EXTENDED RANGE ENSEMBLE FORECASTING AT ECMWF

Č. Branković, F. Molteni, T.N. Palmer, S. Tibaldi* and U. Cubasch**
European Centre for Medium Range Weather Forecasts
Shinfield Park, Reading

1. INTRODUCTION

In the previous papers on the extended range forecasting at ECMWF, we have studied the extended range skill of the ECMWF numerical weather prediction (NWP) model, firstly over one year at a variety of resolutions (Tibaldi et al., 1988), and secondly over three years at the currently operational T106 resolution (Palmer et al., 1988). From Tibaldi et al. it was concluded that the extratropical extended-range skill of the model increased with horizontal resolution, though the performance at T63 was not significantly worse than that at T106 resolution. In Palmer et al. it was shown that variations in skill due to inter- and intra-annual variability of the atmosphere far exceeded variations associated with changes either in model formulation or model resolution.

The comparable results between the T63 and T106 models is perhaps of some relief for future studies, bearing in mind the present computational burden of integrating at the higher resolution. The substantial month to month and year to year variability in skill is also not discouraging. It suggests that with the present forecasting system, there may be occasions when the extended range forecast can give useful guidance. However, clearly this potential cannot be realised unless an a priori estimate of forecast reliability can be given.

The prediction of forecast skill is a topic of much interest at present, both for extended, medium and short range forecasting. Possible techniques for estimating forecast reliability include the use of stochastic dynamical models (Epstein, 1969), Monte Carlo forecasting (Leith, 1974), and statistical prediction techniques (Palmer and Tibaldi, 1988).

Problems of formulation, closure and sheer computational cost still require solution before the stochastic dynamic technique can be considered viable. 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.

---

* Department of Physics, University of Bologna, Italy
**Max-Planck Institut für Meteorologie, Hamburg, Federal Republic of Germany

The notion of Monte Carlo forecasting originated as an alternative to the stochastic dynamic model and was defined by a finite sample or ensemble of integrations of a deterministic model, each one obtained by randomly perturbing the spectral coefficients (or grid point values) of an operational analysis. As such, this technique is not appropriate in a forecast system with nonlinear normal mode initialisation, even when the wind and mass fields of the perturbations are in geostrophic balance (Hollingsworth, 1980).

However, a very simple and convenient way (in an operational forecast environment) of effectively obtaining an ensemble of balanced initial states for a Monte Carlo forecast ensemble, is through the time-lagged approach advocated by Hoffman and Kalnay (1983). In the current operational analysis/forecast system at ECMWF where analyses are produced at 6 hourly intervals, an n-member ensemble at time t=0 would comprise the operational analysis at t=0, the 6 hour 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.

As mentioned above, this technique is particularly easy to implement in an operational environment, and does not suffer from the problems of initial ensemble perturbations projecting onto unbalanced modes. However, neither is it a perfect system. For example, a basic premise is that large ensemble dispersion should give an indication of a poor forecast, and small ensemble dispersion should give an indication of a reliable forecast. In principle, the time-lagged technique could fail in both these respects when applied in the ECMWF operational environment. For example, since the initial perturbations of the ensemble should reflect uncertainty in the initial analysis, the perturbations should be weighted towards data sparse regions. However, in a region of no data, the analysis is taken from the first guess field, which is the 6-hour forecast. Hence, the effective time-lagged Monte Carlo perturbation would be zero in a region of no initial data. In these circumstances, the notion that ensemble dispersion is a predictor of forecast reliability can be seen to be doubtful, at least in the first few days of the integrations. Secondly, when the size N of the ensemble is sufficiently large, the Nth perturbation (at time t=0) becomes an unrealistic measure of analysis uncertainty both because of internal and external error growth.

Whilst these drawbacks may be serious for short and medium range ensemble forecasting, the simplicity of the method may outweigh its disadvantages when studying the extended range forecast problem. For example, when assessing the skill of a monthly forecast, one cannot assume that the skill of today's forecast will be significantly less than the skill of yesterday's forecast, or even the day before yesterday's. Furthermore, the geographical distribution of the initial perturbations, and their association with data sparse areas, may not be so important for the extended range problem. For example, beyond the first few days, initial analysis errors will propagate downstream and excite the dominant modes of instability of the flow field. Ultimately, the structure of forecast errors will depend more on the geographical distribution of these modes of instability than on the geographical distribution of the initial analysis errors.

In the present paper, we therefore study the problem of predicting extended range forecast skill using the time-lagged approach. Following earlier studies (Leith, 1974; Hoffman and Kalnay, 1983) we recognise that about 10 integrations are necessary to form a large enough sample. This would cause practical computing problems if it was necessary to integrate the model at the currently operational T106 resolution. However, as mentioned above, the results in Tibaldi et al. (1988) can be used to justify integrating at the T63 resolution. This paper, then, is a study of extended range time-lagged forecast integrations of the operational ECMWF model truncated at T63.

In section 2 of this paper, we describe the database and the time-lagged ensemble forecasts that have been performed. In section 3 we present some properties of ensemble forecasts which can be obtained within a simple theoretical framework. Summary results on forecast skill and ensemble spread are given in section 4.

