-mean spread can be used to assess forecast reliability. We show the synoptic development in the extended range of some of the most skilful and least skilful ensemble-mean forecasts.

In section 5 we discuss in more detail two of the forecast ensembles from consecutive months. One of these ensembles is the poorest of our set, the other is one of the most skilful. The poor forecast ensemble was initialised about two weeks preceding the development of an intense European block, diagnosed in detail by Hoskins and Sardeshmukh (1987). We discuss the forecast performance using both conventional synoptic maps, and using less conventional phase-space trajectories. Additional experiments are described in an attempt to assess the mechanisms that may have been responsible for the failure of the first of these case studies.

Having established a methodology for extended-range ensemble forecasting, it is necessary to consider techniques for postprocessing the results. Though the ensemble-mean forecast is a simple and convenient way of collating the results from individual members, the RMS error of the ensemble-mean forecast is trivially smaller than the mean RMS error of the members of the ensemble, because the ensemble mean is a smoothed field biased towards the model climate. An estimate of the reliability of the ensemble-mean forecast, in a model without external error, can be obtained from the ensemble standard deviation.

More generally, it is possible to use the information available within the forecast ensemble to indicate possible alternate forecast flows. The association of probabilities to each of the possible alternatives captures the essence of forecasting beyond the limit of deterministic probability. In section 6, we apply a cluster analysis algorithm on the forecast empirical orthogonal functions to produce sub-ensembles. These give some insight into the question of predictability during periods of blocking. Using this cluster analysis technique, 'probability' forecasts are made for the winter ensemble forecasts. Concluding remarks are made in section 7.

2. THE DATABASE AND EXPERIMENT DESIGN

An ensemble of time-lagged forecasts in the present study is composed of 9 members, each member being an extended-range prediction with the T63 version of the ECMWF operational spectral model. The initial data for each member of the ensemble were ECMWF operational analyses separated (lagged) by 6 hours. There is therefore a 48-hour period spanning the first and the last member of the ensemble (Fig. 1). This was a natural choice, since ECMWF analyses are available at 6-hour intervals: 00Z, 06Z, 12Z and 18Z. The last forecast of an ensemble starts from a 12Z analysis at 'D0' and is integrated for 30 days. All

verifying times are relative to D0 and therefore this last forecast is referred to as the control run. The first forecast of an ensemble starts from 12Z analysis at D-2, i.e. 48 hours before the initial date of the last forecasts, and is integrated for 32 days.

The complete list of the ensemble forecast dates is given in Table 1. From September 1985 until March 1986 the time-lagged forecasts were performed every month. After this period they were run at 3-month intervals.

The observed sea surface temperatures (SSTs), which are part of the initial data, were kept constant during the course of integration. Since, in the operational data assimilation scheme, SSTs are updated daily at 12Z, it is in principle possible to have different SSTs within the same forecast ensemble. However, these differences (if any) are negligible, by virtue of their slowly varying nature.

The ensemble mean is computed as a simple arithmetic average from all members. There is no weighting of individual forecasts, (Hoffman and Kalnay, 1983), because of our interest in extended range predictions where all weights would be essentially identical (c.f. Molteni *et al.*, 1986 and below).

Until March 1986, the same version of the ECMWF operational model was used for all forecast ensembles: 16 levels, envelope orography and physical parametrization as defined in May 1985 (Tiedtke *et al.*, 1988). The June 1986 ensemble was run with the model in which vertical resolution was increased to 19 levels by including three additional levels between 10 and 150 mb (Simmons *et al.*, 1989). From September 1986 the ensembles were run with a model which included the parametrization of gravity wave drag (Palmer *et al.*, 1986; Miller *et al.*, 1989). From the beginning of 1988 the vertical diffusion scheme above the planetary boundary layer was removed. This inhomogeneity in model data is an unavoidable consequence of our desire to keep the extended-range programme relevant to the needs of the operational forecasting system, particularly with regard to the diagnosis of systematic error.

The ECMWF operational analyses were used for the objective verification of the ensembles and individual forecasts. A monthly climate, derived from six years (1979 to 1984) of ECMWF analyses, was employed to evaluate the forecast anomaly correlation coefficient of skill.

In the following sections we denote forecasts from December, January and February as 'winter' forecasts; forecasts from June, July and August as 'summer' forecasts; and forecasts from all other times of year as 'transition' forecasts.

No.	Date	Comment/Model change
1.	16 May 1985	May 1985 physics
2.	16 September 1985	
3.	16 October 1985	
4.	16 November 1985	
5.	15 December 1985	
6.	19 January 1986	
7.	16 February 1986	
8.	16 March 1986	
9.	15 June 1986	19-level model
10.	14 September 1986	Gravity wave drag parametrization
11.	14 December 1986	
12.	15 March 1987	
13.	14 June 1987	
14.	13 September 1987	
15.	13 December 1987	
16.	13 March 1988	New vertical diffusion scheme

Table 1

Initial dates of a set of 16 forecast ensembles run with T63 ECMWF operational spectral model. Changes in the model are indicated on the right-hand-side.

3. THEORETICAL BACKGROUND

Before looking at results obtained from the integrations, it is useful to discuss briefly what improvement can be expected, on average, from the ensemble mean of a time-lagged forecast over a single deterministic forecast. An unweighted time-lagged forecast is a particular realization of an ensemble forecast, and the results of Leith (1974) and Seidman (1981) are appropriate in the perfect-model environment.

Practical experience (see for example Molteni et al., 1988, and Murphy, 1988) indicates that, in general, the improvement in the skill of an ensemble-mean forecast compared with that of a deterministic forecast is noticeably smaller than expected from the perfect model theory. In this section we consider the expected improvement that ensemble averaging can make in a non-perfect model environment. Having established basic notation in section 3.1, we discuss properties of a simple analytical model for error growth in which the ECMWF model errors are parametrized.

3.1 Basic mathematical relations

For the purpose of statistical assessment of an ensemble forecast, we first define a basic set of expressions. Let F_i be a forecast field produced by one member of the ensemble (i=1, ..., N). For any given field X (which could be for example the verifying analysis, climate, etc.), the mean square distance of X from the members of the ensemble can be written as

$$\frac{1}{N}\sum_{i=1}^{N}|F_i - X|^2 = |\bar{F} - X|^2 + \frac{1}{N}\sum_{i=1}^{N}|F_i - \bar{F}|^2 \qquad (1)$$

where $\bar{F} = \frac{1}{N}\sum_{i=1}^{N} F_i$ represents the average of the N forecast fields, that is the centroid of the ensemble, and vertical bars denote the modulus. Let A be the analysed field which verifies each F_i, let $\bar{E} = \bar{F} - A$ the error field of the ensemble mean, and let us assume from now on that all the fields are expressed in terms of anomalies, i.e. deviation from the climate. We can define the following variables describing the statistical properties of the ensemble:

$$f^2 = \frac{1}{N}\sum_{i=1}^{N}|F_i|^2 \qquad (2a)$$

$$e^2 = \frac{1}{N} \sum_{i=1}^{N} |F_i - A|^2 \qquad (2b)$$

$$\Delta^2 = \frac{1}{N} \sum_{i=1}^{N} |F_i - \bar{F}|^2 \qquad (2c)$$

$$\delta^2 = \frac{1}{(N-1)N} \sum_{i=1}^{N} \sum_{j=1}^{N} |F_i - F_j|^2 . \qquad (2d)$$

f^2 is the ensemble average of the spatial variance of individual members (or deterministic forecast anomalies), e^2 represents the mean squared error of individual members (again averaged over the ensemble), Δ^2 is the mean squared spread (or dispersion) from the ensemble mean, and δ^2 is the mean squared distance between all pairs of individual forecasts. Using (1) we obtain relationships between variables defined by (2a) - (2d). For X=0 it follows

$$f^2 = |\bar{F}|^2 + \Delta^2, \qquad (3)$$

whereas if we set X=A we obtain

$$e^2 = |\bar{E}|^2 + \Delta^2. \qquad (4)$$

Eq. (4) quantifies the average improvement of the ensemble-mean forecast over the individual members in terms of mean-square error, and Eq. (3) indicates that this improvement is obtained by removing part of the variance from the forecast fields. The practical usefulness of the ensemble-mean forecast depends on whether this removed variance is due only to unpredictable scales of motion; in an ideal situation, Δ^2 should be exactly equal to the variance of the unpredictable components. Finally, a relationship between the squared ensemble spread Δ^2 and the mean squared distance of all pairs δ^2 is obtained by substituting F_j in (1) and summing over all F_j forecasts:

$$\delta^2 = \frac{2N}{N-1} \Delta^2. \qquad (5)$$

Now, in the perfect model hypothesis one assumes that the growth of the mean distance among the members of the ensemble is equal to the average growth of the 'deterministic' error, and that the spread of the ensemble at the initial time is representative of the analysis error; then for every forecast time $e^2 \approx \delta^2$, and from (4) and (5) one deduces

$$|\bar{E}|^2 = (1 - \frac{N-1}{2N})\, e^2. \qquad (6)$$

If N is sufficiently large, one obtains the theoretical 'perfect model' limit for the skill of an ensemble forecast deduced by Leith (1974), that is, the mean-square-error of an ensemble forecast is half of the average mean-square-error of the individual members of the ensemble.

Finally, for comparison, note that the error variance of a 'climate' forecast is $|A|^2$, that is, the magnitude squared of the observed anomaly.

Let us now consider the anomaly correlation coefficient (ACC) as a measure of skill for any deterministic or ensemble forecast. For a single forecast in the ensemble the ACC can be expressed as

$$\rho_i = \frac{F_i \cdot A}{|F_i| \cdot |A|} = \frac{|F_i|^2 + |A|^2 - |F_i - A|^2}{2\,|F_i| \cdot |A|}, \qquad (7a)$$

and for the ensemble mean as

$$\rho(\bar{F}) = \frac{\bar{F} \cdot A}{|\bar{F}| \cdot |A|} = \frac{|\bar{F}|^2 + |A|^2 - |\bar{E}|^2}{2\,|\bar{F}| \cdot |A|}. \qquad (7b)$$

In order to derive a relationship between $\rho(\bar{F})$ and the mean ACC of individual members of the ensemble, the latter being simply defined by

$$\bar{\rho} = \frac{1}{N} \sum_{i=1}^{N} \rho_i, \qquad (8)$$

we must make some assumptions about the ensemble. Thus, if we assume that each deterministic forecast F_i has a greater spatial variance than the ensemble mean ($|F_i| > |\bar{F}|$) and covariance $F_i \cdot A > 0$ for each F_i one can demonstrate that from Eq. (7) $\rho(\bar{F}) > \bar{\rho}$.

Alternatively, if one assumes that all the forecast anomalies in the ensemble have nearly the same amplitude, i.e. $|F_i| \simeq f$ for each i, then, by applying (4) and (3) in Eq. (8), it can be deduced that

$$\rho(\bar{F}) \simeq \bar{\rho} \frac{f}{|\bar{F}|} = \bar{\rho} \left(\frac{f^2}{f^2 - \Delta^2}\right)^{1/2}. \qquad (9)$$

(Note that if the average ACC of individual forecasts is negative, ensemble averaging will make poor scores even worse).

Since the ratio $f/|\bar{F}|$ is always greater than 1 and increases with forecast time due to the growth of the spread, the ensemble mean forecast should have a proportional increase in the ACC over a deterministic forecast (providing that $\bar{\rho}$ is positive!).

Given a theoretical model for the growth of e^2 and δ^2 (see section 3.2) and assuming a climatological constant value for f and for $a = |A|$, theoretical curves for $\bar{\rho}$ and $\rho(\bar{F})$ can be computed from

$$\bar{\rho} = \frac{f^2 + a^2 - e^2}{2fa} \qquad (10a)$$

$$\rho(\bar{F}) = \bar{\rho} \frac{1}{\sqrt{1 - \frac{\Delta^2}{f^2}}}. \qquad (10b)$$

3.2 Impact of the ensemble-mean forecast in an imperfect model

Dalcher and Kalnay (1987), based on earlier work of Leith (1978), have shown that the growth of error and spread of deterministic numerical forecasts can be parametrized by the

following equations:

$$\dot{e}^2 = (\alpha e^2 + \sigma V_e^*)(1 - e^2/V_e^*) \qquad (11a)$$

$$\dot{\delta}^2 = \alpha\delta^2 (1 - \delta^2/V_\delta^*), \qquad (11b)$$

where the dot represents the derivative with respect to forecast time, V_e^* and V_δ^* are the asymptotic (saturation) values of the variance of the deterministic error and spread respectively, α an 'internal' (i.e. intrinsic to the real atmosphere) growth rate of the error and σ an additional 'external' growth rate representing the effects (either random or systematic) of the model approximations.

It is clear from Dalcher and Kalnay and from Lorenz's (1982) work that the term σV_e^* is essential for a good parametrization of error growth, especially at short forecast times. Furthermore, it is common experience in NWP that the spread between forecasts grows slower than the actual error. Consequently, the theoretical, perfect model, limit for the error of an ensemble forecast, i.e. $|\bar{E}|^2 = 0.5\ e^2$, can only be achieved at very long forecast times. Fig. 2a, b shows the theoretical time evolution of e^2, $|\bar{E}|^2$, $\bar{\rho}$ and $\rho(\bar{F})$ computed using Eqs. (11a,b) and setting the values of parameters as follows

$a = f = 1.$

$V_e^* = V_\delta^* = 2.$

$\alpha = 0.4\ \text{day}^{-1}$

$\sigma = 0.025\ \text{day}^{-1}.$

The first four parameters are normalized by setting the observed variance to 1; the growth rates α and σ are chosen to fit the error growth of the (improved) T63 model used in this study and differ from those found in Dalcher and Kalnay. The initial values for e^2 and δ^2 were set to 0.025, which corresponds to the average initial squared distance between members of our time-lagged ensembles.

The thin solid line (A) shows the skill of an individual forecast with the above parametrization. In the mean square error it increases asymptotically to the value V_e=2. The thick solid line (B) represents the skill of the ensemble mean in a perfect model environment (ie assuming σ=0). It also increases monotonically, but from equations (4),

(5) and (11) asymptotes to a value $V_e - \frac{1}{2} V_\delta = 1$. The dashed curve (C) shows the skill of the ensemble mean for the imperfect model where σ is nonzero. During the first half of the forecast period, the reduction in the mean square error is much smaller than with the perfect model (curve B). However, it must asymptote to the same value as the perfect model, i.e. $V_e - \frac{1}{2} V_\delta = 1$. The ensemble-mean score therefore appears to show some apparent 'return of skill' in the extended range.

The potential improvement obtained by ensemble averaging is even smaller when one takes into account that in practice $V_\delta < V_e$ by an amount which is proportional to the variance explained by the systematic error itself. From the results to be presented in the following sections, a typical winter value of V_δ/V_e is about 0.75, decreasing to about 0.5 for an exceptionally poor ensemble forecast. The dotted (D) and dot-dashed (E) lines in Fig 2 show the ensemble-mean skill that would be achieved with our parametrization using these ratios, respectively.

Estimates of the anomaly correlation coefficient using equations (10) for these values of the parameters, are shown in Fig 2b. The reduced impact of the ensemble average in a non-perfect model environment can be seen quite clearly in these calculations.

The curves in Fig 2a,b have been computed assuming that the asymptotic value of the forecast error is twice the climatological variance. In fact, as results below indicate, see Fig 5, about 10% of the variance remains predictable at longer range (possibly due to predictability associated with the persistence of initial sea surface temperature anomalies) and therefore reduces the values of V_e and V_δ accordingly. In this case, one obtains the curves shown in Fig. 2c,d. One can see that the most significant effect of this assumption on ensemble forecast skill scores is to enhance the improvement in terms of anomaly correlation coefficient. A 'return of skill' can now be seen in the anomaly correlation coefficient in the case $V_e = V_\delta$ (curve C).

More generally, V_e and V_δ are functions of the annual cycle, and, within a season, of weather regime. Their dependence on the annual cycle is so strong, that even for models with significant systematic bias, asymptotic error and asymptotic spread are well correlated over an annual cycle. Within a season, on the other hand, the dependence of V_e and V_δ on weather regime could be quite different. For example, as discussed above V_δ reflects the intrinsic instabilities within that weather regime, whereas V_e may, additionally, reflect the impact of systematic deficiencies in the NWP model physics and numerics on forecast quality, given the flow pertaining to that weather regime. Hence, within a season, V_δ and V_e could be quite uncorrelated for models with serious systematic error. In such circumstances, ensemble spread and ensemble-mean skill would be poorly correlated.

In conclusion, the ensemble average is essentially a dynamically tuned spatial filter, which has the effect of removing unpredictable scales. A predictable component of the circulation can be made evident by the time-lagged forecasting technique if it already exists in the deterministic forecasts, but obviously it cannot be created! However, one must always remember that the improvement in the scores of the mean forecast is not the only (and probably not even the main) purpose of ensemble forecasting; its usefulness should be judged from its ability to provide a realistic probability distribution for expected atmospheric states.

4. A SUMMARY OF VERIFICATION RESULTS

In this section we shall present a summary of verification results from the time-lagged ensembles listed in Table 1.

4.1 Skill of ensemble-mean vs. mean skill of individual forecasts

We first address the question of whether the skill of the ensemble-mean forecast has increased over the skill of individual forecasts. Scatter diagrams of the 30-day mean northern hemisphere 500 mb height RMS error and anomaly correlation coefficient, respectively, for the ensemble-mean forecast against the mean skill of the individual forecasts are shown in Fig.3. As discussed in section 3 (equation (4)), the ensemble-mean RMS error is inevitably smaller than the mean RMS error of individual forecasts. This is clearly illustrated in Fig.3a, which shows a nearly linear relationship between individual-mean and ensemble-mean skill. In this, and following scatter diagrams, summer forecasts are shown with open circles, winter forecasts are shown with open boxes, and the transition forecasts are shown with crosses. The distribution of points in the scatter diagram clearly reflects the impact of the annual cycle on the ensemble mean RMS errors, with summer forecasts having smallest RMS errors, and winter forecasts having largest errors.

From Fig.3a, one can note that the reduction in error associated with ensemble averaging is, on average, somewhat larger for the transition season forecasts than for the summer forecasts. This is to be expected under perfect model assumptions since improvement due to ensemble averaging is proportional to the dispersion from the ensemble mean (see equations (5) and (6) for large N). Furthermore, one would expect the RMS amplitude of this dispersion to be strongly influenced by the annual cycle, being smallest in summer and largest in winter. However, the reduction of forecast error in the winter season is not as large as one might have anticipated on this basis. In addition, the smallest impact of ensemble averaging does not occur for a summer ensemble; it occurs for the winter ensemble from January 1986. For this ensemble, the RMS error was reduced from an individual forecasts mean value of 96 m to an ensemble mean value of 95 m. This suggests

that the perfect-model assumptions seriously break down for the winter ensemble forecasts. As will be shown in section 5, individual members of the ensemble from January 1986 failed to forecast the development of a major large-scale anomaly, and, moreover, were each consistent with one another.

As discussed in section 3 (equation (7)), provided the spatial variance of the individual forecast fields is larger than the spatial variance of the ensemble mean field, ensemble averaging will increase the absolute value of the anomaly correlation coefficient. This can be seen in Fig.3b. As in Fig.3a, there is an approximately linear relationship between individual-mean and ensemble-mean skill, though in this case the relationship can be thought of as a rotation of the diagonal about the origin. The winter ensemble lying below the origin is again the January 1986 case, where the mean individual score is negative, and, consistent with the discussion in section 3, ensemble averaging has made the anomaly correlation coefficient even more negative. Since the anomaly correlation coefficient is less strongly influenced by annual cycle effects than RMS error, the comparison of relative predictability in summer, transition and winter season forecasts is more meaningful in Fig.3b than in Fig.3a. In this sense, it would appear that summer ensembles have rather low skill in the prediction of phase compared with other times of year.

4.2 Time evolution of the ensemble-mean skill scores

We now briefly discuss the evolution of skill of the ensemble forecasts during the course of the integration. All ensembles are separated into the 'winter' - October to March (OM), and 'summer' - April to September (AS) periods.

In Fig 4, we show 5-day average scores of each individual ensemble-mean forecast. They show large variability. The poorness of the January 1986 forecast is clearly seen in both RMS error and anomaly correlation coefficient (thin dash-dot line). The ensemble-mean anomaly correlation for this case crosses the 0.6 line by day 5, and it continues to fall rapidly, reaching zero anomaly correlation by day 9. (Here, according to Hollingsworth et al. (1980), the 0.6 value has been taken as the threshold of 'usefulness' for medium-range forecasts.) The ensemble forecast for the next month, February 1986 (thick solid line), is the most skilful in terms of anomaly correlation coefficient, which does not fall below 0.6 until day 15 and stays relatively high at the end of the forecast.

During the AS period, it can be seen that anomaly correlation scores tend to decrease to zero faster than during the OM period. The RMS error is lower, but saturation is reached earlier than in winter.

The potential improvement that time averaging makes to the ensemble-mean forecast is shown in Fig.5. Since RMS is trivially reduced through time averaging, we show in Fig.5 only the impact on anomaly correlation coefficient. It can be seen that both daily and time-average scores fall below the 'useful' 0.6 line before day 10. If the predictability of the ensemble-mean forecast was the same for all time scales, the time averaging should merely result in a smoothing of the daily skill scores. In fact, as Fig.5 shows, there are some improvements for both 5 and 10 day mean forecasts compared with daily values, albeit rather modest ones. For example, during the OM period, daily scores drop below 0.3 around day 10. For 5-day mean fields, the same level of skill is achieved for days 11-15, and for 10-day mean fields, it is achieved for days 11-20. The improvement made by time-averaging is negligibly small at the end of the forecast period.

The fact that the largest improvement in the time-averaged fields is found between day 10-20 could be very significant. At the present, the upper limit for medium-range forecasting is normally considered to be about 10 days. The improvement in skill shown in Fig.5 gives rise to the hope that the upper limit for forecasts might extend beyond day 10 in future years, at least for time-averaged fields.

It can be seen in Figs 4 and 5, that there is some 'return of skill' for some individual ensemble-mean forecasts, and for the AS 10-day average ensemble-mean forecasts. This could be expected, as discussed in section 3, on theoretical grounds provided some of the variance remains predictable.

4.3 Skill of ensembles vs. skill of control forecast, climate and persistence

From a practical point of view, it is important to consider whether the ensemble-mean forecast is superior to the latest member of the ensemble, the control forecast. When compared against a single member of the ensemble, it is no longer inevitable that the ensemble-mean will be superior, even in terms of RMS error. Scatter diagrams of 500 mb height and 850 mb temperature RMS error and anomaly correlation coefficient of the ensemble-mean forecast against the control forecast are shown in Fig.6. In view of the impact of the ten-day average on ensemble-mean skill, discussed in the previous subsection, particularly during the middle ten days of the forecast period, here we concentrate on the skill of ten-day mean fields throughout the forecast period.

For days 1-10 (not shown), the ensemble-mean forecast is not a significant improvement over the control, and in the two of the winter cases (December 1985 and January 1986) it is noticeably worse. This by no means implies that ensemble forecasting does not offer significant opportunity for forecasting within the limit of deterministic predictability, but indicates that, in this forecast time range, time-lagged ensemble perturbations to the initial

state are not a priori equally likely. As recognised by Hoffman and Kalnay (1983) one can infer from this, that in order to show a consistent improvement using the time-lagged technique, the individual members of the ensemble would have to be weighted in order to minimise the impact of the earliest members.

For days 11-20 (Fig.6 left), the ensemble mean is more skilful than the control forecast for most cases. In terms of RMS error, only one of the ensemble-mean forecasts (December 1985) is worse than its control. In terms of anomaly correlation, the ensemble-mean forecast is almost always an improvement when it correlates positively with the verifying analysis. One can again note that, in cases where the ensemble-mean anomaly correlation is negative, it is lower than the anomaly correlation coefficient of the control forecast. For 850 mb temperature it can be seen that three winter forecasts (February 1986, December 1986 and December 1987) have an ensemble-mean anomaly correlation greater than 0.6, and only one (December 1985) is slightly negative. Also there are four transition season ensembles with anomaly correlation greater than 0.5. It would appear that, at this forecast time range, the unweighted ensemble-mean does show some benefit in predictive skill over the single deterministic forecast.

For days 21-30 (Fig.6 right), most of the ensemble-mean forecasts display some reduction in RMS error. However, this overall improvement is not seen in anomaly correlation scores where, in terms of 500 mb height, 6 of the ensemble-mean forecasts show some improvement, 4 show essentially no change, and 6 show the control forecast to be superior to the ensemble mean. One therefore must conclude that the apparent improvement in RMS error results in part from the lower asymptotic variance of the ensemble mean (compare in Fig.2 a,c curve A for individual forecast against other curves at considered time range).

We have also compared the skill of ensemble forecasts against 'zero-cost' forecasts provided by persistence (of ten-day mean anomaly fields) and climate. On average, the RMS error of a climate forecast asymptotes to $1/\sqrt{2}$ of the RMS error of a persistence forecast (since equation (6) holds for e=persistence error, E=climate error and for a very large N). Therefore, the climate RMS error is a more severe test of the skill of a numerical forecast in the extended range. Results are summarised in Table 2 for days 1-10, 6-15, 11-20 and 21-30. The first column for each set of ten day mean forecasts indicates whether the ensemble mean forecast was more skilful than the control forecast in terms of both RMS error and anomaly correlation coefficient. A '+' indicates it was more skilful, a '-' that it was not. In a similar way, the second column indicates whether the ensemble-mean forecast was more skilful than both persistence (RMS and anomaly correlation coefficient) and a climate forecast appropriate to the relevant forecast period (given by the amplitude of the observed anomaly; see section 3). We apply these tests to both 500 mb height (Table 2a) and 850 mb temperature (Table 2b).

Table 2

Z500	1 - 10		6 - 15		11 - 20		21 - 30	
Date	E/Ctr	E/ClP	E/Ctr	E/ClP	E/Ctr	E/ClP	E/Ctr	E/ClP
850516	+	+	+	-	+	-	+	-
850916	+	+	+	+	-	-	+	+
851016	-	+	+	-	+	+	-	-
851116	-	+	+	+	+	+	-	-
851215	-	+	-	-	-	-	+	-
860119	-	-	-	-	-	-	-	-
860216	-	+	+	+	+	+	+	+
860316	-	+	+	+	+	+	-	-
860615	-	+	+	-	-	-	-	-
860914	-	+	+	+	+	+	+	-
861214	-	+	-	+	+	-	+	-
870315	+	+	+	-	+	-	-	-
870614	-	+	+	+	+	-	-	-
870913	-	+	-	+	-	+	-	-
871213	-	+	-	+	+	-	+	-
880313	+	+	+	-	+	-	+	-

Table 2a Northern hemisphere 500 mb height skill of ensemble-mean forecast against skill of control forecast, persistence and climate.

T850	1 - 10		6 - 15		11 - 20		21 - 30	
Date	E/Ctr	E/ClP	E/Ctr	E/ClP	E/Ctr	E/ClP	E/Ctr	E/ClP
850516	+	+	+	+	+	-	+	-
850916	-	+	+	+	+	-	+	-
851016	-	+	+	+	+	+	-	-
851116	+	+	+	+	+	+	-	-
851215	-	+	-	-	-	-	+	-
860119	-	+	-	-	+	-	-	-
860216	-	+	-	+	+	+	+	-
860316	-	+	-	+	+	-	+	-
860615	-	+	-	-	-	-	-	-
860914	-	+	+	+	+	+	+	-
861214	-	+	-	+	+	+	+	+
870315	+	+	+	+	+	-	+	-
870614	-	+	+	-	+	-	-	-
870913	-	+	-	+	+	+	+	-
871213	-	+	-	+	+	+	+	-
880313	+	+	+	-	+	-	+	-

Table 2b As Table 2a but for 850 mb temperature.

For the first ten days, it can be seen that only one of the ensemble-mean forecasts of 500 mb height (January 1986) was less skilful than persistence and climate. Only 4 of the 16 ensemble-mean forecasts are improvements over their control. For days 6-15, 9 of the ensemble-mean forecasts of 500 mb height beat persistence and climate. However, of these, only 6 were also more skilful than their control.

As discussed in section 4.2, the most promising period for the ensemble-mean forecasts appears to be for days 11-20. For this ten-day mean, only one control forecast of 500 mb height beats persistence and climate (not shown), whilst there are 6 ensemble-mean forecasts of 500 mb height that beat persistence and climate. Five of these are also more skilful than their controls, and four of the five (November 1985, February 1986, March 1986 and September 1986) were also deemed skilful for days 6-15. For temperature at 850 mb we find seven ensembles more skilful than both their controls and the zero-cost forecasts.

For days 21-30, only two ensemble-mean forecasts of 500 mb height (September 1985 and February 1986) beat persistence and climate. Of these, only the February 1986 ensemble-mean forecast showed skill in the 1-10, 6-15 and 11-20 day period. This case, together with the exceptionally poor ensemble-mean forecast for January 1986 (having a minus sign in all columns!), will be discussed in more detail in section 5.

The results shown in Table 2 can be summarised as follows. For the first 10 days, the unweighted lagged-average forecast gives no significant advantage over single deterministic forecast from the latest initialisation date (control). Similarly, for the last ten days, the ensemble-mean forecast is not consistently more skilful than its control, except in the sense that its asymptotic variance is lower. The time range in which the unweighted lagged-average technique shows the main benefit is for days 6-15 and 11-20, where in most cases both RMS error and anomaly correlation coefficient scores are improved. In terms of anomaly correlation coefficients the forecasts of 850 mb temperature appear to be more skilful than 500 mb height.

The above discussion can be put in the following context. It is clear, from the theory and from the results above, that model systematic errors seriously undermine the possible gains from ensemble forecasting. However, as the model improves the benefit from ensemble averaging, especially in the time range 11-20 days, will also increase. That is to say, more ensemble-mean forecasts will beat their controls, and, at the same time, will beat climate and persistence. At present, as the above results show, about 30% of ensembles achieve this goal. If in the following years this percentage increases above a threshold of 50%, we believe that the operational use of ensemble forecasting will be justified.

4.4 Synoptic evolution of some ensemble-mean forecasts

In order to gauge synoptically the skill of some of the ensemble-mean forecasts of 500 mb height, judged in the previous subsection to be skilful at days 11-20, we show in Figs.7-10 the forecast and observed anomalies of the 500 mb heights for days 1-10, 11-20 and 21-30 of the November 1985, February 1986 and March 1986 ensembles. In addition we show anomaly maps for the exceptionally poor forecast from January 1986. A brief interpretation of the maps is given below.

There was an intense anomalous ridge over the north Pacific in the verifying analysis for the first ten days of the November 1985 forecast (Fig.7; also discussed in Hollingsworth et al., 1987). This subsided in the second ten days, leaving a relatively zonally uniform band of positive height anomaly in high latitudes. In the third ten days strong ridging occurred over the Gulf of Alaska. The ensemble mean-forecast amplitudes were somewhat weak compared with the observed. However, the main synoptic features were captured reasonably well from the first to second ten days, but largely missed by the third ten days.

The development of a strong European blocking dipole during the second ten-day period of the January 1986 forecast can be clearly seen in Fig.8. This continued into the third ten-day period where ridging over the Gulf of Alaska also developed. The ensemble-mean forecast for the second and third ten days is a complete failure - the forecast flow developed towards a strong zonal circulation.

During the first ten days of the February 1986 forecast (Fig 9), there was a strong positive height anomaly, centred over Greenland. This had disappeared by the second ten days, and a strong negative height anomaly was positioned in polar latitudes. In the ensemble-mean forecast, the polar latitude negative anomaly developed in the second ten days, and the Greenland anomaly weakened, though not as comprehensively as in the observed field. In the third ten days a negative anomaly developed to the south of Greenland with a positive anomaly over northern Europe. The forecast for days 21-30 captured the development of the north Atlantic anomaly, but not the European one.

During the first ten days of the March 1986 ensemble (Fig.10), the verifying analysis showed an intense dipole anomaly across the Atlantic, and a moderate anomalous ridge in the mid-Pacific. By the second ten days, both features had weakened, with the Atlantic positive height anomaly moving westward over the US eastern seaboard. This development was captured reasonably well by the ensemble-mean forecast. The development of ridging over the north Pacific in the final ten days was only approximately captured by the forecast.

4.5 Relationship between spread and skill

Our discussion of objective and subjective measures of skill of the ensemble-mean forecasts indicated that of all ensembles about one third appeared to give useful guidance for the ten-day mean period 11-20. This is broadly consistent with results from the deterministic T106 forecasts discussed in I. Without some a priori indication of reliability, this is an unacceptably small number to be of operational use. By studying the dispersion of the ensemble, we might be able to obtain such an a priori indication.

With perfect model assumptions, combining Eqs (5) and (6) for large N, the spread within an ensemble is correlated with the skill of the ensemble mean. We assess whether this correlation holds in the present realistic conditions of an imperfect model and an imperfectly sampled ensemble.

Scatter diagrams of NH 500 mb height and 850 mb temperature ensemble RMS standard deviation against ensemble-mean RMS error are shown in Fig. 11. For the first ten-day mean (not shown), there is very little correlation between these two quantities. By days 11-20 (Fig.11 left) some correlation exists for the transition season forecasts of 500 mb height, but none for the winter forecasts. For days 21-30, there is a correlation between spread and skill taking forecasts of 500 mb height from all seasons into account, but none for forecasts within a season. The correlation that exists in the full sample clearly reflects the impact of the annual cycle on spread and skill of 500 mb heights. Such an effect cannot be seen for 850 mb temperature, indicating the relatively small impact of the annual cycle on hemispheric-scale low frequency variability of this parameter (see, for example, Lau et al., 1981).

Within the winter season in particular there is no apparent correlation between RMS spread and skill at any time within the forecast period. Hence, we conclude that in our imperfect model environment, hemispheric-scale spread is not a good predictor of hemispheric scale skill. The impact of the model systematic error can be seen from Fig.11 as an offset of markers above the diagonal (for hemispheric 500 mb height this is on average about 20-30 m). The asymptotic ensemble-mean RMS error is typically about 1.5 times larger than the asymptotic ensemble-mean spread (and over twice the ensemble-mean spread for the January case). From equations (4) and (5) this gives values of the ratio V_e/V_δ used in the theoretical calculations in section 3.2. For 850 mb temperature ensemble-mean error, this offset is larger in the last ten days than in the days 11-20 and is explained by a much slower approach of RMS error to its saturation level. In fact, in many cases saturation is not reached even by the end of the integration period (not shown).

It has been argued that a spread/skill correlation may be more apparent when the data is calculated over a smaller region than the whole hemisphere. Fig.11 e,f show the scatter points for 500 mb height forecasts over the Atlantic sector (22.5-85.0°N, 90.0°W-27.5°E). It can be seen that there is no overall improvement in the level of correlation when compared with hemispheric results.

Let us return to the practical question raised at the beginning of this section. We identified three of the ensemble-mean forecasts (November 1985, February 1986 and March 1986) that had apparent skill in the ten day mean 11-20, and one that had no skill (January 1986). Can we use the spread of the ensemble to give an a priori indication of their apparent skill? The scatter points associated with these three forecasts are indicated on Fig.11. It can be seen that the skilful forecasts are not distinguished by particularly small spread; indeed the spread for the November 1985 ensemble is the largest of this set. Also, spread for the (poor) January 1986 ensemble is not very much different from the others.

Thus, we must conclude that this spread diagnostic is an uncertain predictor of ensemble-mean skill. However, we shall return to this topic in section 6, when we discuss a cluster analysis of these ensembles.

5. CASE STUDIES

From the skill assessment above, two forecasts are distinguished by their skill, or lack of skill. Already within the first ten days the January 1986 ensemble-mean forecast was poorer than a 'zero-cost' forecast. Conversely, the February 1986 forecast was skilful throughout the forecast period. In an attempt to shed more light on characteristics of successful and unsuccessful ensembles, we investigate in more detail these two wintertime forecasts. As already mentioned, this period is also of considerable interest as it covered the development of a strong European block, which has been the subject of intensive diagnostic study (Hoskins and Sardeshmukh, 1986; Sardeshmukh, 1988).

A synoptic evolution of the ensemble-mean forecast and verifying analysis anomalies for these two cases has been made in the previous section. To summarise, during the period covered by the January 1986 ensemble, the atmosphere underwent a transition to strongly meridional flow, and the ensemble-mean forecast largely failed to predict this transition. By contrast, at the beginning of the period covered by the February 1986 ensemble, the atmosphere was already in a 'meridional' regime, and predictions were quite skilful.

In section 5.1, we display some reduced phase space trajectories of each ensemble, calculated from the principal empirical orthogonal functions (EOFs) for each of the two forecast ensembles. These trajectories show clearly the evolution of the ensemble

dispersion for the two case studies, and highlight the failure of the January 1986 forecast ensemble in the medium range.

In section 5.2, we discuss reasons for the failure of the January 1986 ensemble, and show results from two further experiments, in an attempt to find causes of the failure. Firstly, we have rerun one member of the ensemble with a more recent version of the ECMWF model (including, for example, a parametrization of orographic gravity wave drag). Secondly, motivated by Hoskins and Sardeshmukh's diagnosis of the January 1986 block, we have run a forecast in which model variables in the tropics are relaxed towards their analysed values.

5.1 Phase space trajectories of forecast ensembles

As discussed in section 4.5, the 10-day mean RMS spread of the winter ensembles was not well correlated with their skill. In this section we study the evolution of the ensemble dispersion in more detail. Our initial attempts were through maps of standard deviation of forecast 500 mb height. As expected, these showed maxima in regions of large low-frequency variability (e.g. over Europe), but the local magnitudes of standard deviation did not clearly distinguish between the skilful and unskilful ensembles. We therefore decided to study the dispersal of only the large-scale features of the ensemble forecasts. A convenient way to study this is through the evolution of the leading EOF coefficients that explain the highest proportion of total forecast variability within the ensemble. The time-evolution of an ensemble can be illustrated as trajectories on a low-dimensional phase space cross-section. In our case, for each ensemble, the axes of the phase space trajectories are coefficients of the first two EOFs of the 5-day mean forecast 500 mb height fields (from days 1-5 to days 26-30 inclusive). These cross sections can be thought as optimal in the sense that, of all two-dimensional cuts through the phase space of the ensemble-mean forecast, they capture the most variability of 5-day mean 500 mb height within each forecast ensemble. These two EOFs are shown in Fig.12a for the January ensemble, and in Fig.12b for the February ensemble. The percentage of explained variance associated with each EOF is given in each diagram. (The EOF decomposition, with less severe truncation, has also been used to define possible clustering behaviour of subsets of each forecast ensemble. Results from this clustering analysis is described in detail in section 6.)

Whilst the EOFs are hemispheric in extent, the first EOF of the January ensemble shows maximum (negative) amplitude over the north east Pacific. Indeed the pattern of geopotential height over the Pacific and North America projects significantly onto the Pacific/North American (PNA) mode (Wallace and Gutzler, 1981; Barnston and Livezey, 1987). The second January EOF appears to be composed of a number of north/south

oriented dipoles, one of which (with negative EOF coefficient) characterizes the European blocking dipole. The first February EOF has most amplitude in high latitudes, whilst the second February EOF also has projection onto the PNA mode, though in opposite sense to the first January EOF.