Having established a methodology for extended range ensemble forecasting, it is necessary to establish a procedure for postprocessing the results. The ensemble mean forecast is a simple and convenient way of collating the results from individual members. The RMS error of such a consensus forecast is trivially smaller than the mean RMS error of the members of the ensemble. (As will be discussed below, the magnitude of the anomaly correlation coefficient of the ensemble mean forecast is also, under certain circumstances, trivially

greater than the magnitude of the mean anomaly correlation coefficient of the individual forecasts). Maps of the standard deviation of the ensemble can give an estimate of the reliability of the consensus forecast but give no indication of possible alternative forecast flows to the ensemble mean. A simple way of indicating possible alternatives is through maps showing the probability that a forecast variable lies within certain predefined bounds. In this way, the probability of an extreme event (e.g. wind speeds exceeding three standard deviations from climatology) can be indicated. Such a probabilistic analysis will be given in section 5 for one selected ensemble. In section 6 we describe results from three selected ensembles in terms of somewhat unconventional diagnostics - phase space trajectories. We use these in an attempt to understand the physical differences between the evolution of the ensembles. Concluding remarks will be made in section 7.

2. THE DATABASE AND EXPERIMENT DESIGN

An ensemble of time-lagged forecasts (TLF) in the present study is comprised 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 it 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 TLF dates is given in Table 1. From September 1985 until march 1986 the time-lagged forecasting was performed every month. After this period they were run in 3-month intervals. Asterisks in Table 1 denote those ensembles which are discussed in some detail in sections 5 and 6.

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 TLF ensemble.

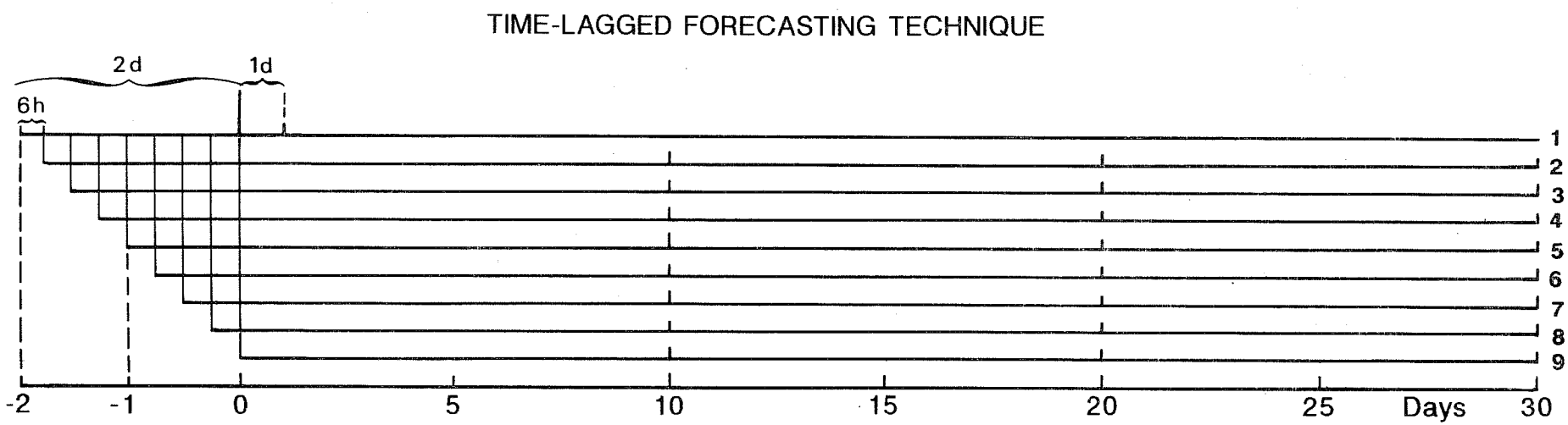


Fig. 1 Schematic of the construction of a time-lagged forecasting ensemble with nine individual deterministic forecasts starting from adjacent initial analyses separated by 6 hours.

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	Introduction of the 19-level model
10.	14 September 1986	Introduction of the GWD 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	Introduction of new vertical diffusion scheme

Table 1

Initial dates of a set of 16 TLF ensembles run with T63 ECMWF operational spectral model. Changes in the model are indicated on the right-hand-side. Those initial dates denoted by an asterisk are cases discussed in detail in sections 5 and 6.

However, these differences (if any), by virtue of their slowly varying nature, are negligible.

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).

Until March 1986, the same version of the ECMWF operational model was used for all TLF 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., 1988). From September 1986 the TLF ensembles were run with a model which included parametrization of the gravity wave drag (Palmer et al., 1986; Miller et al., 1988). 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 TLF 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.

3. THEORETICAL BACKGROUND

Before looking at the actual results obtained from our T63 time-lagged 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 simply a particular realization of an ensemble forecast, and the results deduced by Leith (1974) and Seidman (1981) can be considered as representative of its theoretical behaviour, at least in the perfect model assumption.

The practical experience of Molteni et al. (1986) and Murphy (1988) showed that in general, an ensemble-mean forecast is indeed more skilful than a deterministic forecast but this improvement is far from being as large as expected from the perfect model theory. Molteni et al. also noted a slight 'return' of the forecast skill, that is a weak increase in skill scores, towards the end of the integrations period, the cause of which was not clear at the time. (Such a return has sometimes been observed also in purely deterministic forecasts, see for example Cubasch and Wiin-Nielsen (1986) and Section 3.3 in Molteni et al. (1986) for a possible explanation).

In this section we shall show that a parametrization of NWP errors in a simple analytical model for error growth can lead not only to more realistic theoretical values for the skill of an ensemble forecast, but also to an explanation of its apparent return of skill.