The trajectories themselves are illustrated in Fig.13 for the January ensemble, and in Fig.14 for the February ensemble. The trajectories of the individual forecasts are numbered by the arrow head. The forecast from the earliest analysis is numbered 1, the forecast from the latest (i.e. most recent, see Fig.1) analysis is numbered 9. The trajectory of the projection of the verifying analysis onto these EOFs is also shown in these diagrams (trajectory with open arrow head, numbered 0).

The trajectories for the January 1986 ensemble show that between the first and second 5-day mean each member was evolving in a consistent but erroneous manner. This error became greatly amplified between the second and third pentads (days 6-10 and days 11-15), when the direction of analysis and ensemble were quite contrary. Note that up to this point the spread of the ensemble in this low dimension phase space appears to be modest compared with the error of each forecast. After the third pentad, the forecasts do scatter quite significantly, and are apparently chaotic by the last pentad.

The relative scatter in the January ensemble can also be gauged by comparing the percentage of explained variance of the first two EOFs (shown in Fig.12). For January it is 36%, whilst for February it is 53%. Hence the large scatter observed in the second half of the January ensemble apparently occurred in a significantly larger dimensional space than shown in Fig.13.

Nevertheless, it is interesting to note that the members of the January ensemble began to disperse significantly when the trajectories migrated from the right hand part of the phase-space plane to the left hand part, i.e. from positive to negative values of the first EOF coefficient. From the discussion above we can say that significant ensemble spread occurred when the forecasts began to develop negative PNA index. This behaviour of the January 1986 ensemble trajectories is consistent with results from Palmer (1988), who found that atmospheric barotropic instability is enhanced for flows with negative PNA index.

Because of this eventual large scatter, not all of the individual forecasts from the January ensemble developed towards anomalies typical of the ensemble-mean forecast. From Fig.13 it is apparent that, in this reduced EOF phase space, forecast 2 is closest to the analysis towards the end of the forecast period. Indeed, its position at days 21-25 is very close to the position of the analysis at days 26-30. At days 26-30 it can be seen that only the

analysis and forecast 2 have a strongly negative EOF2 coefficient. Forecast 2 has the lowest 500 mb height RMS error in the last ten days and in the whole 30-day period, 122 m and 76 m respectively. This RMS error is about 20 m less than the mean RMS error from all individual forecasts in both periods.

The 10-day mean anomalies for forecast 2 from the January 1986 ensemble are shown in Fig.15 (for analysis anomalies refer to Fig.8). It can be seen that, like other members of the ensemble, the forecast fails to predict the Euro/Atlantic Block. However, unlike the ensemble mean (Fig.8), forecast 2 does develop positive height anomalies over Alaska and the north-east Pacific in days 11-20. These intensify in the last ten days, and the large-scale forecast anomaly field over the N.America bears some resemblence to reality. Forecast 2 is quite unlike the others in that it never develops into an anomaly pattern characteristic of the ensemble-mean (or indeed the model's systematic error; see below).

The accuracy of the two-dimensional phase space trajectories of the February 1986 ensemble (Fig.14) is impressive, showing that the evolution of the large scale flow even in individual forecasts is captured well right to the end of the forecast. By the end of the forecast period, the February ensemble has noticeably less scatter than at the end of the January forecast period. However, it can be seen that at the beginning of the forecast period, the scatter of the February forecast is not noticeably less than the January forecast.

It is interesting to note, in addition, that the spread in the February ensemble appears to increase most substantially between days 16-20 and days 21-25 (see Fig.14). During this period the first EOF coefficient was roughly constant, whilst the second was increasing from negative to positive values. Again, from the discussion above, this is consistent with the PNA index decreasing from positive to negative values.

The scatter of the phase space trajectories towards the end of the two 30-day forecasts does in fact give an indication of the relative skill of the two forecasts. However, this is not the case for the first half. It could be argued that the failure of the January 1986 ensemble reflected a fundamental inability of the model to simulate blocking activity (see, for example, Tibaldi and Molteni, 1988). On the other hand, it is possible that none of the initial perturbations introduced by the time-lagged forecasting technique was sufficient to cause at least some of the members of the ensemble to develop into blocking patterns. We investigate these issues further in the next subsection.

5.2 Further experimentation for the January 1986 case

Figs 8 and 9 suggest that to first order, the different skill of the January and February ensemble-mean forecasts can be accounted for by the fact that, broadly speaking, the

verifying analysis anomalies developed in the same sense as the model systematic error in February, and developed in the opposite (different) sense as the model systematic error in January. However, this explanation is too superficial. Why did all the January forecasts miss the development of the Euro/Atlantic block in the third pentad? In order to study this further, we have performed two additional experiments from 19 January 1986.

Since the time of integration of the January 1986 ensemble forecast, the operational model at ECMWF has three additional vertical levels, a parametrization of orographic gravity wave drag and revised vertical diffusion scheme in the free atmosphere (see Table 1). All of these modifications are known to have reduced the model's systematic error, and by alleviating the problem of systematic too strong westerlies improved the model climatology (Miller et al., 1989, Simmons et al., 1989). In addition, longer timescale integrations suggest that the gravity wave drag parametrization improves a GCM's ability to simulate low-frequency variability, particularly over the European region (Slingo and Pearson, 1987; Palmer, 1987).

In order to test whether these modifications could improve the skill of the January 1986 forecast, we have rerun the last member of the ensemble (12Z 19 January 1986) with the late 1988 version of the ECMWF model ('cycle 30'). The 10-day mean anomalies of this additional experiment for days 11-20 are shown in Fig 16. Despite the model changes, it can be seen that there is no essential improvement in the forecast skill. Development of the block is still missed. (The relatively skilful forecast 2 of the January 1986 ensemble was also rerun with model cycle 30, but yielded no substantial improvement over results shown in Fig 16).

The impact of the improvements in model formulation is felt principally in the extratropics. However, serious systematic deficiencies in the tropical flow associated with insufficient large scale divergent flow still persist. These deficiencies affect the model's ability to simulate low-frequency variability in the tropics. This may be significant, since Hoskins and Sardeshmukh (1987) have concluded in a diagnostic study of the Euro/Atlantic block during the winter 1985/6, that whilst changes in the momentum and heat fluxes due to synoptic weather systems were crucial, a catalyst for the block could have been provided by an unusual diabatic forcing in the S.American - Caribbean region, and in particular to the development of anomalously intense convection over the northern part of S.America associated with strong convergence over the Caribbean.

Fig.17 shows the 5-day mean 200 mb divergence over the Caribbean area for the analysis and the latest member of the forecast ensemble. The strong Caribbean convergence pattern begins to develop during days 6-10 of the forecast, and is well-established during

days 11-15 (see Fig.17, top). On the other hand, the latest individual forecast (Fig.17, bottom) has hardly forecast this area of convergence, and errors are large, both in amplitude and scale. The pattern of divergence field for the January 1986 ensemble-mean forecast is essentially similar to that shown in Fig.17, bottom. However, due to ensemble averaging, amplitude is weaker than in individual forecasts.

In order to test the importance of the loss of skill in the tropics on the prediction of the Euro/Atlantic block, a 20 day forecast experiment from 12Z on 19 January 1986 has been run, in which the model fields in the tropics were relaxed strongly towards their analysed values. This technique has already been used by Haseler (1982) and Klinker (1989) to study the impact of the tropics on medium-range extratropical forecast skill. For technical reasons, this integration had to be made at T42 resolution, and so a control experiment was also run at T42. Both integrations were run with cycle 30 of the ECMWF model. The basic technical details of the relaxation are as follows. The forecast prognostic equations contain the additional term $-c(X-X_a)$, where X is a prognostic variable (vorticity, divergence or temperature), X_a is the appropriate analysed value of X, and c is the relaxation coefficient. Equatorward of 15 degrees latitude, c took the value of 1/8hr . Between 15 and 25 degrees c varied smoothly between 1/8hr and 0. There was no relaxation poleward of 25 degrees. To achieve a continuous relaxation, the 6 hourly analysed fields were interpolated to values matching the timestep of the model. Full details of this and other relaxation experiments are described in Ferranti et al. (1989).

Fig.18 shows the 500 mb height fields for the control (left) and relaxed (top right) experiments for days 16-20, together with the verifying analysis (bottom right). Whilst the control experiment develops towards strong zonal flow over the Euro/Atlantic region, the relaxed forecast has correctly developed cut-off low over Europe (though its position is not perfect), and a resonably strong ridge over the north-east Atlantic. It should be noted that 500 mb height over the Pacific and N.America are also more realistic in the relaxed experiment than in the control experiment (notice in particular the intensity of the ridge off the west coast of N.America in the relaxed experiment compared with the control).

The improvements obtained with the relaxation experiment are even more clearly seen in the northern hemisphere skill scores (Fig.19). (Skill scores are calculated north of 30°, thus avoiding overlapping with the area of relaxation). Apart from the first 5-day mean, the RMS error in the relaxation experiment is reduced progressively in all 5-day mean periods towards the end of integration (20 days). The 20-day mean RMS error (far right lines in Fig.19) has been reduced by 13 m; this improvement is even more impressive in the

anomaly correlation coefficient which rises from 0.23 in the control experiment to 0.50 in the relaxation experiment (see Ferranti et al., 1989). The additional graph seen in Fig.19 (dashed) refers to another relaxation experiment in which model fields in the tropics were relaxed throughout the integration towards the initial data. Clearly, the worsening of the skill for this experiment indicates that correct prediction of tropical low-frequency variability is essential for extratropical extended-range forecasting.

These results certainly lend some support to Hoskins and Sardeshmukh's conclusions; an incorrect forecast of the tropical flow during this period led to a deterioration of forecast skill over Europe. However, one cannot say that the intensity of the block was well captured in the relaxed experiment. One must conclude that the failure to forecast accurately aspects of purely extratropical variability (for example associated with the transport of low potential vorticity northward to the region of blocking) was crucial in accounting for the poor skill of the January 1986 ensemble.

Phase-space trajectories for the January 1986 ensemble-mean forecast, the 'cycle 30 model' forecast (shown in Fig.16), and the tropical relaxation forecast are displayed in Fig.20 using the same two EOFs illustrated in Fig.12. Here, an open arrow (numbered 0) indicates verifying analysis, the January 1986 ensemble-mean is denoted by 1, the cycle 30 experiment by number 2 and the relaxation experiment by number 3. Fig.20a shows that in the medium-range (days 6-10), the cycle 30 integration has had more impact, relative to the ensemble mean than the tropical relaxation experiment, the latter following approximately the same trajectory as the ensemble-mean forecast. Despite the fact that the direction of the cycle 30 experiment trajectory is incorrect, it follows a path not traced out by any of the ensemble forecasts. In other words, it would appear that the dispersion of trajectories associated with a perturbation in model formulation is not simulated by the time-lagged perturbations (of initial conditions). This indicates that future methodologies for Monte Carlo forecasting should have the potential for both types of perturbation (model formulation and initial conditions).

Between days 6-10 and 11-15, the relaxation forecast trajectory completely reverses its direction, and between days 11-15 and 16-20 follows approximately the direction of the analysis (Fig.20b). In this reduced phase space it can be clearly seen that at days 16-20 the relaxed experiment is closest to the analysis. This is further confirmed by the forecast trajectories projected onto the planes of EOF1 and EOF3 (Fig.20c) and EOF2 and EOF3 (Fig.20d). In both of these planes, the relative closeness of the relaxation experiment trajectory to the analysis trajectory can be seen.

6. CLUSTER ANALYSIS OF ENSEMBLE FORECASTS

As discussed in the introduction, it is necessary to establish procedures for condensing the enormous information content of an ensemble of forecasts. From a practical point of view, one potentially important advantage of the ensemble technique is the ability to give probabilities of possible alternative developments. If an extreme event is predicted in just one member of the ensemble, it would be associated with a small probability. Nevertheless, this may be valuable information for a user. Information about possible extreme events would not be available from the ensemble-mean forecast.

In formulating a probabilistic approach to the forecast ensemble, one can calculate probabilities that a forecast variable falls within different predefined categories. This technique has been explored using the ECMWF time-lagged ensembles by Brankovic et al. (1988). In this paper we propose an alternative strategy where clusters of ensemble forecasts are objectively identified, and assign probabilities to each cluster according to the density of population.

Clearly, the skill of these probability forecasts is only as good as the forecasting system. If the model has serious systematic errors, or if the technique for constructing the ensemble is unrepresentative in some way, then the estimated probabilities could be quite misleading. Indeed, as discussed in the previous section, the model is not free of systematic errors, and there are still critical questions concerning the required size of an ensemble, and the representivity of perturbations generated by the time-lagged technique. Despite these caveats, the results described below indicate that cluster techniques can help to improve the potential of ensemble forecasting.

In section 5.1 we described phase space trajectories of the January 1986 ensemble by calculating EOFs of the forecast 500 mb height fields. We noted from a visual inspection of the trajectories of the ensemble in the two-dimensional plane spanned by the first two EOFs, that forecast 2 appeared to be somewhat unique, and quite different from the ensemble mean. In section 6.1, we shall describe the technique used to study objectively the tendency of forecasts to cluster, and discuss the application of the technique to the January 1986 and February 1986 cases. In section 6.2 we describe the results of the analysis applied to all the winter cases, together with the two skilful forecasts from November 1985 and March 1986.

6.1 Cluster analysis and the January 1986 ensemble

The technique we use is the application of a Ward hierarchical clustering algorithm to the EOF coefficients of each forecast ensemble (see Anderberg, 1973). For the calculations shown here we use as many EOFs as is necessary to explain about 80% of the variance of

500 mb height within the forecast ensemble. (For example, for the January 1986 ensemble 11 EOFs are used; for the February 1986 ensemble 8 EOFs are used.) Hence these calculations use many more degrees of freedom than are shown in the phase space trajectories in Figs 12 and 13. Calculations have been performed with different EOF truncations to confirm the approximate independence of results to truncation.

The cluster algorithm is applied separately to the EOF coefficients for 10-day mean periods, i.e. for days 1-10, 11-20 and 21-30. For each 10-day mean the algorithm works in nine separate steps. In the first step it finds those two members of the forecast ensemble which are the most similar in the sense that their RMS difference is minimized. These two forecasts are then merged to form a single sub-ensemble (cluster). In the next step, the algorithm merges either two more forecasts, or any of the remaining forecasts with the cluster obtained in the first step, with the objective of finding the combination in which the internal variance of the new cluster is minimized. Hence the nine steps of the algorithm have a tree-like structure; at each step the number of clusters is reduced by one. In the last step, all nine members of the ensemble are merged to form the ensemble mean. At an earlier step n, $1<n<9$, probabilities could be assigned to each cluster according to the number of individual members of the ensemble that have been merged into that ensemble. The tree diagrams for the three 10-day means of the January 1986 and February 1986 ensembles are shown in Fig.21.

Let us consider first the January 1986 ensemble. The tree diagram for the first ten day mean is the least interesting. At each step of the clustering algorithm, adjacent forecasts are merged; that is to say, the procedure merges forecasts which are initialised from adjacent initial analyses. This is what one would expect, at least in the short range, and demonstrates that the algorithm is operating sensibly. In the penultimate step, the two sub-ensembles are formed from forecasts (1,2,3,4) and (5,6,7,8,9) respectively.

For the second ten day mean, the clustering algorithm produces more interesting results. At the fourth step, two non-adjacent forecasts are merged, numbers 2 and 4. At step 7, the first forecast is merged with a sub-ensemble which includes the last member of the ensemble, forecast 9. In fact, at step 7 the first of the three clusters is composed of forecasts 2 and 4; the second consists only of forecast 7 and the third is composed of forecasts 1,3,5,6,8,9. In the penultimate step, the two sub-ensembles are formed from forecasts (2,4) and (1,3,5,6,7,8,9) respectively.

For the third 10-day mean, forecasts from non-adjacent analyses have been merged by the third step of the algorithm (where forecast 3 is merged with the mean of forecasts 5 and 6). This demonstrates quite clearly that in the extended range, there is no obvious way to

'weight' individual members of the ensemble as a means of reflecting the time-lagged technique used to construct the initial perturbations.

The uniqueness of forecast 2 is further highlighted in the tree diagram for the third ten-day period. By the seventh step, the first of the three clusters consists purely of forecast 2, the second consists of forecasts 4 and 7, the third is comprised of all the others. By the eight step, only forecast 2 continues to remain distinct from the others; the two sub-ensembles are formed from forecasts (2) and (1,3,4,5,6,7,8,9) respectively. This cluster analysis has confirmed our assessment in section 5.1 of the uniqueness of forecast 2. As discussed earlier, forecast 2 has the lowest 500 mb height RMS error of any of the ensemble forecasts at days 21-30.

As mentioned above, probabilities could be assigned to the clusters according their density of population. However, this may not be appropriate in a practical, rather than perfect model environment. If a densely populated cluster indicates an anomaly field strongly correlated with the known model systematic error, then its probability might be overestimated by the clustering technique. On the other hand, if a cluster develops away from the known model systematic error, then its probability is likely to be underestimated. Hence there is some scope for a statistical adjustment of cluster probabilities to take into account the synoptic development of the clusters. On this basis, one would in practice give considerably more weight to forecast 2 of the January 1986 ensemble than the ratio 1:8 suggested by the cluster analysis at step 8.

The tree diagrams for the February 1986 ensemble show that in the first ten days, with the exception of one step, only adjacent forecasts are merged. In the second ten days non-adjacent forecasts are merged by the third step. For the third ten days there is considerable merging of non-adjacent forecasts, and, by the penultimate step, the two ensembles are formed from forecasts (1,3,5,7) and (2,4,6,8,9) respectively.

6.2 Probability forecasts for the extended winter period

In Table 3 we give the anomaly correlation scores at days 11-20, and 21-30, respectively, of forecast northern hemisphere 500 mb heights for the centroids of the three clusters obtained at step 7 of the clustering analysis technique outlined above. We show the set of five winter ensemble forecasts, supplemented by the forecasts from November 1985, March 1986 and September 1986, deemed to be 'skilful' in section 4. The first three columns give the scores calculated for three forecast clusters; most populated cluster always being in column A, and the individual forecasts comprising each cluster are shown in brackets. We discuss here the stage in the cluster analysis algorithm when only three clusters are defined, because it has been found that, on average, with three clusters about 50% of the

Table 3

11-20	Cluster			Ensemble mean
Date	A	B	C	
851116	(2,3,4,6,9) .51	(5,7,8) .76	(1) .09	.66
851215	(1,2,3,4) -.16	(5,6,8) -.24	(7,9) .06	-.15
860119	(1,3,5,6,8,9) -.28	(2,4) -.43	(7) -.44	-.36
860216	(1,3,4,5) .58	(7,8,9) .60	(2,6) .62	.64
860316	(1,3,6,7,8) .46	(2,5) .55	(4,9) .27	.49
860914	(1,3,4,5,6) .29	(2,7,8) .58	(9) -.10	.44
861214	(2,4,5,8,9) .18	(1,6,7) .23	(3) .19	.22
871213	(3,6,7,8,9) .35	(2,4,5) .31	(1) .04	.32

Table 3a Days 11-20 northern hemisphere 500 mb height anomaly correlation coefficients for three clusters and ensemble-mean.

variance of the full ensemble is explained (100% being explained by nine 'clusters', i.e. all individual forecasts). In the fourth column, the score for the single cluster (ensemble mean) is shown.

For the means of both days 11-20 and days 21-30, in 7 out of 8 ensembles studied, at least one of the three clusters is superior to the ensemble mean. However, for days 11-20, in only 2 of the cases considered in Table 3, is the most skilful cluster also the most densely populated cluster. For days 21-30 in 5 cases the most skilful cluster was the most dense; but in three of these the density was equal to at least one other cluster (November 1985, December 1986 and December 1987). This suggests a problem of sampling because of the relatively small number of ensemble members.

To demonstrate the technique, anomaly and full maps of 500 mb height for days 11-20 for the three clusters are shown in Figs 22 and 23 for the ensemble forecasts from November 1985, and February 1986; the verifying fields are shown as two top panels.

21-30	C l u s t e r			Ensemble mean
Date	A	B	C	
851116	(5,6,7,9) .08	(2,3,4,8) -.14	(1) -.10	-.06
851215	(2,3,4,5,8) .35	(6,7,9) .36	(1) .17	.37
860119	(1,3,5,6,8,9) -.34	(4,7) -.15	(2) -.07	-.30
860216	(2,4,6,8,9) .44	(1,3) .20	(5,7) .25	.39
860316	(2,4,5,7,9) .36	(1,6,8) .28	(3) -.07	.32
860914	(1,4,5,9) -.15	(3,7,8) .19	(2,6) .32	.11
861214	(2,5,8) .30	(1,6,7) .14	(3,4,9) .28	.29
871213	(1,2,5) .34	(3,8,9) -.05	(4,6,7) .13	.15

Table 3b Same as Table 3a but for days 21-30.

For the November 1985 case (Fig.22), it can be seen that the height fields of the three clusters are dramatically different. The first cluster, (with highest density) has positive anomalies over high latitudes but fails to correctly develop the negative anomalies over the north Pacific and north Atlantic. The second cluster is clearly the most skilful (see Table 3a); the major anomaly centres are correctly predicted. The third cluster (with lowest density) develops intense ridging over Europe and the north Pacific. On the basis of the discussion for the clustering of the January 1986 ensemble in the previous section, it might be imagined that since the latter cluster represents a development away from the model systematic error, this development should be given higher weighting than suggested by the cluster density. In this case, however, it represents incorrect development.

For the February 1986 case (Fig.23), all clusters correctly show the extensive region of negative height anomaly over the high latitude Euro/Asian continent. They differ in other areas, however, for example in the strength of the Rockies ridge, and the trough over south-western Europe. Overall it can be seen, and is confirmed in Table 3a, that the third (least dense) cluster is the most skilful.

These results are certainly encouraging. However, they cannot be considered definitive. Firstly, problems associated with cluster density may well be associated with poor sampling. This suggests that the size of the ensemble should be much larger than the size studied here. This would probably require a different Monte Carlo technique to generate the ensembles. Secondly, the practical efficacy of the cluster analysis is certainly reduced by the presence of systematic model error. Nevertheless, we envisage continuing this type of analysis using future generations of ECMWF model.

7. CONCLUSIONS AND SUMMARY

A set of 16 ensembles of time-lagged extended-range forecasts have been run at different times of year using the T63 version of the ECMWF operational model. Each ensemble was composed of 9 integrations from consecutive 6 hourly analyses.

By definition, the RMS error of the ensemble mean forecast is inevitably smaller than the mean RMS error of the individual forecasts; the magnitude of the anomaly correlation coefficient of the ensemble mean forecast is similarly larger than the mean magnitude of the anomaly correlation coefficient of the individual forecasts, provided that the ensemble-mean forecast has smaller spatial variance than any of the individual forecasts.

Skill scores from the forecast ensembles confirm these results, though they show a smaller improvement then expected from perfect model theory, in particular during the winter period. However, they are consistent when the theory is extended to an imperfect model.

For the middle ten days (days 11-20), when all members of ensemble forecast can be considered as a priori equally likely and no weighting of individual forecasts is required, the ensemble mean is in most cases more skilful than the latest member of the ensemble (control forecast). However, at this time range only about a third of the ensembles were more skilful than both climate and persistence. The ensemble-mean forecasts of 850 mb temperature are more skilful than 500 mb height.

Under perfect model assumption, the ensemble spread and ensemble-mean skill are perfectly correlated for a large, properly sampled ensemble. For the integrations, there is an overall correlation between 10-day mean spread and 10-day mean skill; however, a substantial part of this correlation reflects the impact of the annual cycle on both spread and skill. Only within the transition season (i.e. neither winter, nor summer) does there still appear to be some positive correlation for 500 mb height at days 11-20 which does not bear any obvious relation to the annual cycle. In the winter period, however, there was no clear correlation between 10-day mean spread and 10-day mean skill. Therefore, the ensemble spread cannot be considered as reliable predictor for ensemble skill.

Within the winter period, there was considerable case-to-case variability in forecast skill. The January 1986 ensemble was the poorest of all the ensembles; the February 1986 ensemble was one of the most skilful. These two ensembles have been studied in more detail to try to find some reasons to explain this apparent lack of spread/skill correlation. This period has also been the subject of a diagnostic study of the large-scale flow, as it encompassed blocking events in both the Euro/Atlantic sector, and the north Pacific sector.

The January 1986 ensemble forecasts were initialised before these events developed. None of the forecasts correctly predicted the onset of the Euro/Atlantic block between days 10-15 of the forecast period, though at least one of the forecasts (but not the latest one) predicted the development of the north Pacific ridging. By contrast, the flow was already quite anomalous at the beginning of the February 1986 ensemble, and subsequent developments were well predicted, at least to day 20.

The different character of these two ensembles was clearly shown by considering phase space trajectories of the ensemble forecasts in the plane spanned by the two principal forecast EOFs of 500 mb height. During the first 15 days, the trajectories of the January 1986 ensemble forecasts were all very consistent with each other, but contrary to the observed atmospheric trajectory (such a behaviour would explain the lack of spread/skill correlation). During the second half of the forecast period, as the forecasts migrated from positive to negative PNA index, the trajectories dispersed quite strongly, becoming chaotic. By contrast, the trajectories for the February 1986 ensemble forecasts remained throughout most of the forecast period, both mutually consistent and in agreement with the real atmosphere's trajectory.

It can be said, therefore, that in terms of this large-scale measure of spread, the forecast dispersion at the end of the January 1986 forecast period gave an indication of the forecast skill. However, the dispersion in the first half did not. This is clearly related to the fact that all members of the ensemble consistently failed to predict the Euro/Atlantic block.

In order to investigate possible reasons for this, we conducted two additional experiments. In the first we reran the control integration with a later version of the ECMWF operational model (which has higher vertical resolution than the earlier model, a parametrization of orographic gravity wave drag, and a revision of the vertical diffusion scheme in the free atmosphere). The prediction of the Euro/Atlantic block was still missed, though the phase-space trajectory of this forecast was quite different to any of the members of the original ensemble.

In the second experiment, we ran an integration in which the model's tropical fields were relaxed towards the verifying analysis. This was motivated partly in the knowledge that the model's tropical divergence fields suffer serious systematic error (see II), and partly by the suggestion of Hoskins and Sardeshmukh (1987) that the development of anomalous upper convergence over the Caribbean region provided a catalyst for the block, and the fact that the large scale divergent flow in the tropics is poorly simulated in the model. Enhanced ridging over the Euro/Atlantic area certainly occurred in this experiment, and extratropical skill scores improved; however, it could not be said that the intensity of the block was well captured.

It would appear therefore that the failure of the model to predict the onset of the January 1986 block was partly associated with systematic deficiencies in the model and partly associated with a problem of predictability in the presence of imperfect sampling. We have seen that one of the members of the ensemble was able to capture some of the developments over the Pacific region in the extended range. A similar contradictory "success" has been reported by Hollingsworth et al. (1987) for November 1985 case.

By definition, extended range forecasting is concerned with prediction beyond the limit of deterministic predictability. In this sense it has an inherently probabilistic component. We used a clustering algorithm based on the EOFs of some ensemble forecasts to define sub-ensembles, and hence possible alternative developments of the large-scale flow. At any step of the cluster analysis algorithm, probabilities could be assigned according to the density of population of the cluster. For the cases studied one of the three clusters considered was invariably more skilful than the ensemble mean. However, this cluster was not always the more densely populated. It is possible that this is associated with a sampling problem, and that more realistic probability estimates could be obtained with a much larger sample. If this is the case, then the time-lagged method of generating ensembles may not be the most suitable. Systematic errors also play an important role in altering the probabilities of various circulation regimes.

Whilst these results do not suggest that extended-range ensemble forecasting is at present viable operationally, the rapid advances currently being made to reduce systematic error in NWP models, particularly in the tropics, together with new techniques for generating forecast ensembles that can identify a priori the most rapidly growing perturbations, suggest that probabilistic forecasting of extratropical time-mean weather using ensemble forecasting up to three weeks into the future is a feasible goal.

ACKNOWLEDGMENT

We wish to thank Laura Ferranti for allowing us to use some of her relaxation experiments. Also we thank A. Hollingsworth for constructive discussions and comments.

REFERENCES

Anderberg, M.R., 1973: Cluster analysis for applications. Academic Press, New York, 359 pp.

Barnston, A.B. and R.E.Livezey, 1987: Classification, seasonality and persistence of low-frequency atmospheric circulation patterns. Mon. Wea. Rev., 115, 1083-1126.

Baumhefner, D.P., 1988: Forecast skill and predictability at the extended range using Monte Carlo ensemble integrations from a general circulation model. In: Proceedings of the ECMWF Workshop on Predictability in the medium and extended range, 25-43. ECMWF, Reading, UK.

Brankovic, C., F.Molteni, T.N.Palmer, S.Tibaldi and U.Cubasch, 1988: Extended range ensemble forecasting at ECMWF. In: Proceedings of the ECMWF Workshop on Predictability in the medium and extended range, 45-87. ECMWF, Reading, UK.

Dalcher, A. and E.Kalnay, 1987: Error growth and predictability in operational ECMWF forecasts. Tellus, 39A, 474-491.

ECMWF, 1988: Reports of the working groups. In: Proceedings of the ECMWF Workshop on Predictability in the medium and extended range, 3-23. ECMWF, Reading, UK.

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

Ferranti, L., T.N.Palmer, F.Molteni and E.Klinker, 1989: Tropical - extratropical interaction associated with the 30-60 day oscillation, and its impact on medium and extended range predictability. (submitted to J. Atmos. Sci.)

Haseler, J., 1982: An investigation of the impact at middle and high latitudes of tropical forecast errors. ECMWF Tech. Rep., No.31, 40 pp.

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

Hollingsworth, A., 1980: An experiment in Monte Carlo forecasting. In: Proceedings of the ECMWF Workshop on Stochastic dynamic forecasting, 65-86. ECMWF, Reading, UK.

Hollingsworth, A., K.Arpe, M.Tiedtke, M.Capaldo and H.Savijarvi, 1980: The performance of a medium-range model in winter - impact of physical parameterizations. Mon. Wea. Rev., 108, 1736-1773.

Hollingsworth, A., U. Cubasch, S. Tibaldi, C. Brankovic, T.N. Palmer and L. Campbell, 1987: Mid-latitude atmospheric prediction on time scales of 10-30 days. In: Atmosheric and oceanic variability, ed. H. Cattle, Royal Meteorological Society Monograph, Bracknell, U.K., 117-151.

Hoskins, B.J. and P.D.Sardeshmukh, 1987: A diagnostic study of the dynamics of the northern hemisphere winter of 1985-86. Quart. J. R. Meteor. Soc., 113, 759-778.

Klinker, E., 1989: Investigation of systematic errors by relaxation experiments. (submitted to Quart. J. R. Meteor. Soc.)

Lau, N.C., G.H.White and R.L.Jenne, 1981: Circulation statistics for the extratropical northern hemisphere based on NMC analyses. NCAR Tech.Note TN-171+STR, 138 pp.

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

Leith, C.E., 1978: Objective methods for weather prediction. Ann. Rev. Fluid. Mech., 10, 107-128.

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

Miller, M.J., T.N.Palmer and R.Swinbank, 1989: Parametrization and influence of subgridscale orography in general circulation and numerical weather prediction models. Met. and Atmos. Phys., 84-109.

Molteni, F., U.Cubasch and S.Tibaldi, 1988: Monthly forecast experiments with the ECMWF spectral models. In: Persistent meteo-oceanographic anomalies and teleconnections. Edited by C.Chagas and G.Puppi, Pontificae Academiae Scientiarum Scripta Varia, 69, 505-555, Vatican City.

Murphy, J.M., 1988: The impact of ensemble forecasting on predictability. Quart. J. R. Meteor. Soc., 114, 463-494.

Palmer, T.N., 1987: Modelling low frequency variability of the atmosphere. In: 'Atmospheric and Oceanic Variability', Ed. H.Cattle, Royal Meteorological Society, Bracknell, 75-103.

Palmer, T.N., 1988: Medium and extended range predictability and stability of the Pacific North American mode. Quart. J. R. Meteor. Soc., 114, 691-714.

Palmer, T.N., G.J.Shutts and R.Swinbank, 1986: Alleviation of a systematic westerly bias in general circulation and numerical weather prediction models through an orographic gravity wave drag parametrization. Quart. J. R. Meteor. Soc., 112, 1001-1039.

Palmer, T.N. and S.Tibaldi, 1988: On the prediction of forecast skill. Mon. Wea. Rev., 2453-2480.

Palmer, T.N., C.Brankovic, F.Molteni and S.Tibaldi, 1989: Extende-range prediction with ECMWF models: I Interannual variability in operational model integrations. (submitted to Quart. J. R. Meteor. Soc.)

Sardesmukh, P.D., 1988: Another look at the European block of February 1986. In: Proceedings of the ECMWF Seminar on the nature and prediction of extra-tropical weather systems, Vol.2., 11-26. ECMWF, Reading, UK.

Seidman, A.N., 1981: Averaging techniques in long range weather forecasting. Mon. Wea. Rev., 109, 1367-1379.

Simmons, A.J., D.M.Burridge, M.Jarraud, C.Girard and W.Wergen, 1989: The ECMWF medium-range predicition models: Development of the numerical formulations and the impact of increased resolution. Met. and Atmos. Phys., 28-60.

Slingo, A. and D.W.Pearson, 1987: A comparison of the impact of an envelope orography and of a parametrization of orographic gravity-wave drag on model simulations. Quart. J. R. Meteor. Soc., 113, 847-870.

Tibaldi, S. and F.Molteni, 1988: On the operational predictability of blocking. In: Proceedings of the ECMWF Seminar on the nature and prediction of extra-tropical weather systems, Vol.2., 329-371. ECMWF, Reading, UK.

Tibaldi, S., T.N.Palmer, C.Brankovic, and U.Cubasch, 1989: Extende-range prediction with ECMWF models: II Influence of horizontal resolution on systematic error and forecast skill (submitted to Quart. J. R. Meteor. Soc.)

Tiedtke, M., W.A.Heckley and J.Slingo, 1988: Tropical forecasting at ECMWF: The influence of physical parametrization on the mean structure of forecasts and analyses. Quart. J. R. Meteor. Soc., 114, 639-664.

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

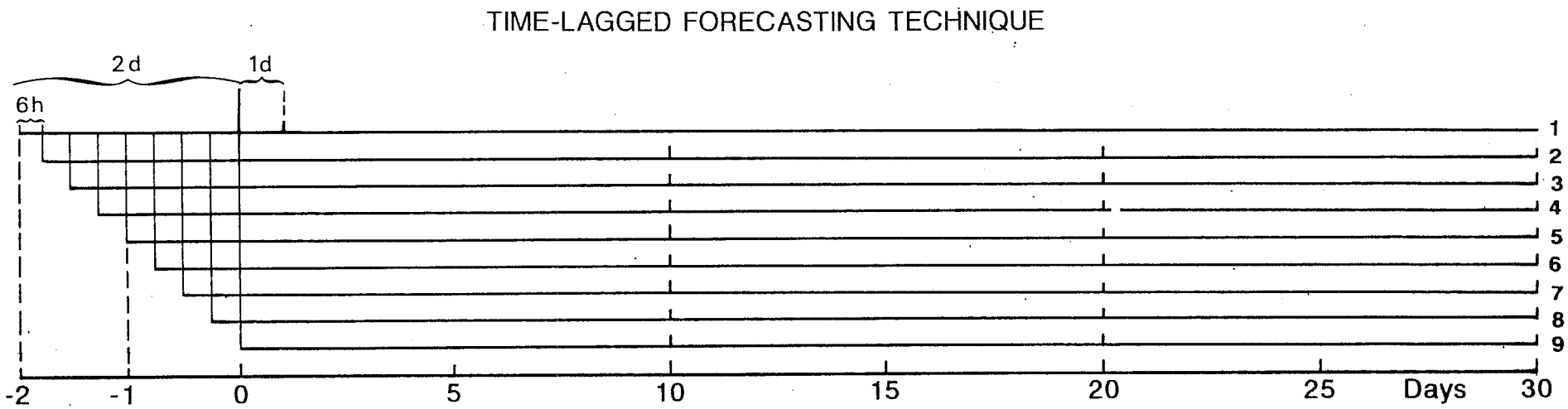

Fig.1 Schematic of the construction of a time-lagged forecasting ensemble with nine individual deterministic forecasts starting from adjacent initial analyses separated by 6 hours.

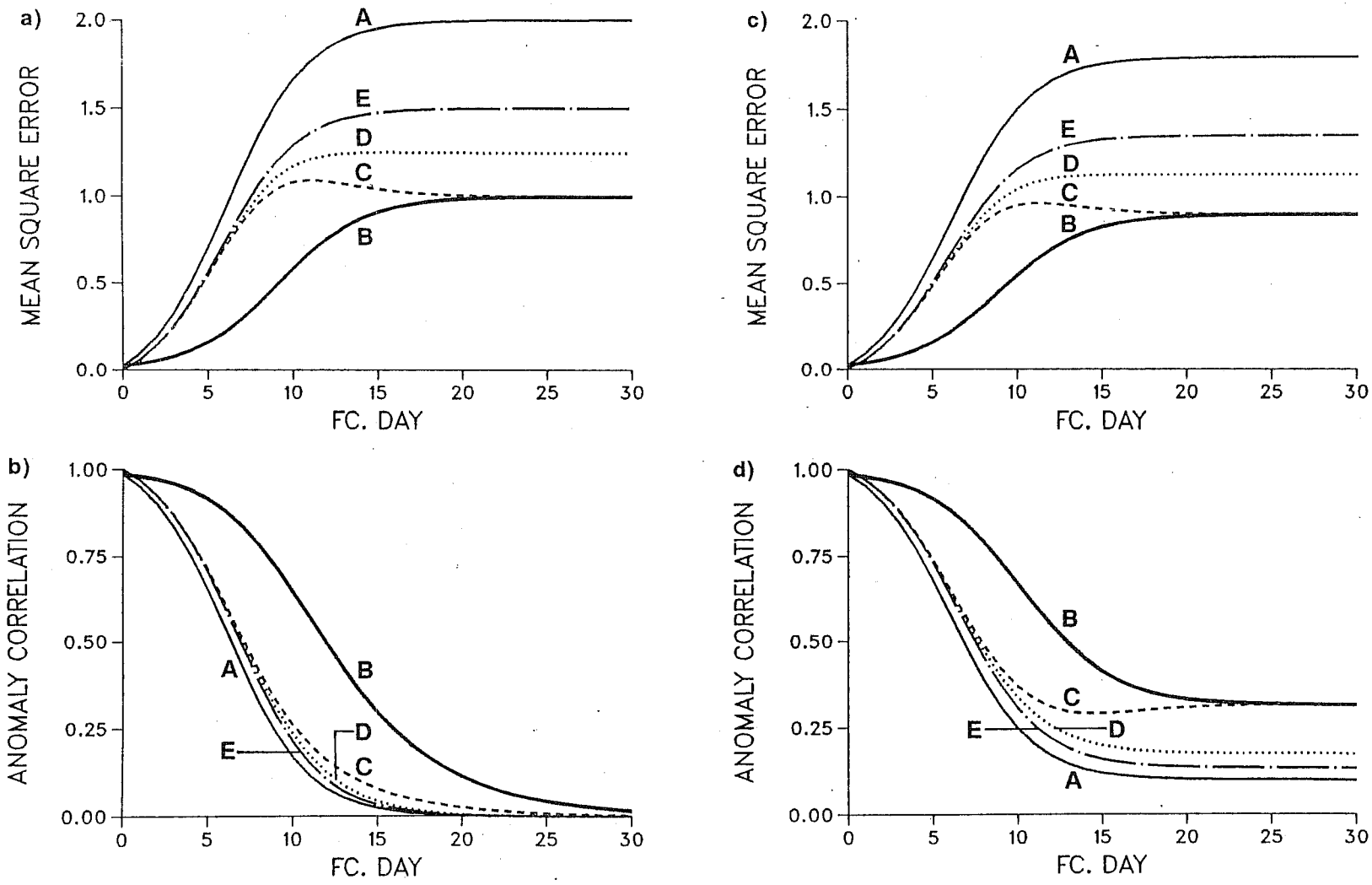

Fig.2 Theoretical error growth curves using an idealised parametrization of ECMWF model deficiencies. a) and b) RMS error and anomaly correlation respectively, c) and d) RMS error and anomaly correlation respectively assuming 10% of variance is predictable due to persistence of boundary forcing. Curve legend: A - skill of individual forecast, B - skill of ensemble-mean forecast under perfect model assumptions, C - skill of ensemble-mean forecast in model with external error growth, but no climate drift, D - skill of ensemble-mean forecast for the model with climate drift and V_δ/V_e=0.75, E - same as D but for V_δ/V_e=0.5.