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 ($i=1, \ldots, N$). 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, $\bar{E} = \bar{F} - A$ 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 observed climate X_c (i.e. put $X' = X - X_c$ for all X, and drop all primes hereafter). We can define the following variables describing the statistical properties of the ensemble:

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

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

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

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

f^2 is the ensemble average of the spatial variance of individual members (or deterministic forecasts), 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) and setting X to various fields we obtain relationships between variables defined by (2a) - (2d). For X=0 it follows

$$f^2 = |\bar{F}|^2 + \Delta^2, \tag{3}$$

whereas if we set X=A we obtain

$$e^2 = |\bar{E}|^2 + \Delta^2. \tag{4}$$

Eq. (4) quantifies the average improvement of the ensemble-mean forecast over the individual members in terms of mean-square error, and Eq. (3) clarifies 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:

$$\delta^2 = \frac{2N}{N-1} \Delta^2. \tag{5}$$

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

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

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

$$\rho_i = \frac{F_i \cdot A}{|F_i| \cdot |A|}$$

$$= \frac{|F_i|^2 + |A|^2 - |F_i - A|^2}{2\, |F_i| \cdot |A|} \quad , \tag{7a}$$

and for the ensemble mean as

$$\rho(\bar{F}) = \frac{\bar{F} \cdot A}{|\bar{F}| \cdot |A|}$$

$$= \frac{|\bar{F}|^2 + |A|^2 - |E|^2}{2\, |\bar{F}| \cdot |A|} \quad . \tag{7b}$$

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

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

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 ($|F_i| > |\bar{F}|$), then from Eq. (7) $|\rho(\bar{F})| > |\bar{\rho}|$. Whilst in practice this assumption holds in many situations, it is easy to construct an artificial case where it does not. Consider two anomaly fields $F_1 = C_1 \cos\lambda$, $F_2 = C_2 \cos(\lambda + \pi/4)$ where λ is longitude, and we ignore latitude and height variations. If the verifying field $A = C_3 \cos\lambda$, we clearly have $\bar{\rho} = \frac{1}{2}(1 + 1/\sqrt{2})$. A simple calculation gives

$$\rho(\bar{F}) = \frac{\varepsilon + 1/\sqrt{2}}{(\varepsilon^2 + \sqrt{2}\ \varepsilon + 1)^{\frac{1}{2}}}$$

where $\varepsilon = C_1/C_2$. In the limit $\varepsilon \to 0$, $\rho(\bar{F})$ is clearly less than $\bar{\rho}$, and more generally it is easily shown that $\rho(\bar{F}) < \bar{\rho}$ when $\varepsilon < \varepsilon_o = .82$ Clearly 'ensemble averaging', when $\varepsilon < \varepsilon_o$, fails to 'improve' the mean skill because the spatial variance of F_1 is smaller than the ensemble mean.

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

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

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 (and consequently of Δ^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

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

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

3.2 A theoretical model for skill and spread

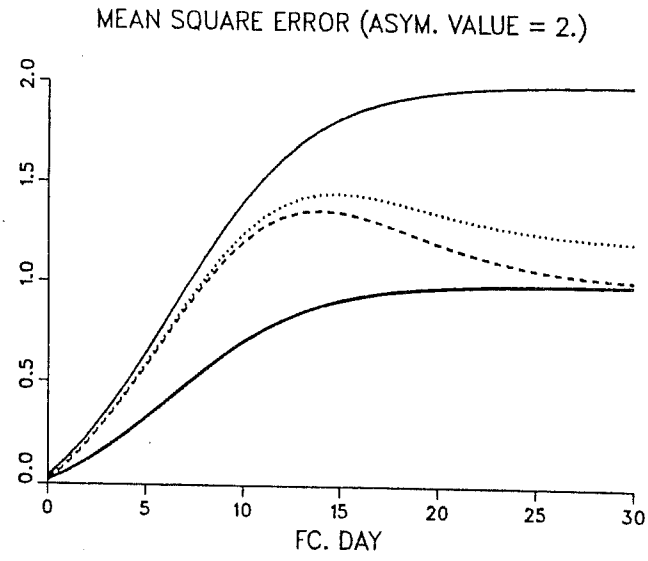
Dalcher and Kalnay (1987), based on earlier work of Leith (1978), have shown that the growth of error and spread of deterministic numerical forecasts can be parametrized by the following equations:

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

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

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 effect of the model approximations. Using ECMWF forecasts of 500 mb heights for winter 1980/81 and setting $V_\delta^* = V_e^*$, they obtained α=0.39 day^{-1} and σ = 0.045 day^{-1}.

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 at short forecast times. Furthermore, it is common experience in numerical weather prediction 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. 2 shows the theoretical time evolution of e^2, $|\bar{E}|^2$, $\bar{\rho}$



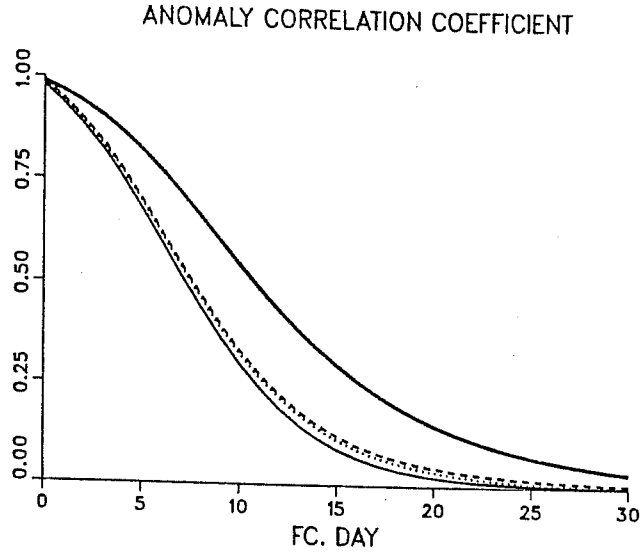


Fig. 2 Theoretical error growth rates when asymptotic values of variances are set to 2. See text for explanation.