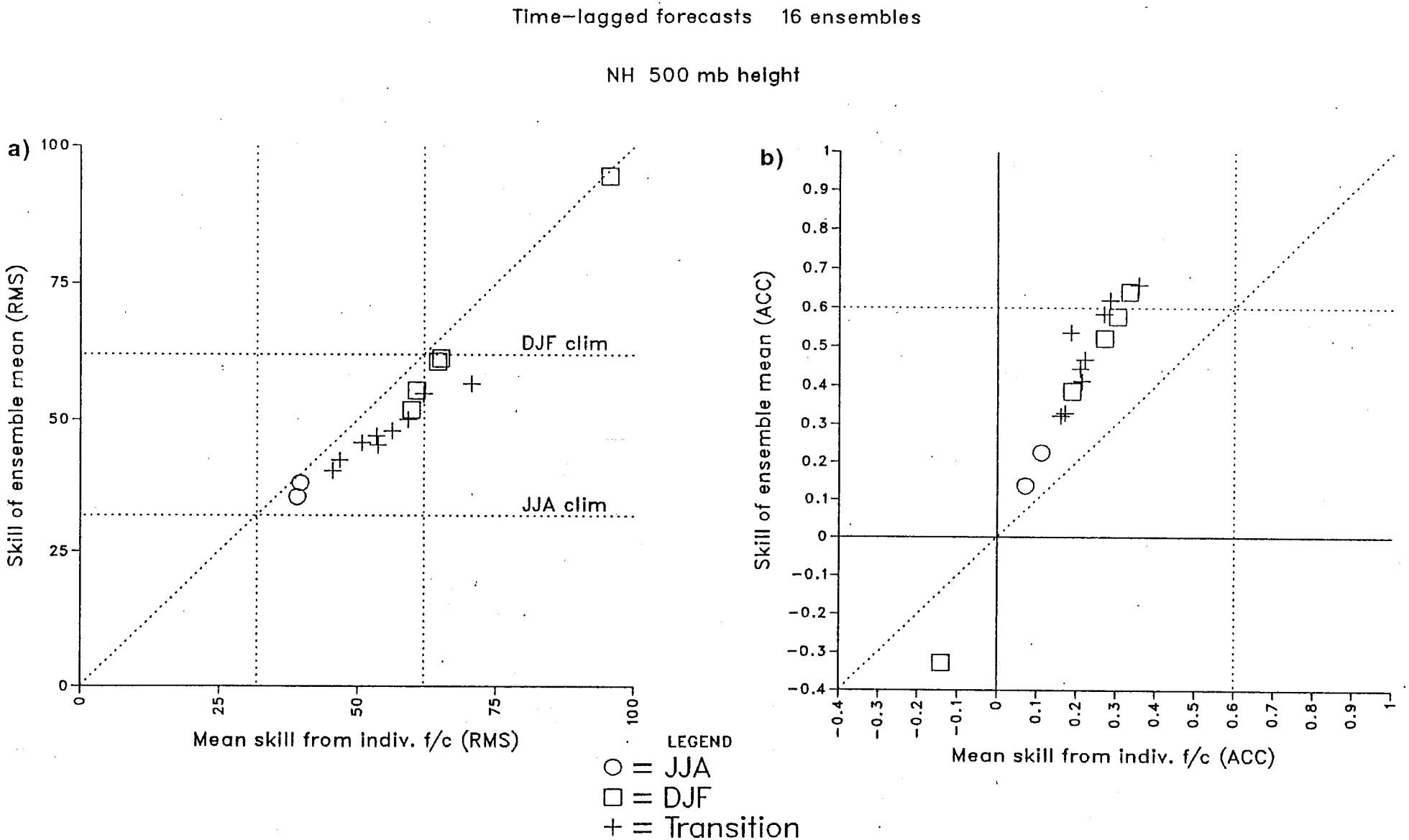

Fig.3 Scatter diagram of the northern hemisphere 30-day mean 500 mb heights: a) RMS error, and b) anomaly correlation coefficient for ensemble-mean forecast vs. mean of all individual forecasts. RMS error in metres.

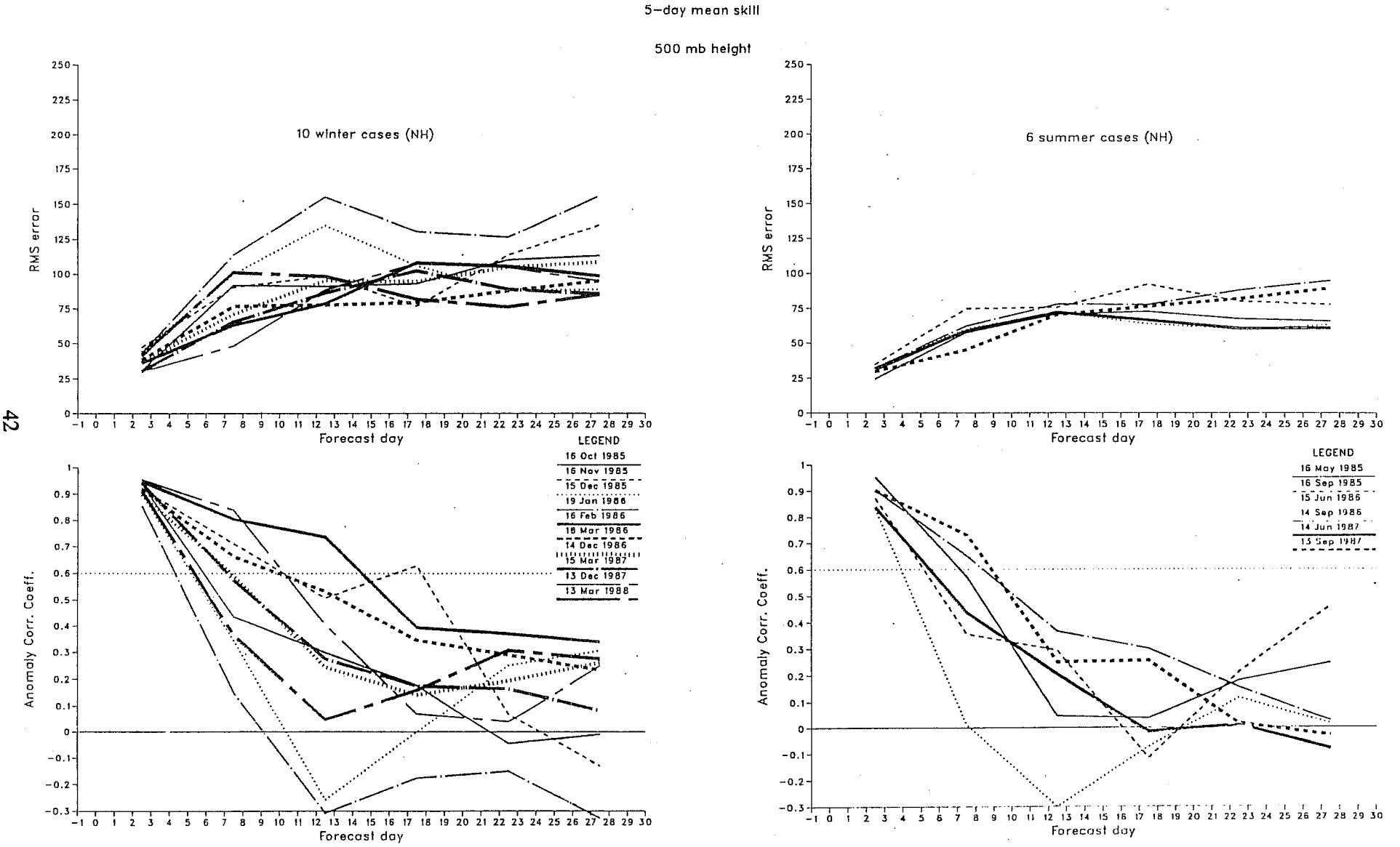

Fig.4 Time evolution of October to March (left) and April to September (right) ensemble-mean 5-day average skill for northern hemisphere 500 mb heights. Top: RMS error (m); bottom: anomaly correlation coefficient.

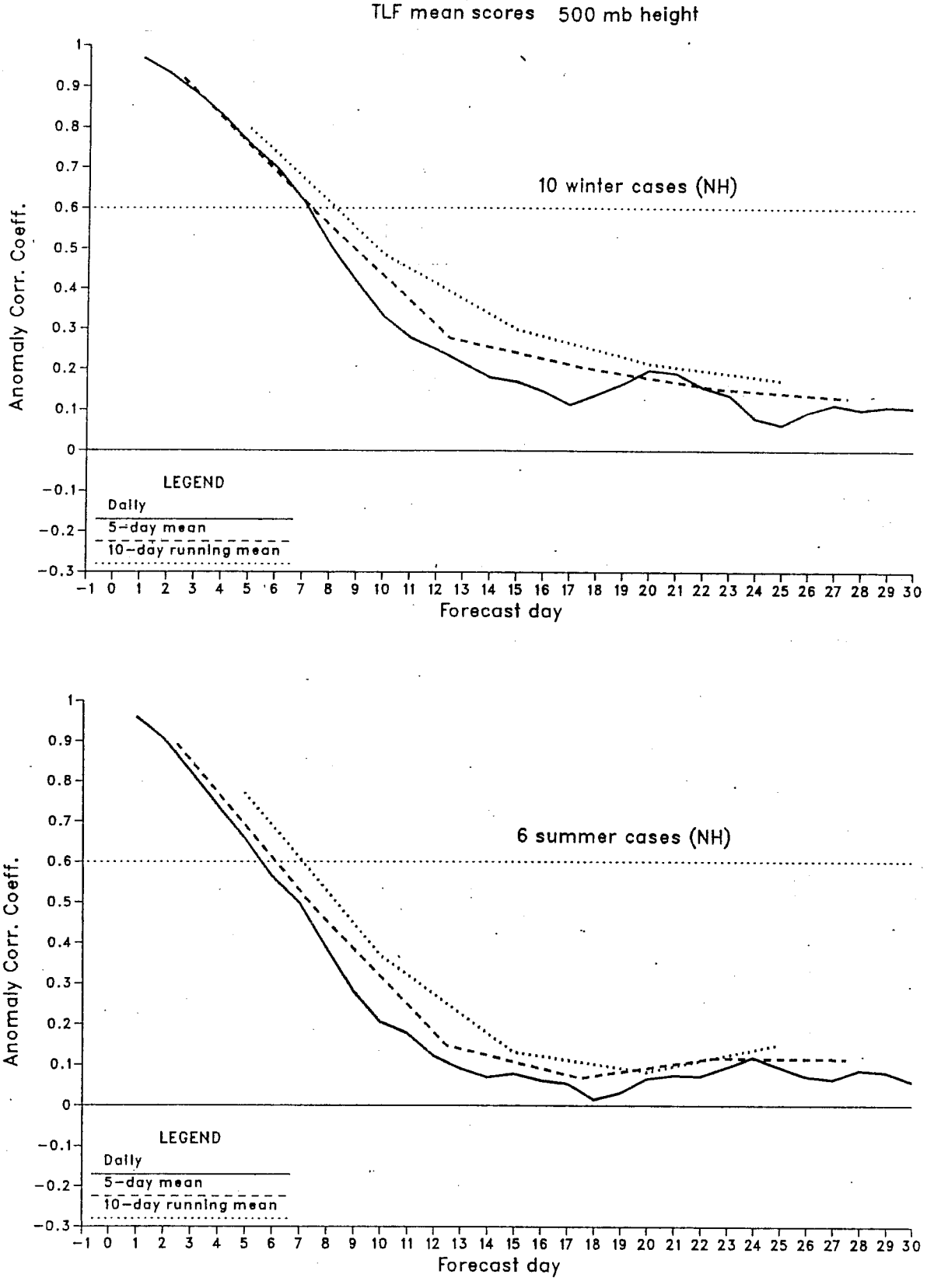

Fig.5 Time evolution of daily, 5-day and 10-day running average northern hemisphere 500 mb height ensemble-mean anomaly correlation coefficient averaged over cases from October to March (top) and April to September (bottom).

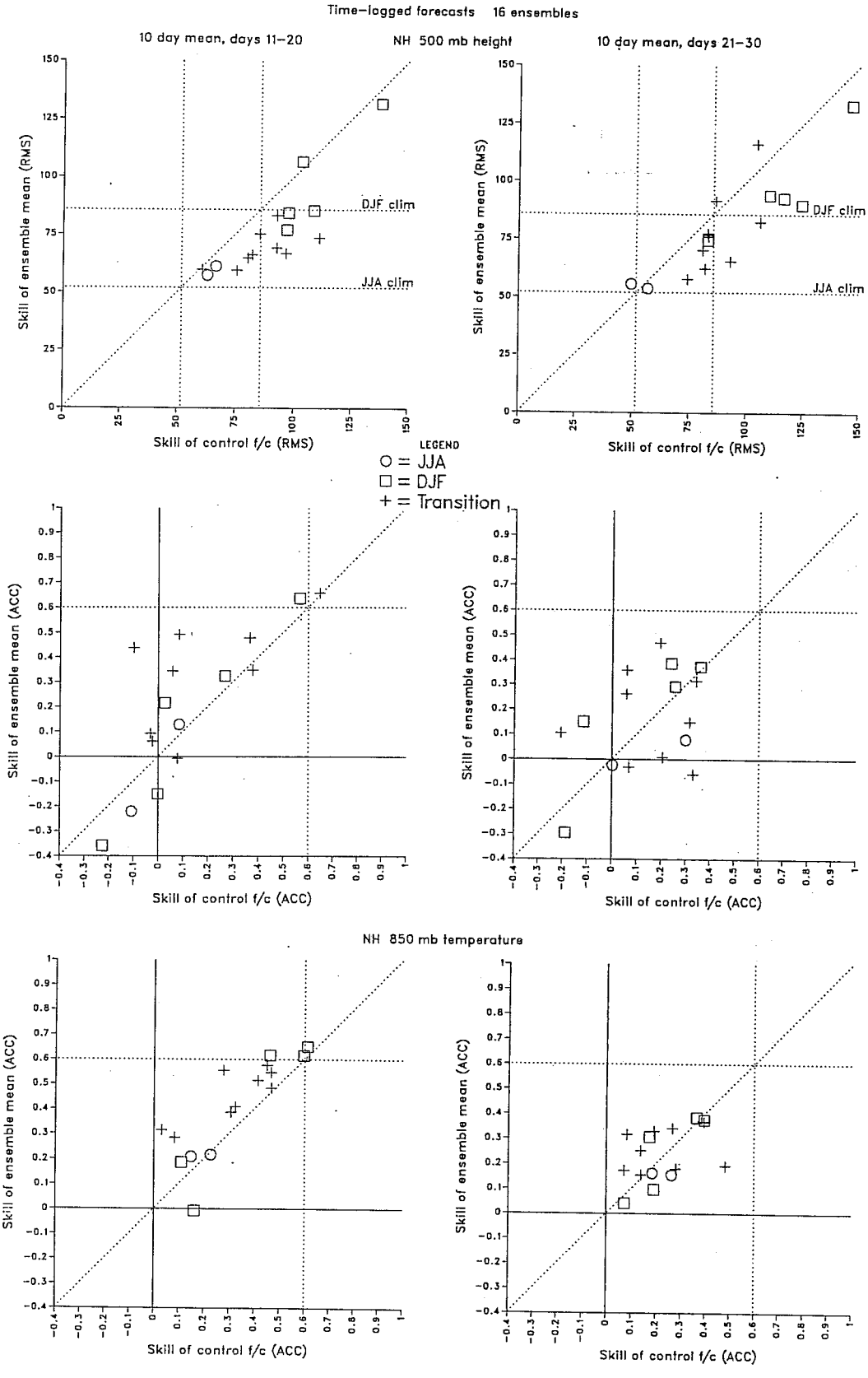

Fig.6 Scatter diagram of RMS error (top) and anomaly correlation coefficient (middle) of the northern hemisphere 500 mb heights, and anomaly correlation coefficient of the northern hemisphere 850 mb temperature (bottom) for ensemble-mean forecast vs. control forecast. Left hand column: days 11-20. Right hand column: days 21-30. RMS error in metres.

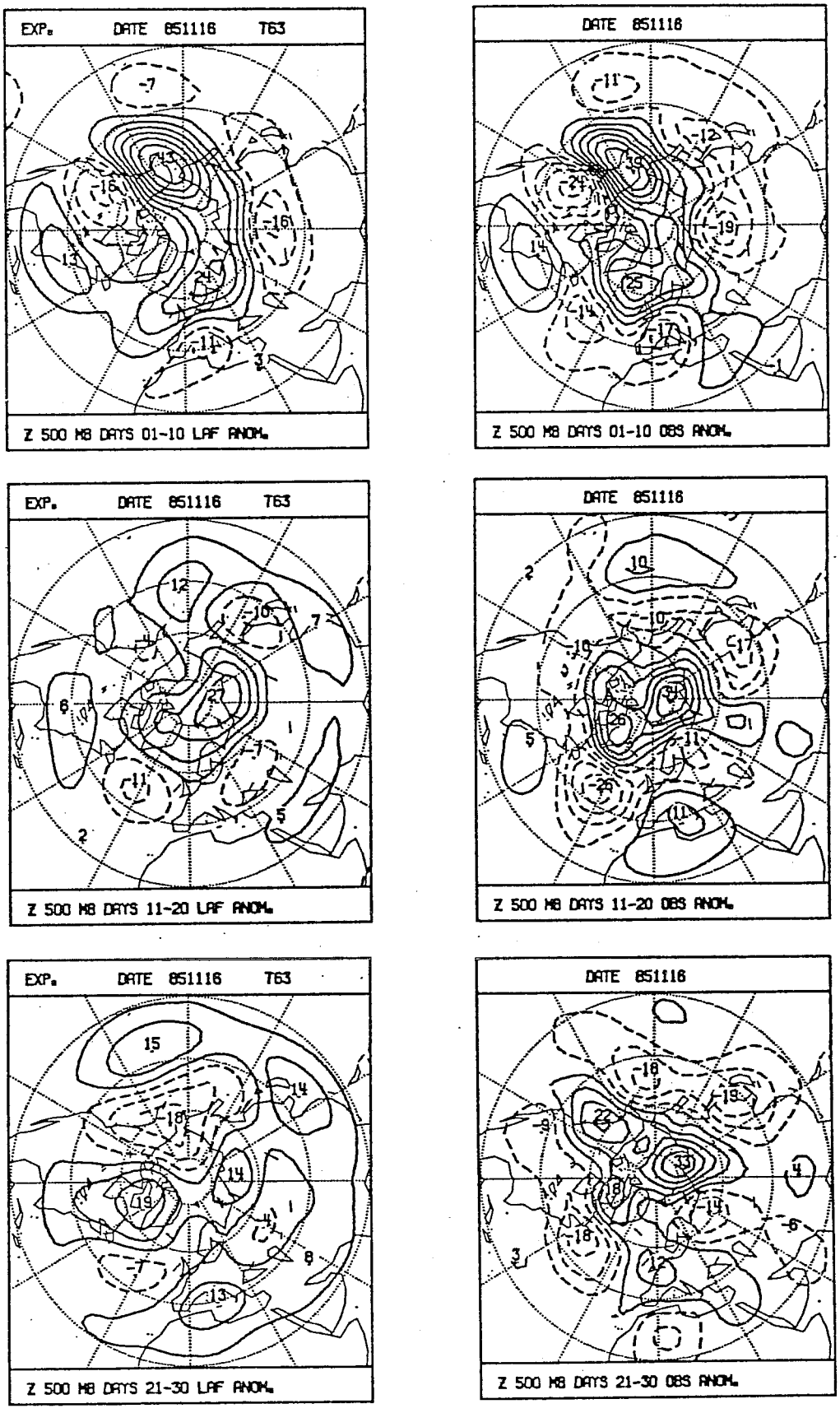

Fig.7 Ensemble-mean forecast (left) and verifying analysis (right) of days 1-10, 11-20 and 21-30 500 mb height anomaly for November 1985 ensemble. Contours every 6 dam starting from +3 (-3), positive anomalies solid, negative dashed; zero contour not drawn.

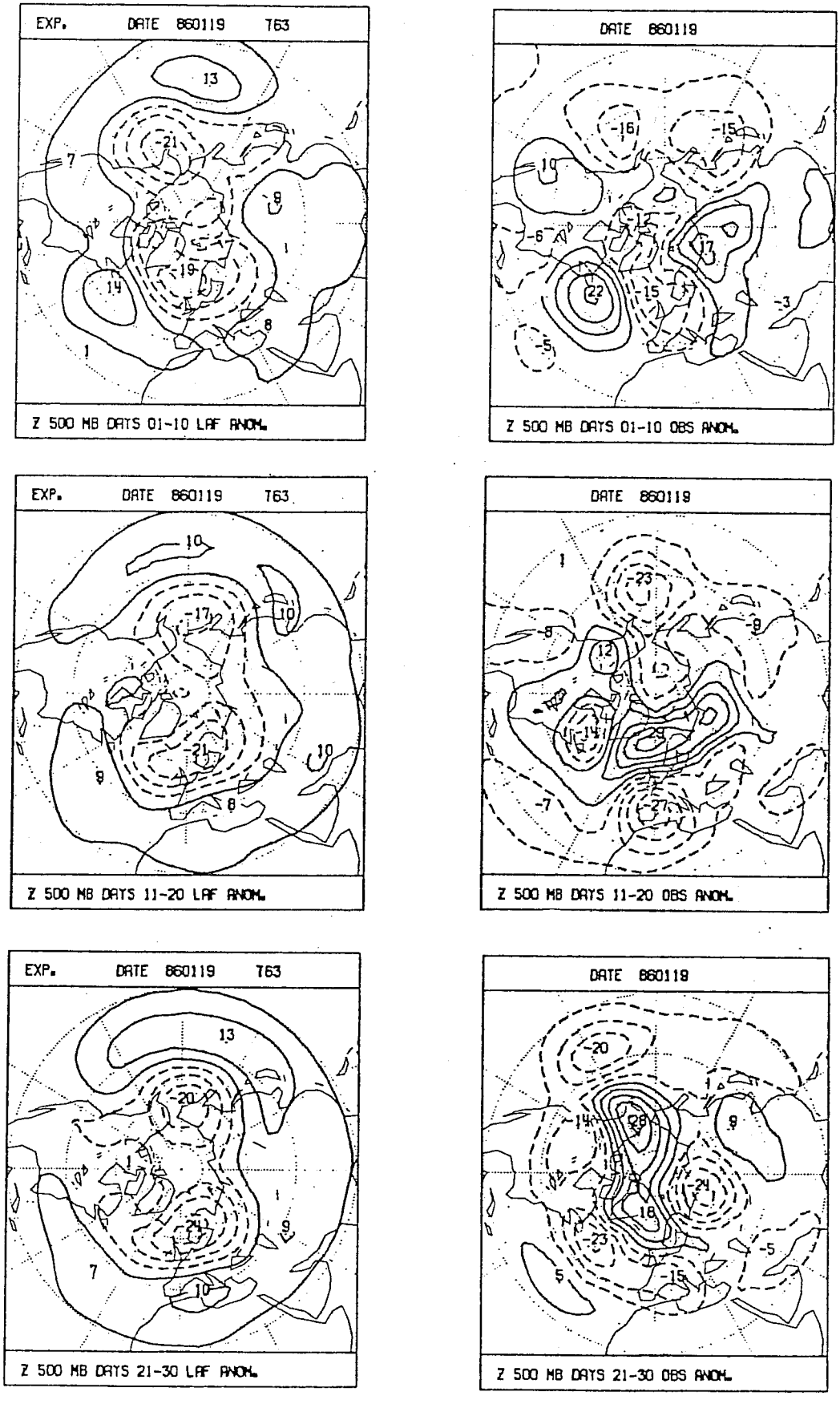

Fig.8 As Fig.7 but for January 1986 ensemble.

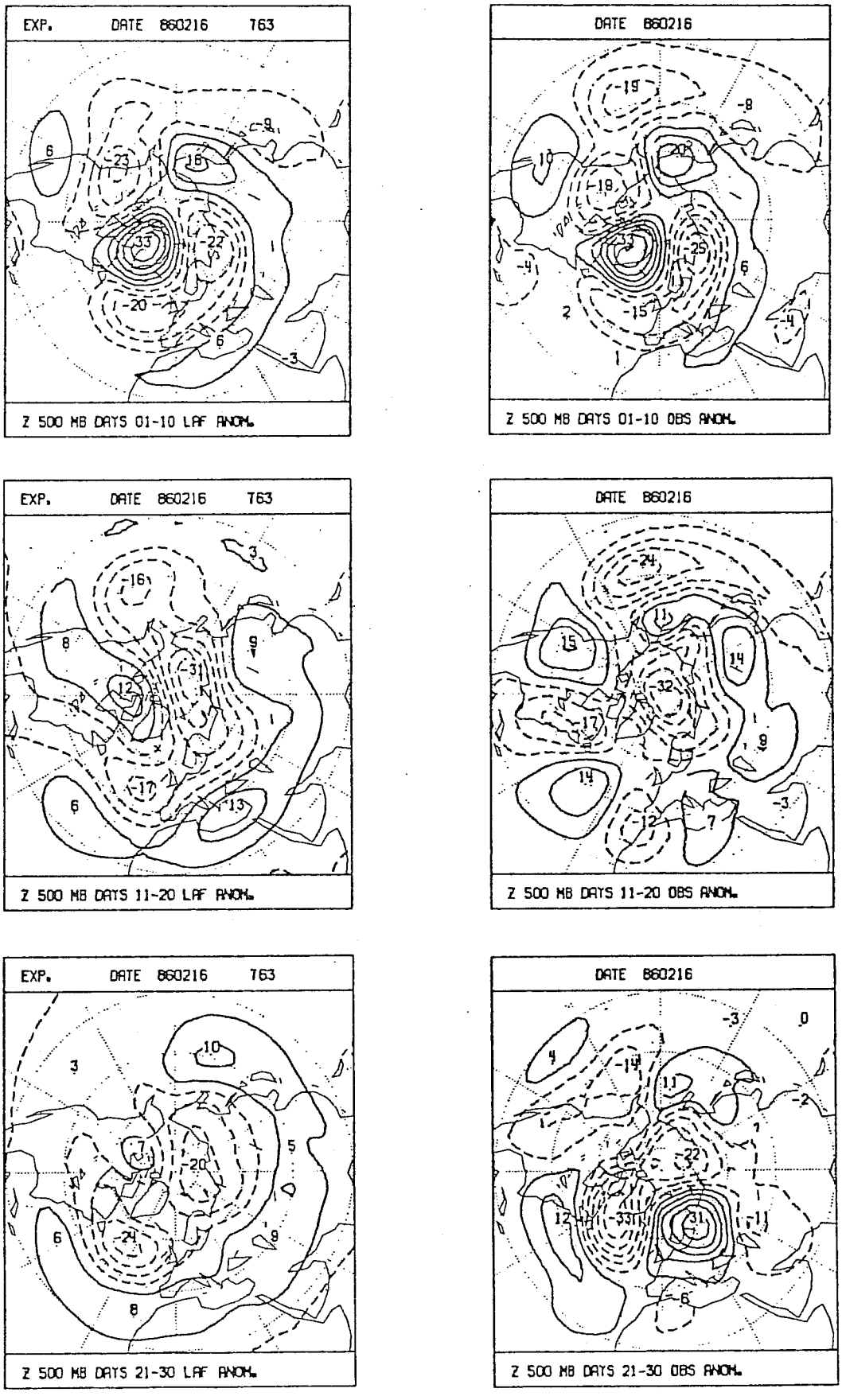

Fig.9 As Fig.7 but for February 1986 ensemble.

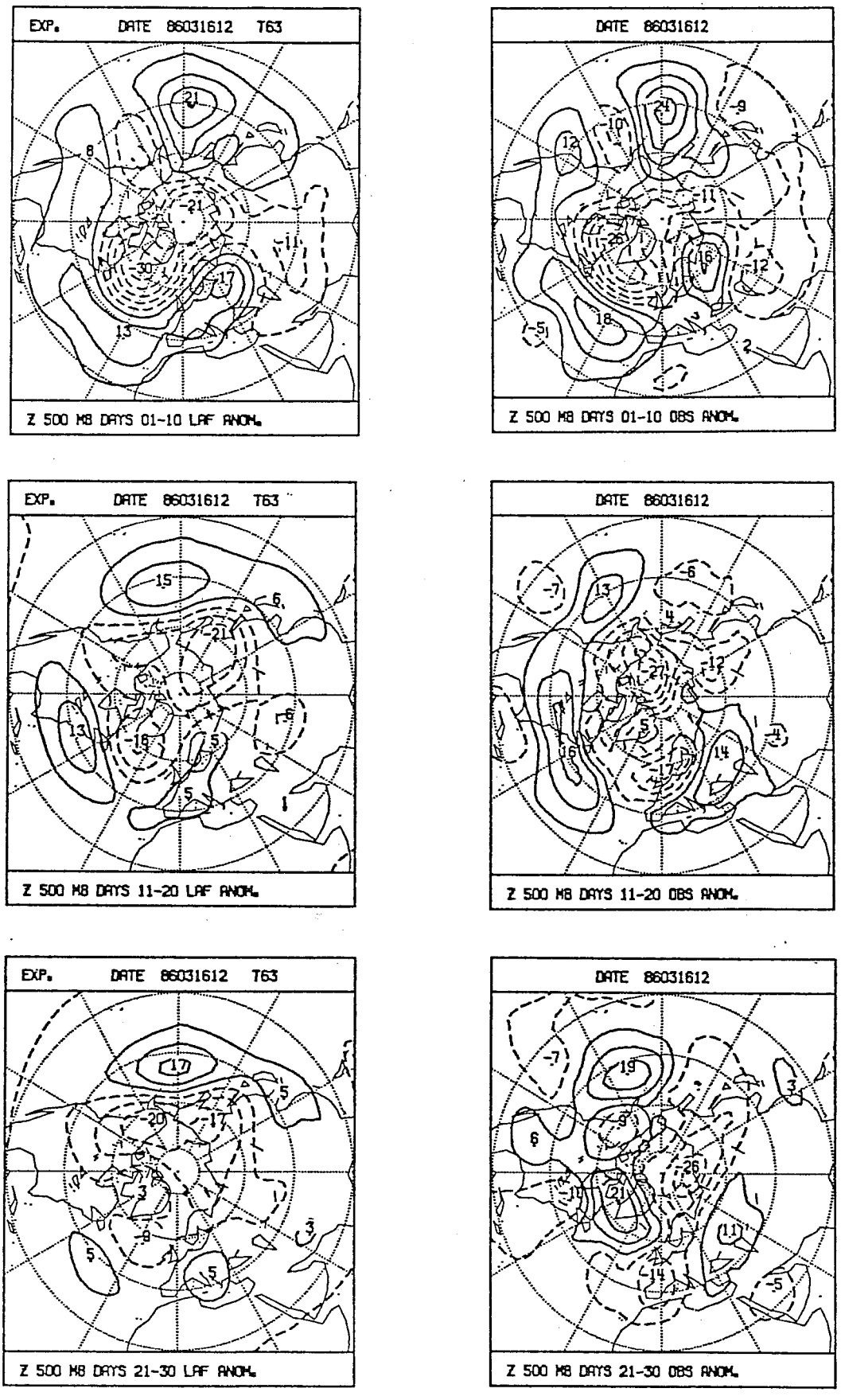

Fig.10 As Fig.7 but for March 1986 ensemble.

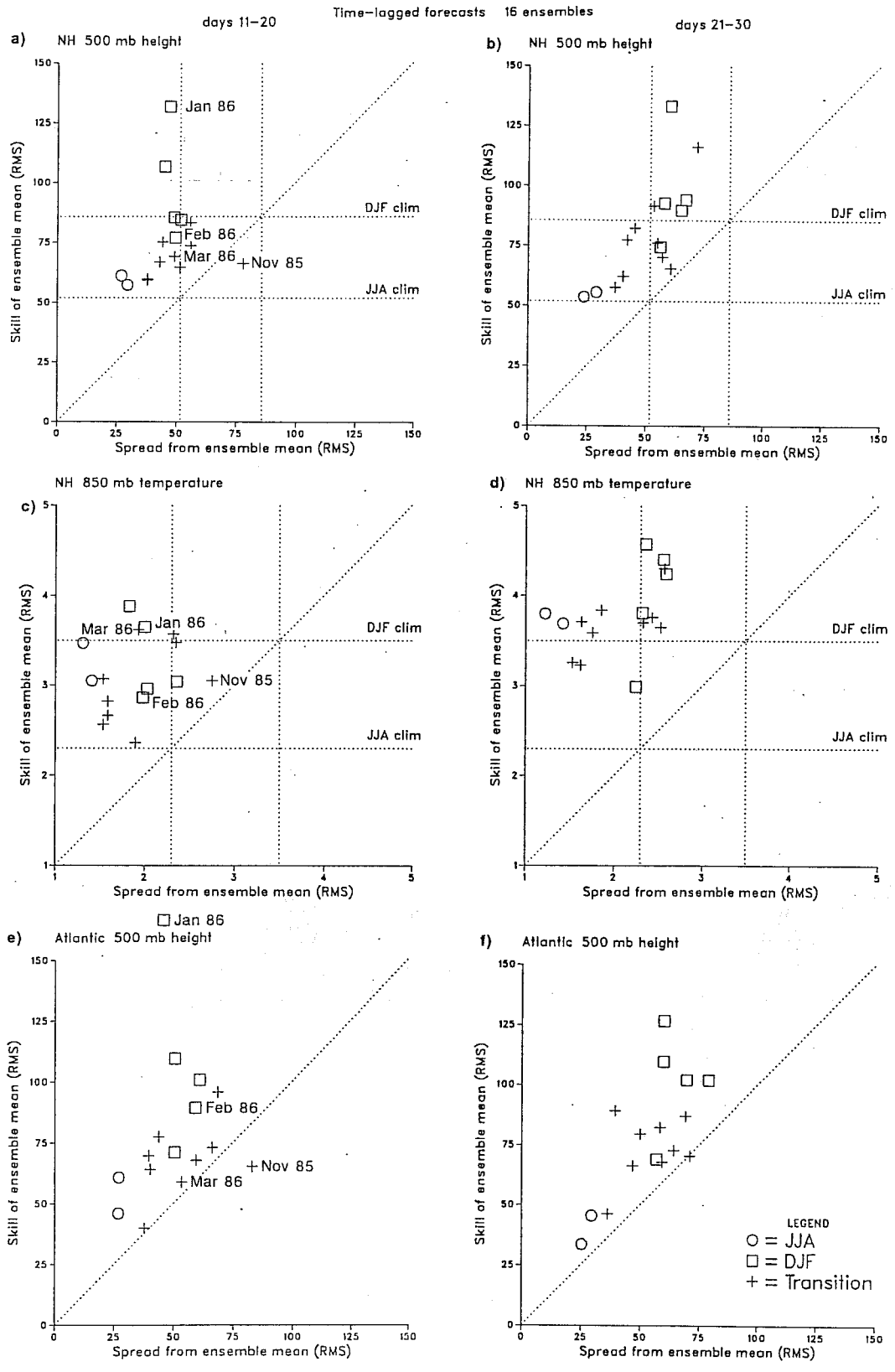

Fig.11 Scatter diagram of RMS ensemble spread (m) vs. RMS error (m): days 11-20 (left) and days 21-30 (right). Northern hemisphere 500 mb height (top), northern hemisphere 850 mb temperature (middle) and Atlantic region 500 mb height (bottom).

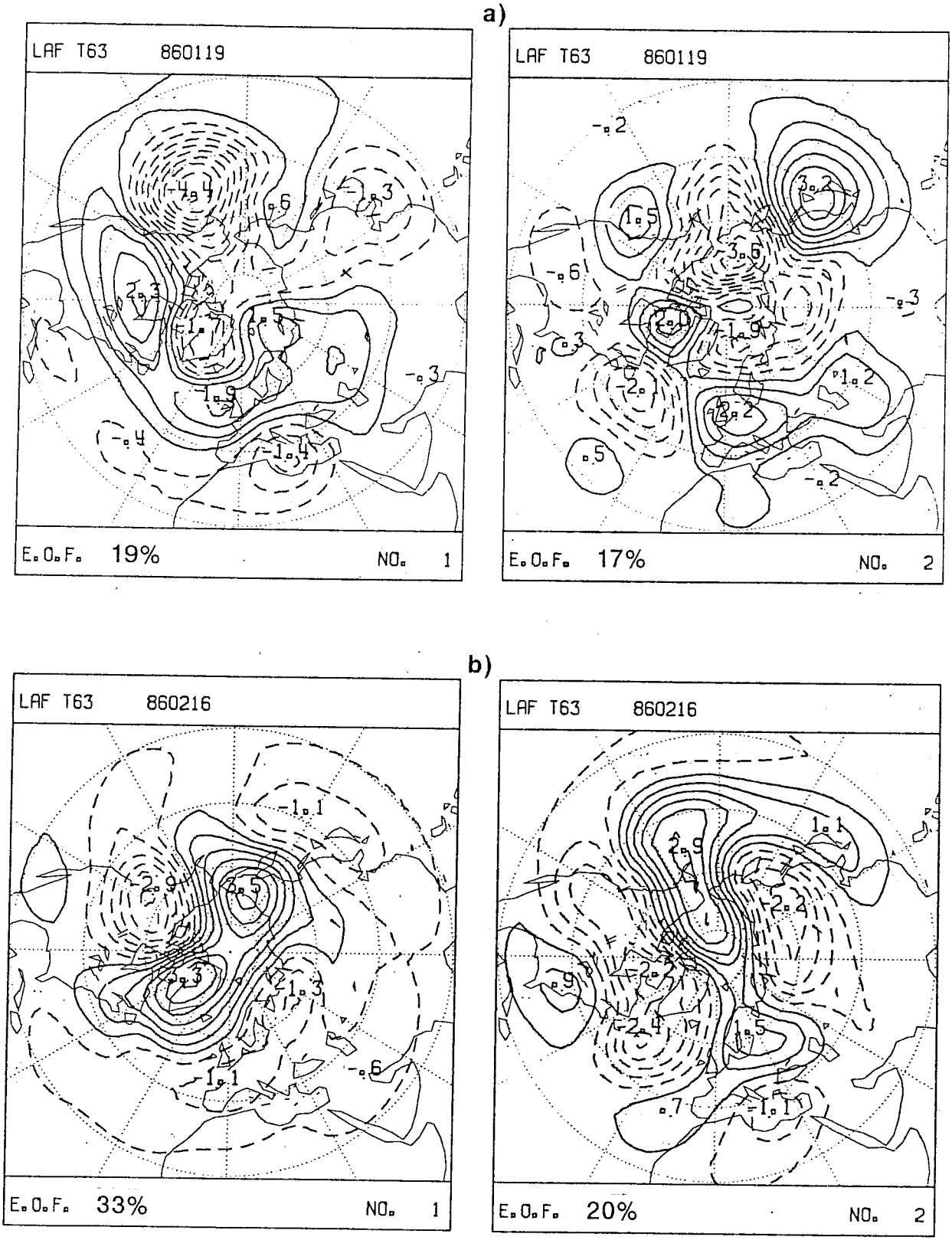

Fig.12 The first two EOFs of the 5-day mean 500 mb heights for a) January 1986, and b) February 1986 ensemble. Contour interval 0.5, starting at +0.25 (-0.25); zero contour not drawn.