and $\rho(\bar{F})$ computed using Eqs. (11a,b) and setting the values of parameters as follows

$$a = f = 1.$$
$$V_e^* = V_\delta^* = 2.$$
$$\alpha = 0.25 \text{ day}^{-1}$$
$$\sigma = 0.035 \text{ day}^{-1}.$$

The first four parameters are normalized by setting the observed variance to 1; the growth rates are set to lower values than those found in Dalcher and Kalnay in order to have a better fit to the error growth of the T63 model used in this study (i.e. to take into account improvements in the ECMWF model). Comparing the full lines (e^2 and $\bar{\rho}$) with the dashed lines ($|\bar{E}|^2$ and $\rho(\bar{F})$) one can see that the improvement due to ensemble forecasting is considerably less than the perfect-model limit (heavy solid line) in the first 10 days of the forecast. This theoretical limit is only approached after ~ 20 days, due to the slower saturation of the spread, which causes a slower decrease of the ensemble mean error (according to Eq (4)) in the last part of the forecast. However, looking at the anomaly correlation coefficient (lower panel), the curve for $\rho(\bar{F})$ remains monotonically decreasing with the modest improvement over $\bar{\rho}$.

The situation becomes even less favourable for the ensemble forecasting if one takes into account that the assumption $V_\delta^* = V_e^*$ is suitable for a model with random but no systematic errors. If a systematic error exists, then V_δ^* will be lower than V_e^* by an amount which is proportional to the variance explained by systematic error itself. Assuming again $V_e^* = 2.$, but $V_\delta^* = 1.6$ (corresponding to a normalized variance of the systematic error equal to 0.2; see Appendix in Tibaldi et al., 1987), we obtain for $|\bar{E}|^2$ and $\rho(\bar{F})$ the dotted curves shown in Fig. 2.

As previously said, all the curves in Fig. 2 have been computed assuming that the asymptotic value of the forecast error is twice the climatological variance. In fact, the recent experience in extended range forecasting indicates that, on the monthly time scale, some predictability may exist even at day ~ 30, due to the thermal interaction with the lower boundary. This predictability is evident in cases of strong SST anomaly, but even in

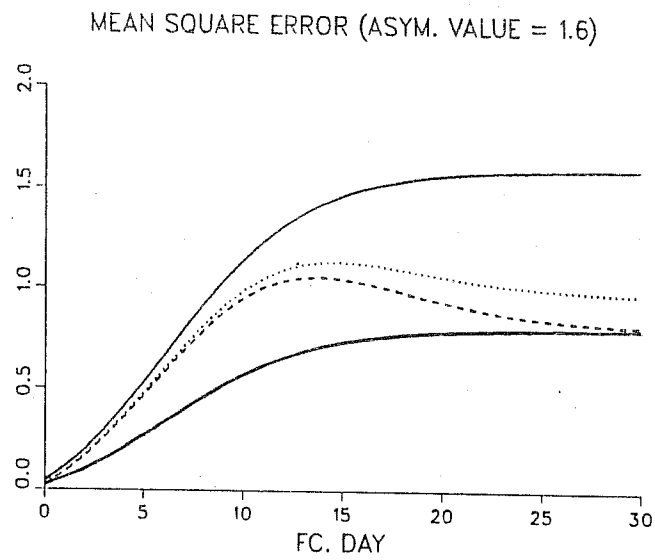
'near-average' situations it cannot be totally neglected and is probably responsible for the relatively high skill of extended range prediction of temperature in the lower troposphere, especially in the southern hemisphere (see Molteni et al., 1986; Tibaldi et al., 1987).

If one assumes that, due to this interaction, 20% of the variance remains predictable at longer range, and reduces the values of V_e^* and V_δ^* accordingly, one obtains the curves shown in Fig. 3. One can see that the most relevant effect of this assumption on the ensemble forecast skill scores is to enhance the improvement in terms of ACC. A return of skill can now be seen also in the ACC, which is much more evident in the case of no systematic error (dashed curve), confirming the results of Molteni et al. (1986). This is due to the fact that while $\bar{\rho}$ is practically constant from day 15 onwards (and greater than zero), $|\bar{F}|$ continues to decrease and from Eq (9), $\rho(\bar{F})$ increases in the second half of the forecast range.

In conclusion, the slower growth of the spread of an ensemble compared with the error growth, restricts the improvement in skill of the ensemble mean forecasts over that of individual members of the ensemble, and gives rise to a 'return of skill' apparent in both RMS error and ACC if the thermal interaction with the surface explains a non-negligible proportion of the observed and forecast variance.

In general V_e and V_δ are function of annual cycle, and, within a season, of weather regime. Their dependence on 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, 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_e and V_δ could be quite uncorrelated for models with serious systematic error.

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



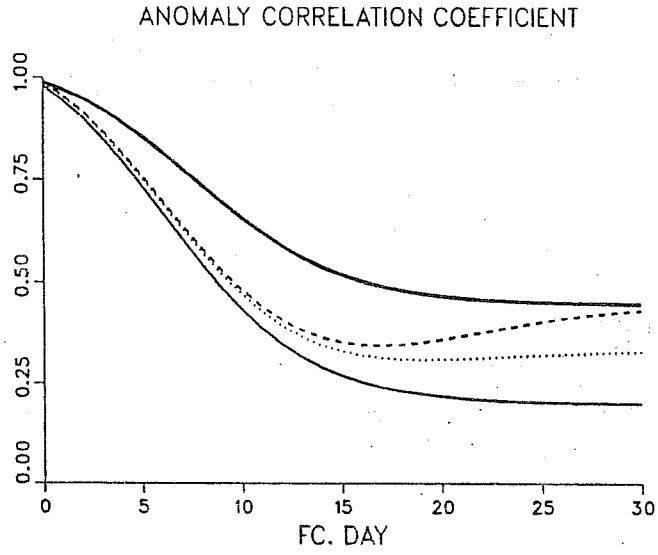


Fig. 3 Same as Fig. 2 but when asymptotic values of variances are set to 1.6.

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 the expected atmospheric states.

4. A SUMMARY OF RESULTS

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

4.1 Skill of ensemble forecasts

We first address the question of whether the skill of the ensemble-mean forecast has increased over the skill of individual forecasts. In Fig 4a, b, we show scatter diagrams of the 30-day mean northern hemisphere 500mb height RMS error and anomaly correlation coefficient for the ensemble-mean forecast against the mean skill of the individual forecasts. As discussed in section 3, the ensemble-mean RMS error is inevitably smaller than the mean RMS error of individual forecasts. This is clearly illustrated in Fig 4a, 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 of the ensemble mean RMS errors, with summer forecasts having smallest RMS errors, winter forecasts having largest errors.

From Fig 4a, 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; by combining equations 5 and 6, for large N, we have $|e-E|=(\sqrt{2}-1)|\Delta|$, i.e. the improvement due to ensemble averaging is proportional to the dispersion Δ from the ensemble mean. We would expect this dispersion to be strongly influenced by the annual cycle, and consequently to contribute to a larger improvement to ensemble forecasting in the transition season than in the summer season. The largest impact of ensemble averaging is