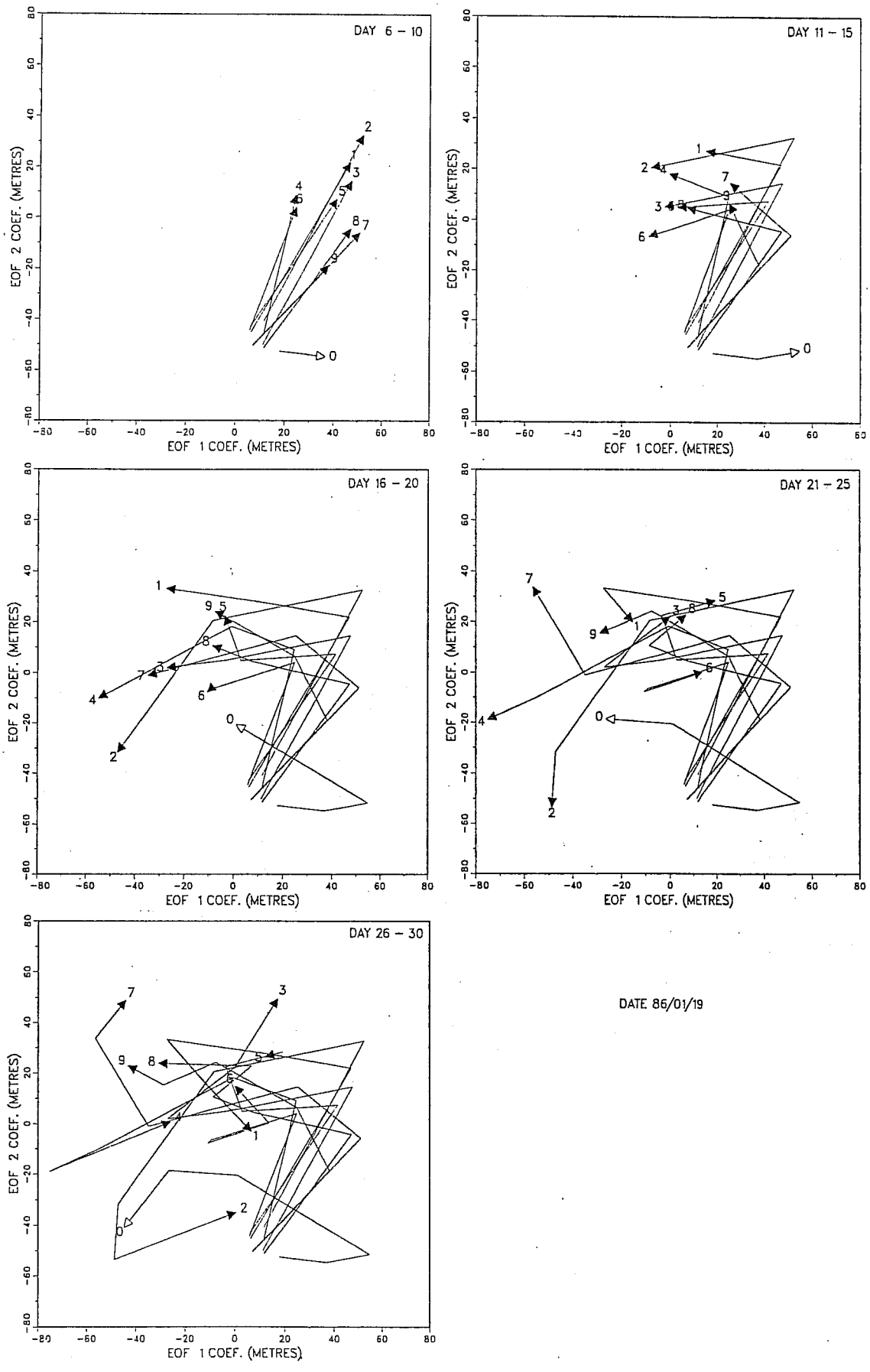

Fig.13 Phase space trajectories of the 5-day average northern hemisphere 500 mb heights in the plane defined by the first two EOFs of the January 1986 ensemble (see Fig.12). 1 to 9 denote individual forecasts as depicted in Fig.1. Zero and open arrow denote verifying analysis.

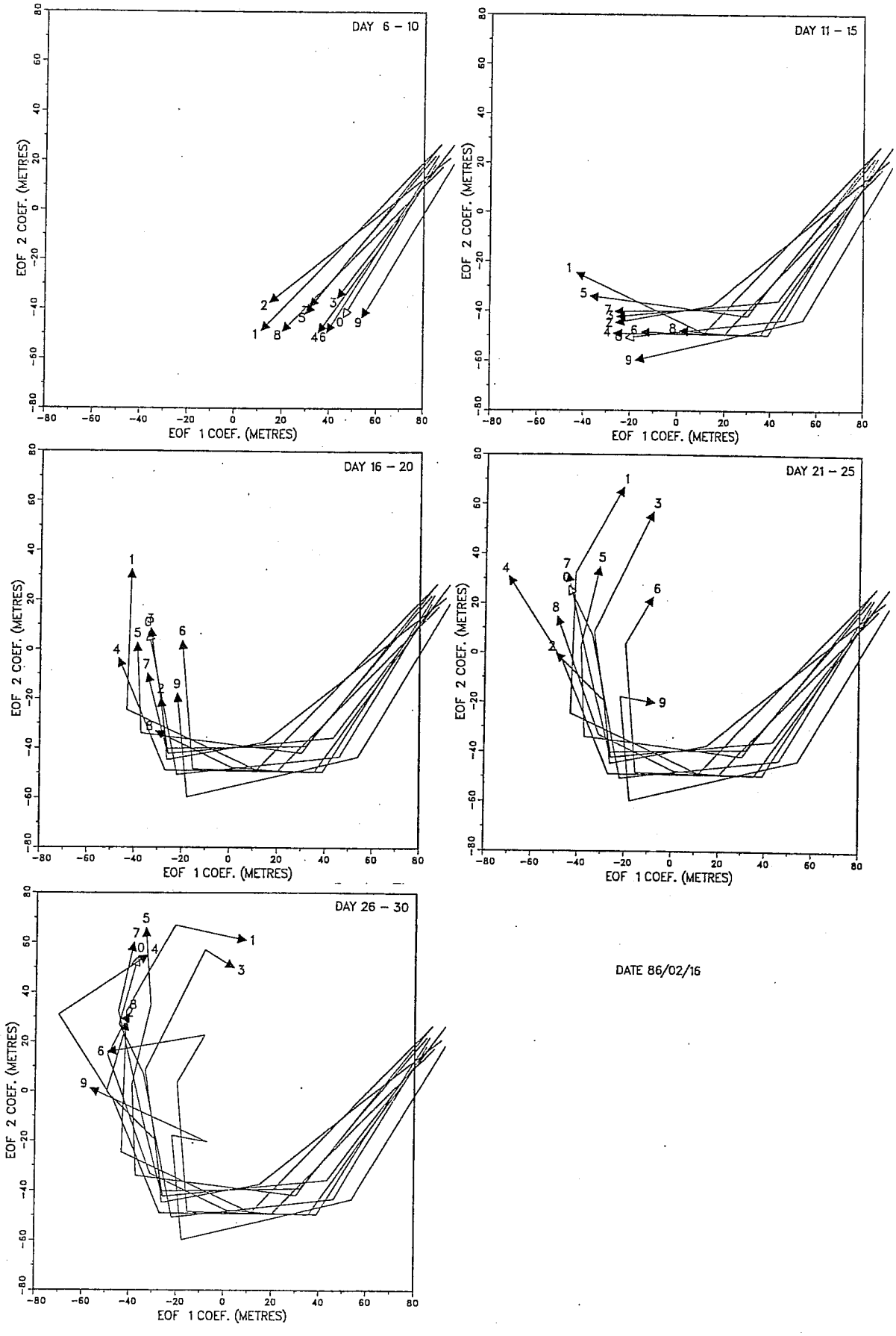

Fig.14 As Fig.13 but for February 1986 ensemble.

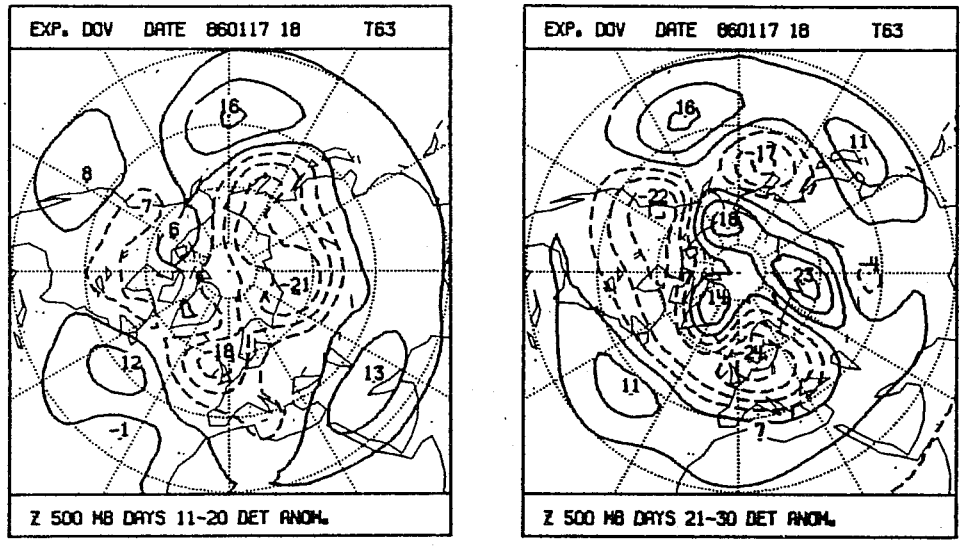

Fig.15 500 mb height anomaly of days 11-20 and 21-30 for forecast 2 of the January 1986 ensemble. Contours as in Fig.7.

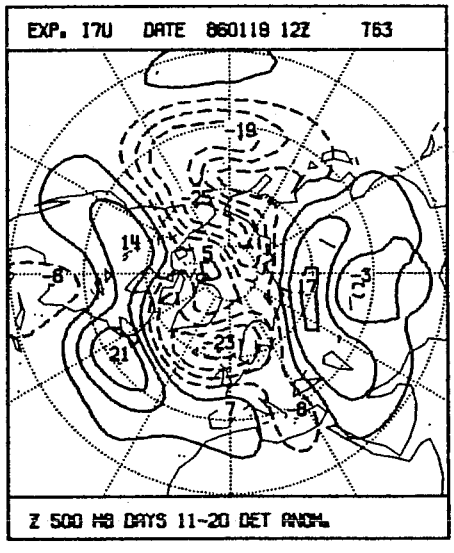

Fig.16 500 mb height anomaly of days 11-20 for repeated control forecast (19 January 1986 at 12Z) with the model cycle 30 (see text). Contours as in Fig.7.

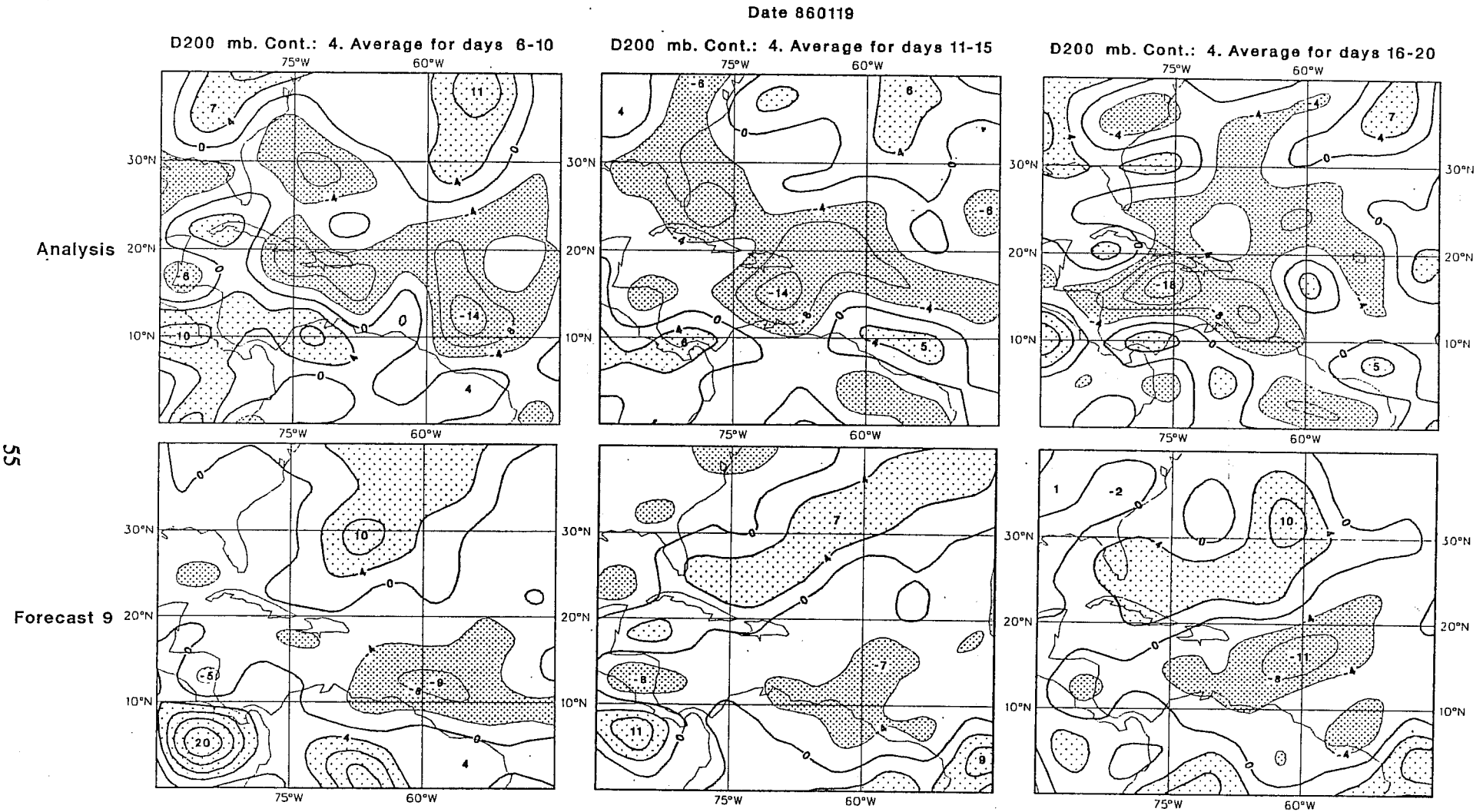

Fig.17 Analysis (top) and forecast from control integration of January 1986 ensemble (bottom) of the 5-day mean 200 mb divergence field (10^{-6} s^{-1}). Mean days 6-10, 11-16 and 16-20 shown. Contour interval 4. Areas above (below) +4 (-4) shaded; convergence dense stipple, divergence light stipple.

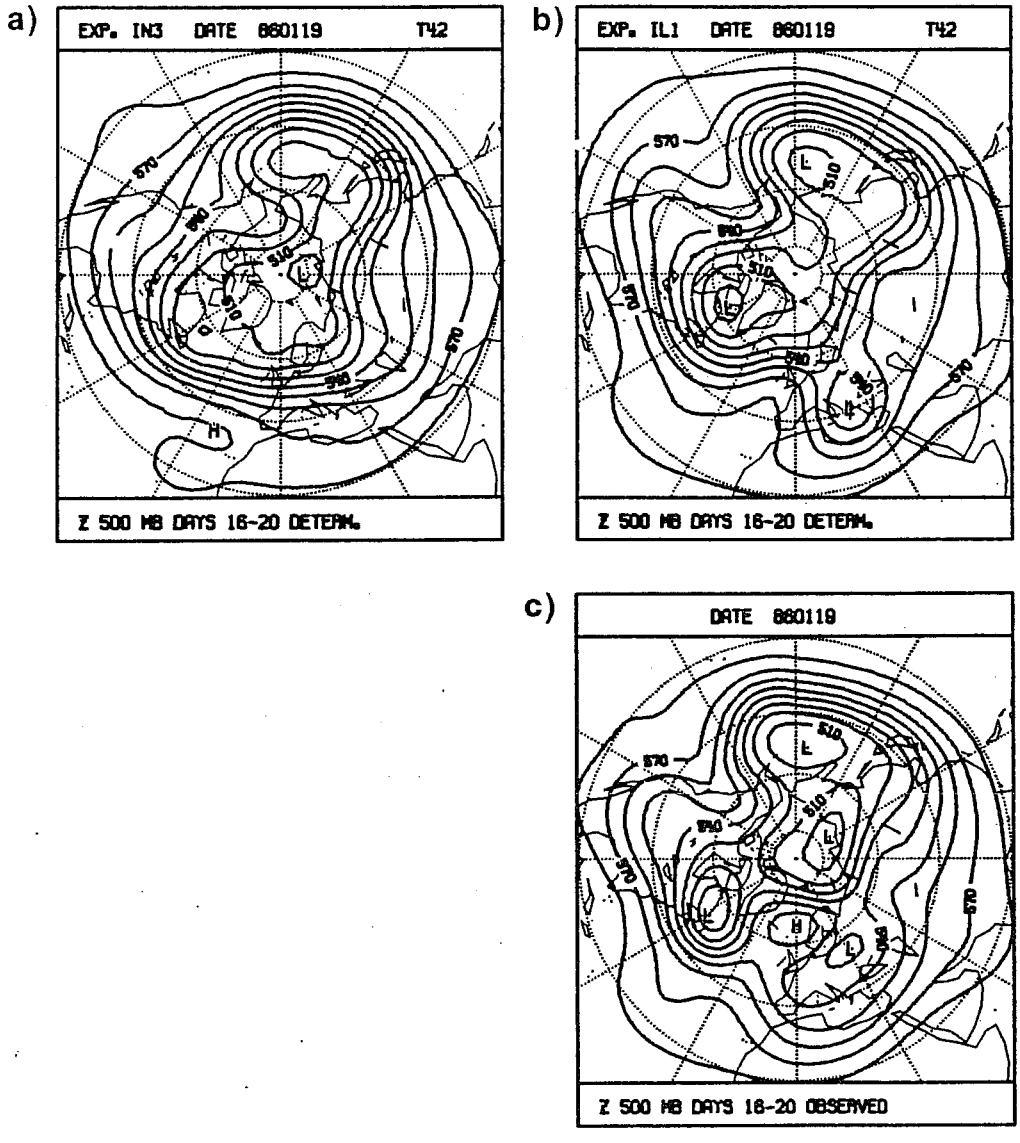

Fig.18 500 mb heights for days 16-20 of a) T42 control forecast, b) T42 relaxation experiment and c) verifying analysis. Initial date was 19 January 1986, 12Z. Contours every 10 dam.