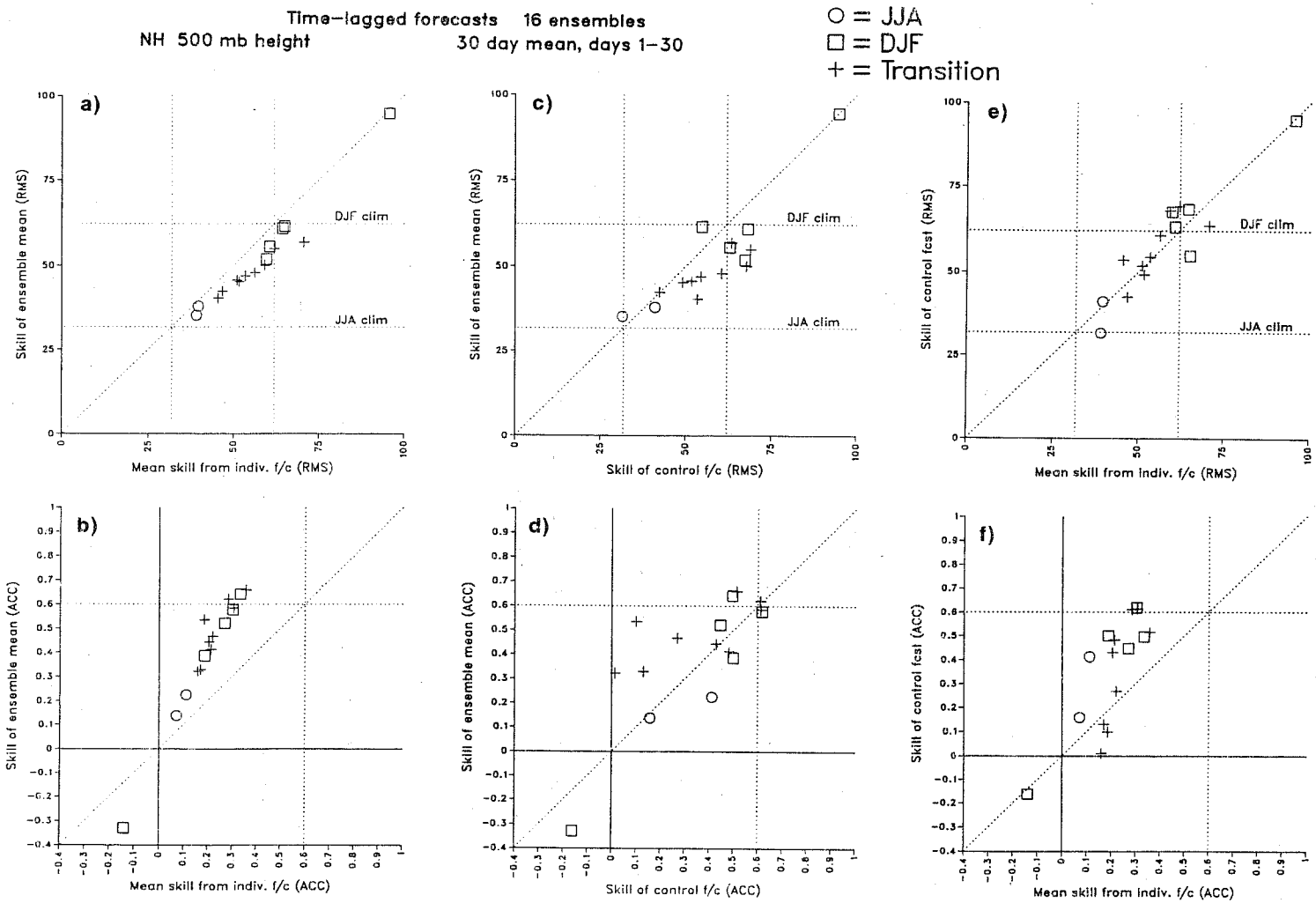


Fig. 4 30-day mean RMS error (top) and anomaly correlation coefficient (bottom) of the northern hemisphere 500 mb heights for: a) and b) ensemble mean forecast vs. mean of all individual forecasts, c) and d) ensemble mean forecast vs. control forecast, e) and f) control forecast vs. mean of all individual forecasts. RMS error in metres.

associated with the ensemble from November 1985 (part of the transition period), when RMS error fell from a mean value of 71 m to an ensemble average value of 57 m (implying Δ=42 m).

However, it can also be seen from Fig 4a, that the impact of the seasonal cycle on reduction of forecast error does not extend into the winter season. The smallest impact of ensemble averaging does not, as one might have expected, 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 mean value of 96 m to an ensemble mean value of 95 m (implying Δ=14 m). This suggests that the perfect model assumptions seriously break down for the winter ensemble forecasts. As can be envisaged already, individual members of the ensemble from January 1986 failed to forecast the development of a major large scale anomaly, and moreover were each remarkably consistent with one another.

As discussed in section 3, the improvement to the anomaly correlation coefficient is not inevitable. However, 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 effect can be seen in Fig. 4b. As in Fig 4a, 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 our expectation, ensemble averaging has made the anomaly correlation coefficient even more negative. It is interesting to note that the annual cycle distinguishes less clearly, than for RMS error, the skill of the ensembles from different seasons.

Of the winter and transition ensembles, exactly half have 30 day mean anomaly correlation coefficient in excess of 0.5. (A value of 0.5 has some objective significance, since, if the forecast and observed fields have equal spatial variance, an anomaly correlation coefficient of 0.5 corresponds to a climatological RMS error.) 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 4b, than in Fig 4a. In this sense, it would appear that summer ensembles have rather low skill in the prediction of phase compared with other times of year.

It could be argued that a more stringent test on the effect of ensemble averaging would be to compare the skill of the ensemble mean forecast with the skill of the latest member of the ensemble, the control forecast. Scatter plots of this, for RMS error and anomaly correlation coefficient, are shown in Fig 4c and d respectively. The improvement in skill through ensemble averaging is now not inevitable, and for RMS error, one of the winter control forecasts (December 1985) and one of the summer control forecasts (June 1986) are more skilful than the ensemble mean. For anomaly correlation, there are cases from all times of year when the control forecast is superior to the ensemble mean. However, there are more cases where ensemble averaging improves skill, and there are no cases of substantial worsening.

In view of these results it is of interest to ascertain whether the cases of higher control-versus-ensemble skill in Fig 4c,d are merely associated with random sampling, or whether the control forecast is systematically more skilful than the mean skill of individual forecasts. Scatter plots to determine this are shown in Fig 4e,f. For RMS error, there is no particular evidence to favour the control forecast; the points are distributed reasonably uniformly about the diagonal line. However, for anomaly correlation coefficients, there is evidence that the control forecast is favoured, particularly for winter and summer cases. Clearly, with such a small sample, one must exercise caution, but this result lends further support to the notion outlined in the introduction, that with the time-lagged technique, the individual members of the ensemble are not equally likely, even for the extended range, at least as far as maximisation of anomaly correlation coefficient is concerned. (For the anomaly correlation coefficient, this improvement is very clear in the first 10-day mean, less clear in later 10-day means. For RMS error the improvent of the control in the first 10-day mean can be just discerned. For later 10-day means it cannot be detected.)

Note from Fig 4a, four of the five winter ensemble-mean forecasts have lower RMS errors than the average skill of a wintertime climate forecast (shown as the dashed horizontal line marked DJF climatology). However, it could be

argued that a stricter comparison can be obtained when the 30-day mean scores are compared with the 30-day RMS error of a climate forecast appropriate to the relevant forecast period. From section 3, the RMS error of a climate forecast is equal to the amplitude of the observed anomaly. A scatter plot of ensemble mean RMS error against amplitude of observed anomaly is shown in Fig 5a. There are only two winter ensembles (December 1986 and December 1987), and two transitional ensembles (November 1985 and September 1986) which can be said to be better than a climatological forecast. However, these results should perhaps be tempered by two considerations. Firstly, when the magnitude of the observed anomaly is small, it is clearly unlikely that any model could improve over a climatological forecast. The skill of a numerical model should be gauged in circumstances when the flow is anomalous. In this respect, it can be said that the model outperforms climatology for the four most anomalous situations. Secondly, for other model fields, a more substantial proportion of forecasts do better than climatology. For example, for 850mb temperature (see Fig 5b) all but one of the winter forecasts is superior to climatology (the case from January 1986 being inferior).

Another objective comparison is shown in Fig. 5c, d, where the 30-day mean skill of ensemble mean forecast is compared against the 30-day persistence forecast. Here we use persistence of anomaly fields rather than persistence of full fields, in order to remove the seasonal trend and thus to optimize the persistence forecast. One can note that the persistence RMS errors lie approximately within the bounds defined by the perfect model assumption, that is $\sqrt{2}$ times the climate norm (Hollingsworth et al., 1987). Clearly, in all but one case (January 1986) the model outperforms the persistence forecast. This is also true in the anomaly correlations, though the seasonal cycle is now less evident.

The general regional improvement of ensemble mean forecasts over the control forecasts can be seen in Fig 6a-c. It is interesting to note that the effect of the seasonal cycle is much more apparent for the Atlantic sector than the Asian sector. This is consistent with the observed ratio of low-frequency variability in the winter and summer, large over the Atlantic, smaller over Asia (see, for example, Lau et al., 1981). In the southern hemisphere (Fig 6d), there is even less distinction between the summer, winter and transition ensembles, again consistent with the relative weakness of the annual cycle on the general circulation of the southern hemisphere.

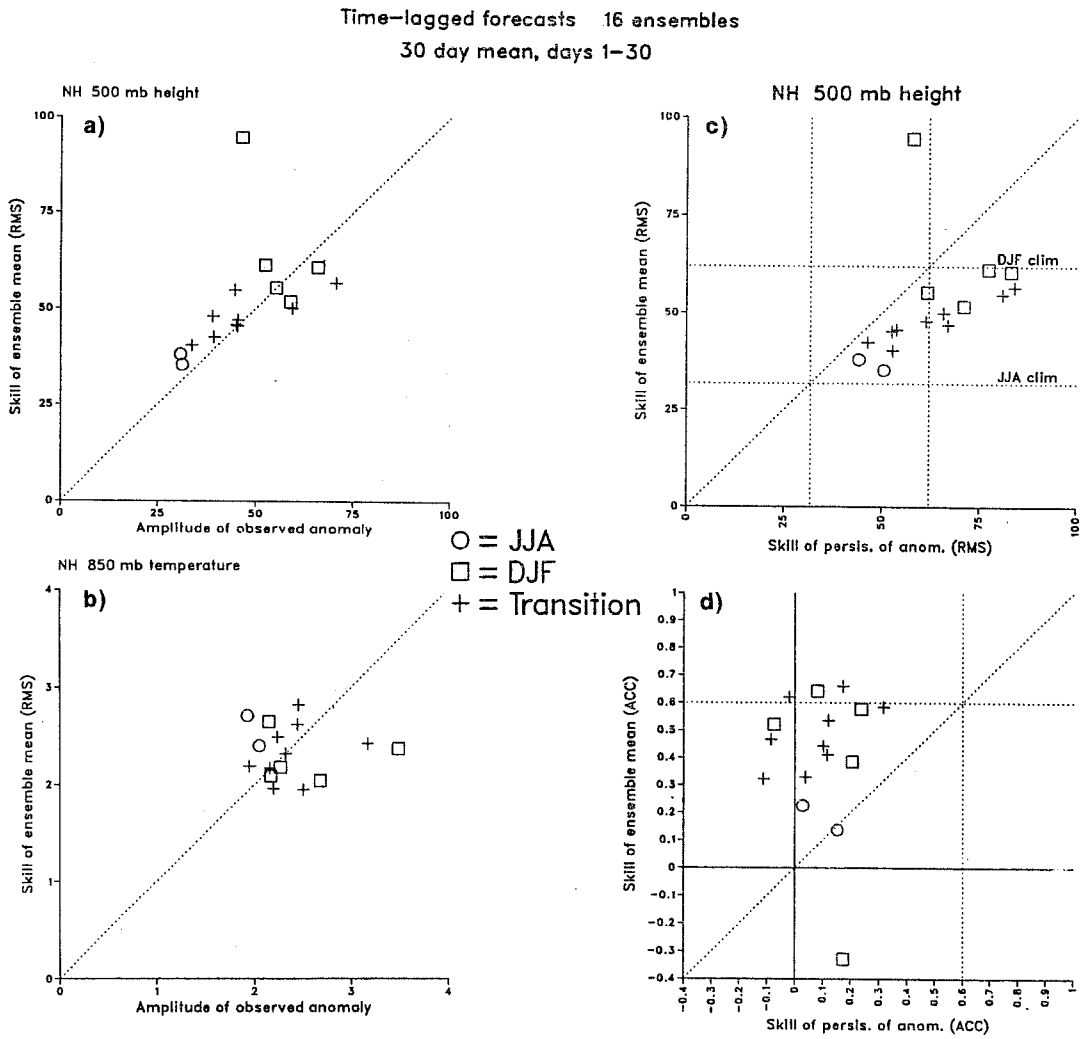


Fig. 5 30-day mean RMS error for ensemble mean forecast vs. amplitude of observed anomaly of: a) 500 mb height (m), b) 850 mb temperature (deg K); 30-day mean skill of the 500 mb heights for ensemble mean forecast vs. persistence of anomalies: c) RMS error (m), d) anomaly correlation coefficient.