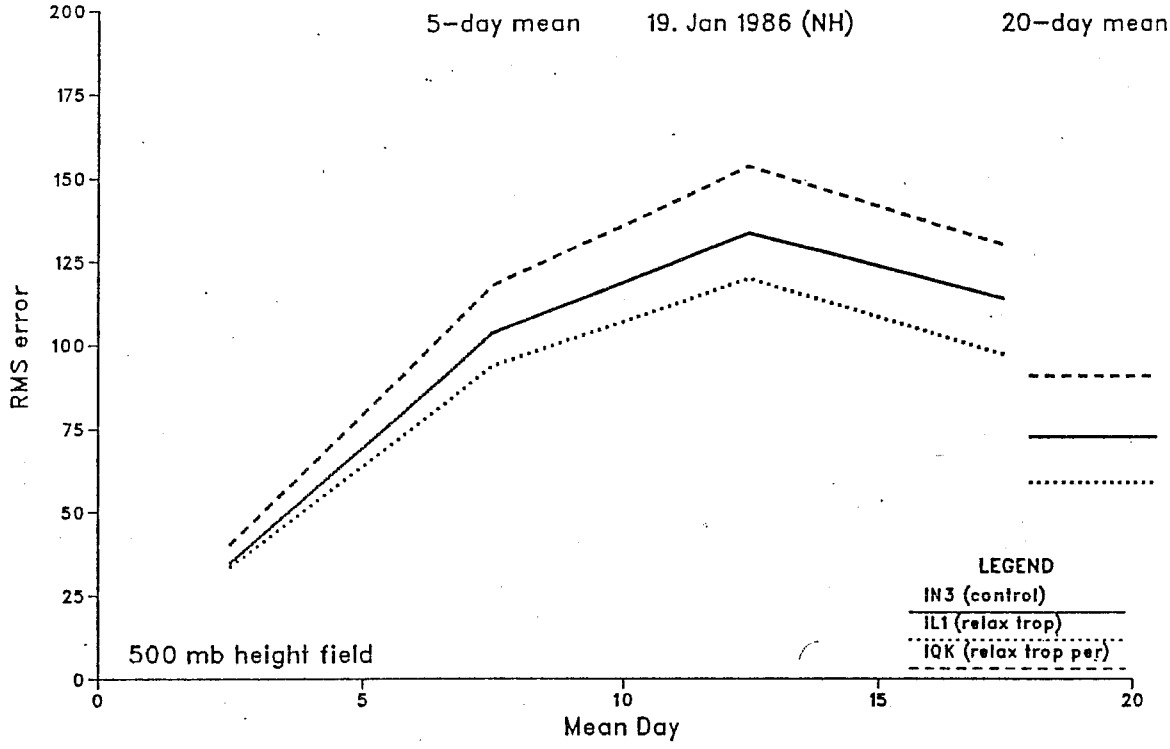

Fig.19 RMS error (m) of the 5-day and 20-day mean northern hemisphere 500 mb heights for T42 experiments from 19 January 1986, 12Z. Control solid, relaxation towards analysis dotted, relaxation towards persistence dashed.

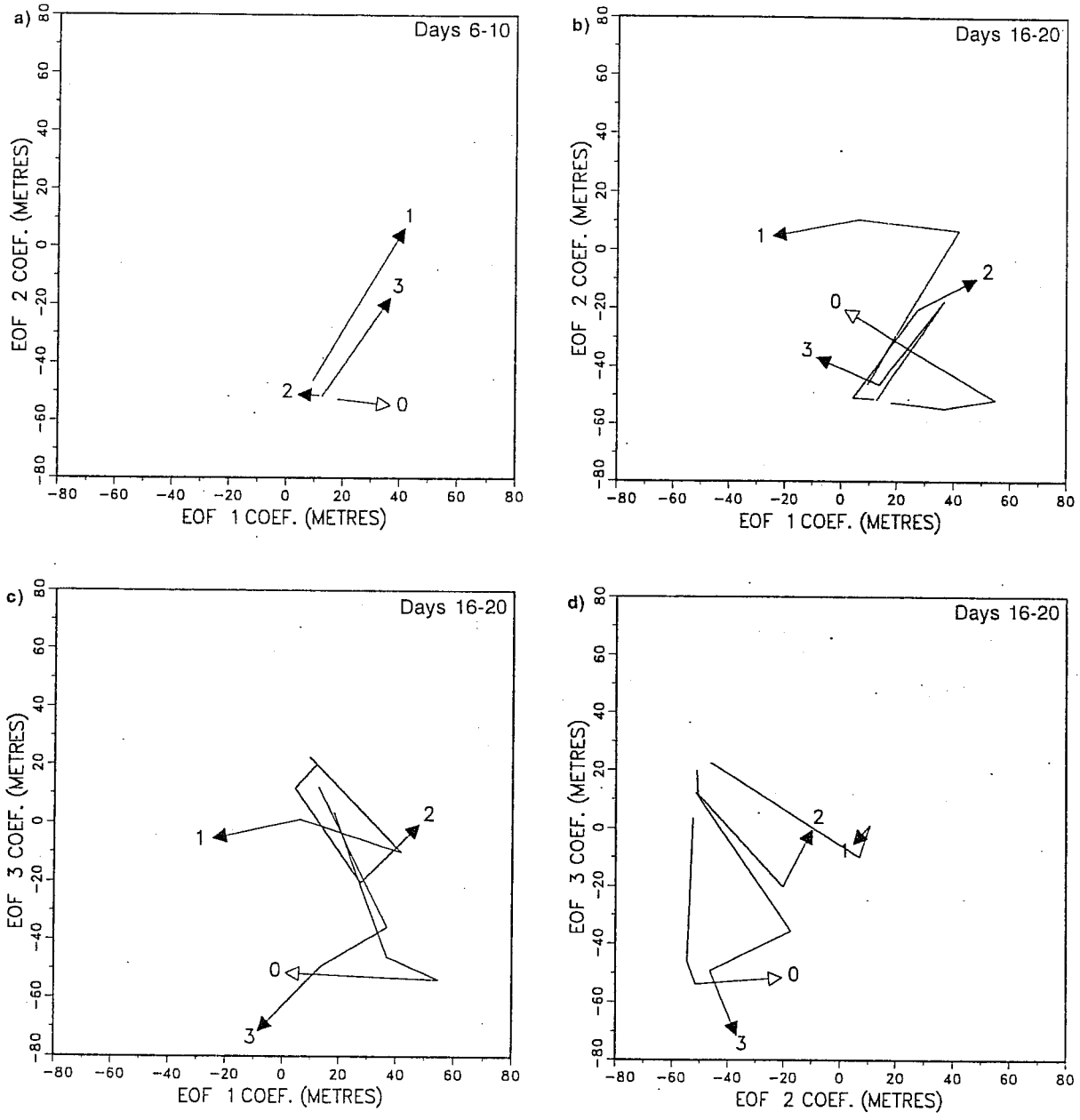

Fig.20 Phase space trajectories of the 5-day average northern hemisphere 500 mb heights: a) EOF1/EOF2 plane days 6-10, b) EOF1/EOF2 plane days 16-20, c) EOF1/EOF3 plane days 16-20, d) EOF2/EOF3 plane days 16-20. Arrow heads legend: '0' - verifying analysis, '1' - January 1986 ensemble-mean forecast, '2' - integration from 19 January with cycle 30 model, '3' - integration from 19 January with tropics relaxed towards analysis.

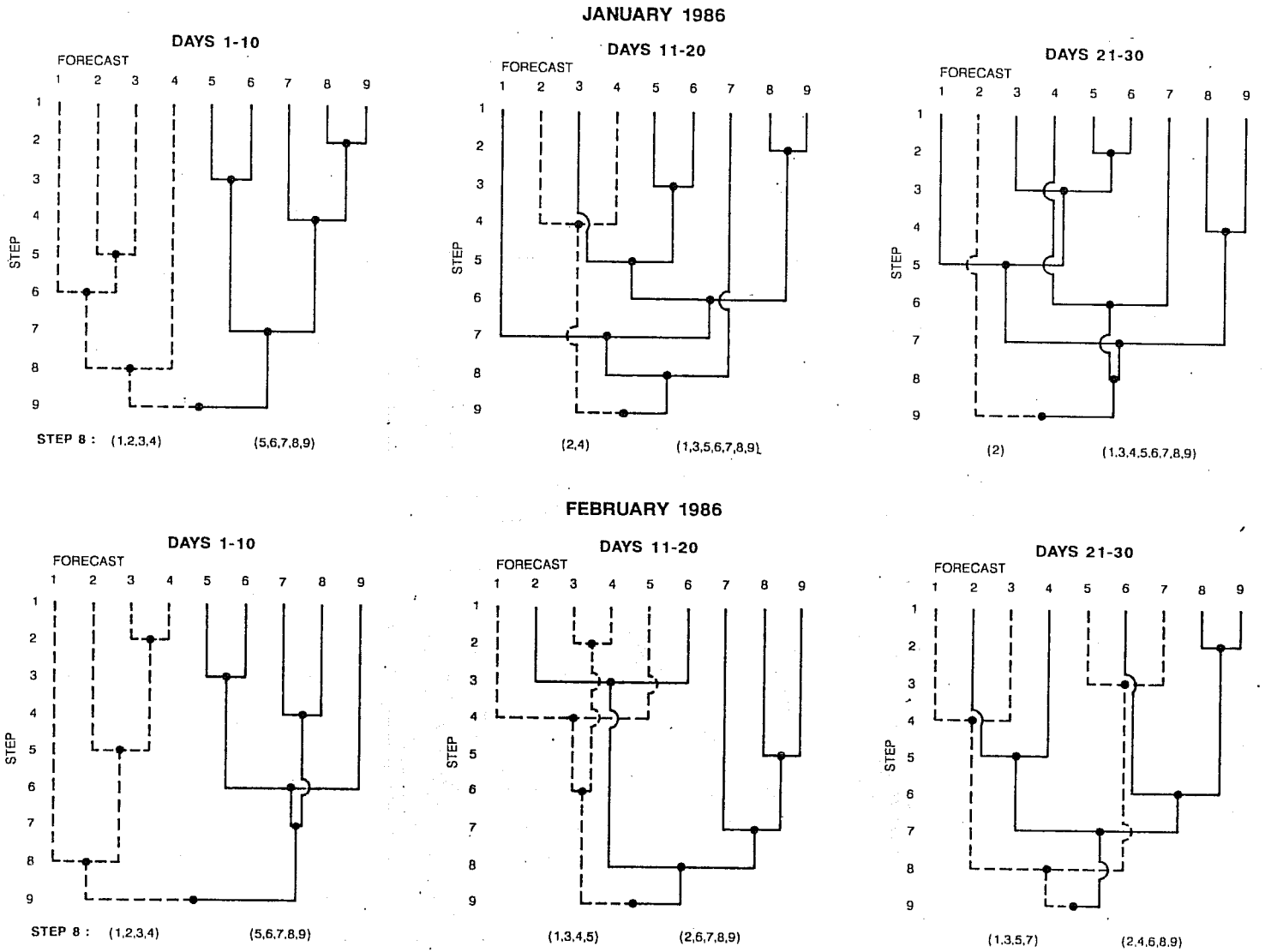

Fig.21 Clustering 'tree' diagram for the January 1986 ensemble (top) and February 1986 ensemble (bottom). Left hand column days 1-10; middle column days 11-20; right hand column days 21-30. Two clusters at penultimate (8th) step of each 10-day period shown at the bottom and forecasts belonging to these clusters are joined together with the same line style.

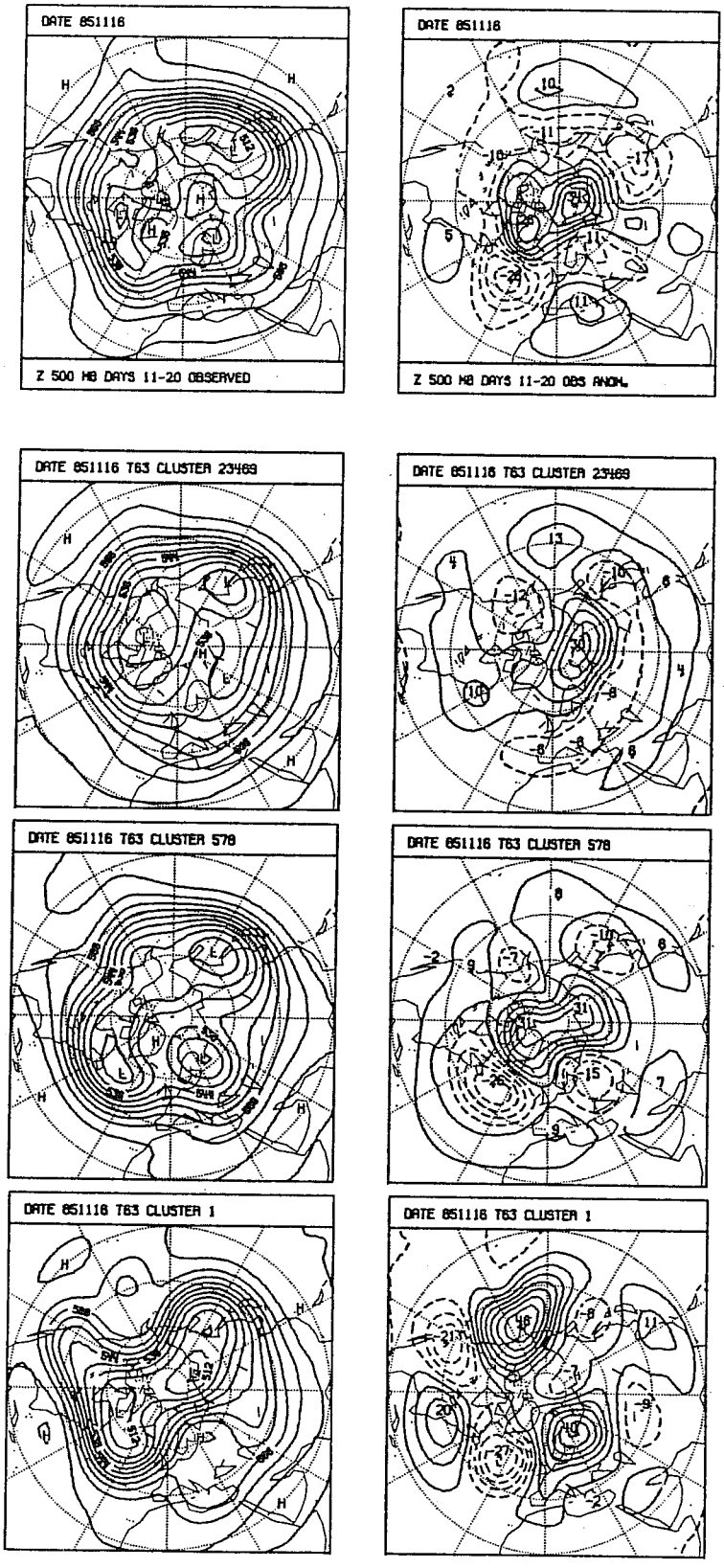

Fig.22 Verifying analysis of 500 mb height (top) and day 11-20 for the three forecast clusters associated with the November 1985 ensemble. Full field (left) and anomaly (right). Contours as in Figs.18 and 7 respectively.

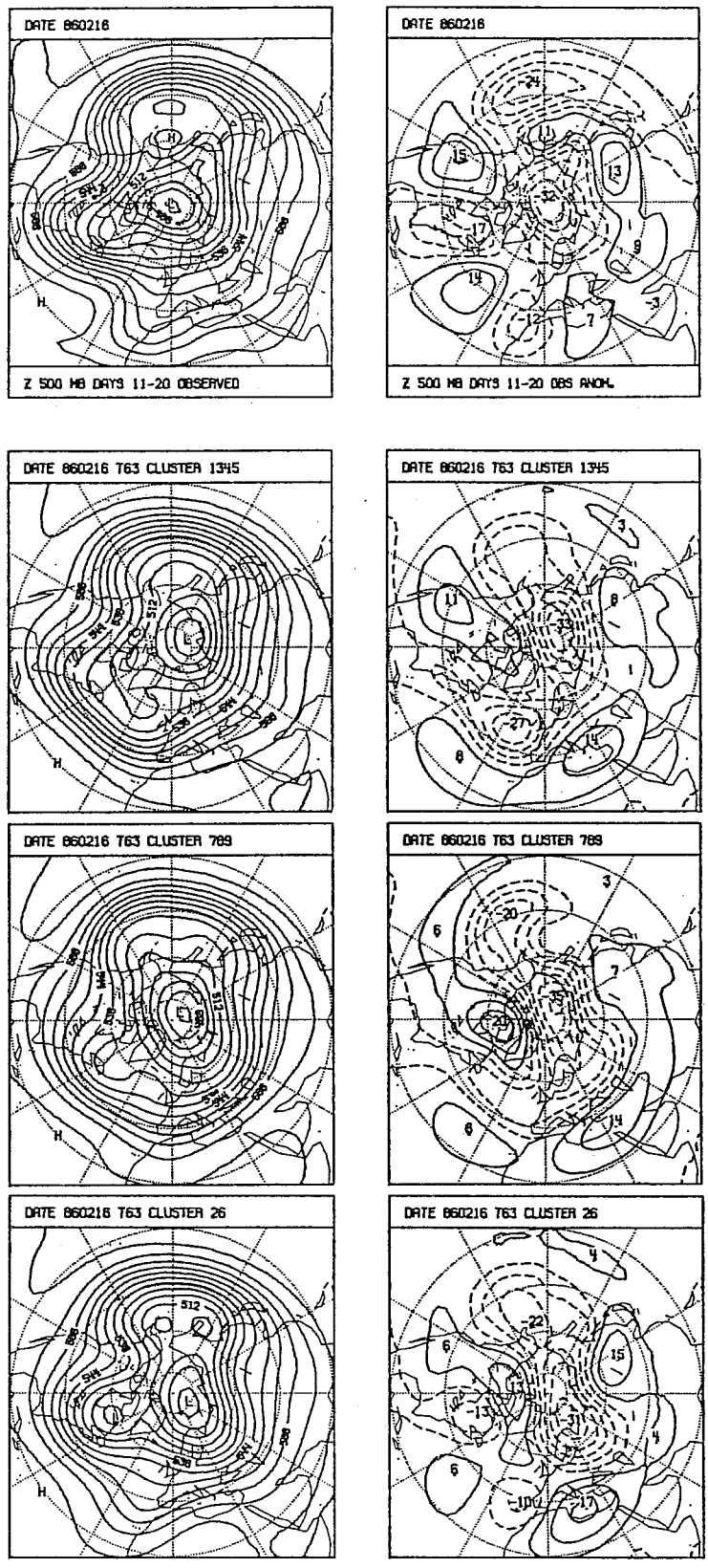

Fig.23 As Fig.22 but for February 1986 ensemble.