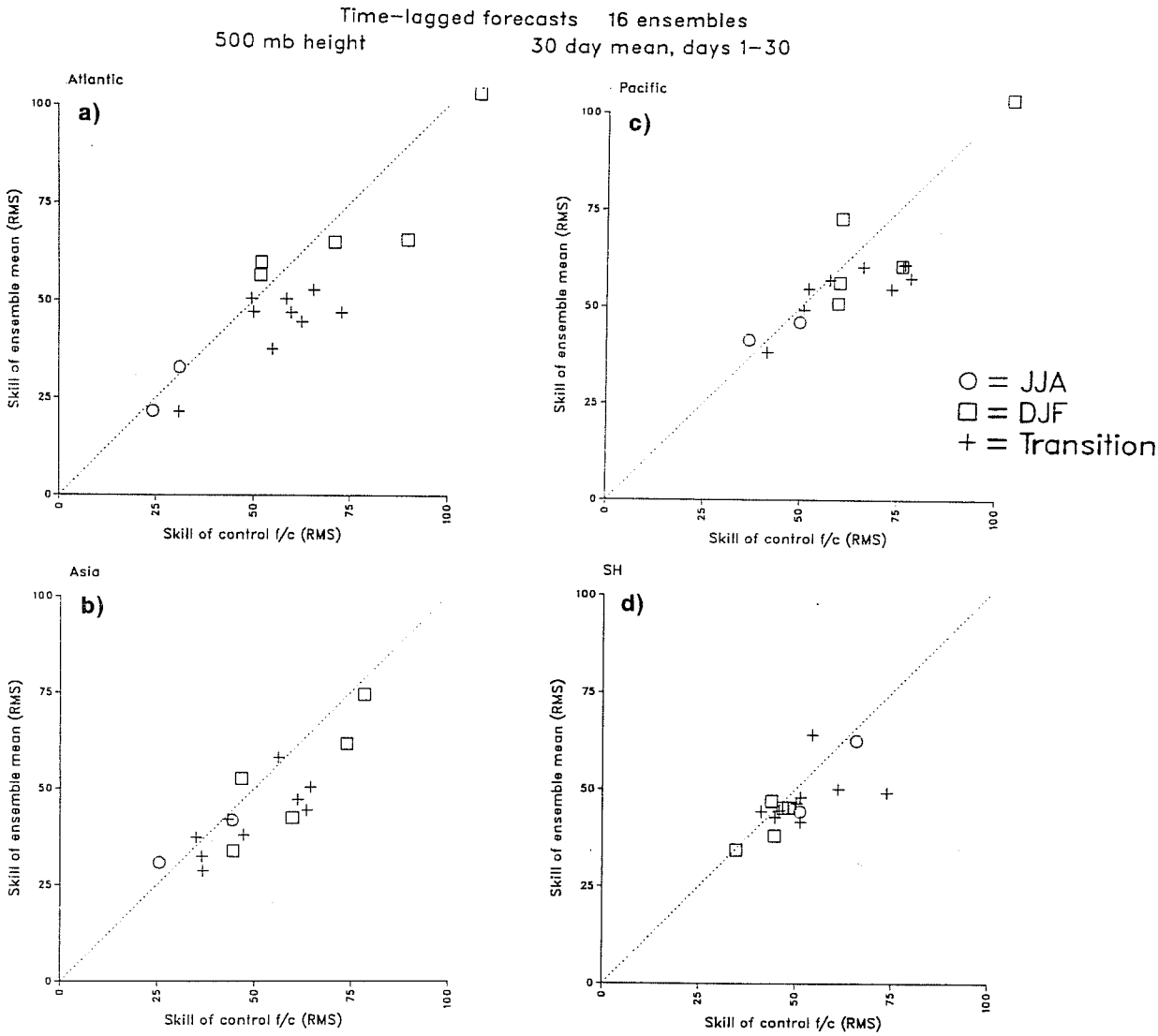


Fig. 6 30-day mean RMS error (m) of the 500 mb heights for the ensemble mean forecast vs. control forecast over: a) Atlantic, b) Asia, c) Pacific and d) southern hemisphere.

4.2 Relationship between spread and skill

The notion that the spread and skill of the ensemble forecasts are positively correlated is central to the usefulness of the ensemble technique. Under perfect model assumptions, as discussed in section 3, $\Delta=E$ for large N. The extent to which this is achieved in practice can be seen in the scatter diagram in Fig 7. In fact, far from being equal, it can be seen that, on average, the mean square forecast error is about 3 times the mean square ensemble spread. For the exceptional case of January 1986, the mean square error is about 9 times the ensemble spread. This general inequality is suggestive of systematic deficiencies in the model, particularly in wintertime. This will be discussed below.

Overall, it can be seen that there is a clear positive correlation between spread and skill. However, a substantial part of this correlation reflects the impact of the seasonal cycle on spread and skill. Taking the summer and transition seasons together, spread and skill are approximately equal, subject to an offset of about 20 m. The offset can be thought of as an effect of the model systematic error, that is to say, there is an 'external' source of error of about 20 m which is largely independent of weather regime.

Within the winter season, on the other hand, there appears to be an anticorrelation between (RMS) spread and skill. This demonstrates quite clearly, that in a nonperfect model environment spread is not necessarily a good predictor of skill. This anticorrelation is probably also related to systematic error. In wintertime, it is well known that model systematic deficiencies simultaneously affect the drift both of the mean state, and of low-frequency variability about this mean state (Palmer, 1987). If the climate drift of the model has the effect of producing an erroneous and deep 'attractor basin' in some portion of the phase space of model states, then trajectories of a forecast ensemble falling into this basin will be consistent and invariably wrong; whilst trajectories of an ensemble avoiding this basin will be less consistent though at least occasionally correct. Clearly the latter type of systematic error is much less 'benign' than the 20 m 'offset' (which could be thought of as a mere translation of the attractor basins in some direction of phase space, without affecting their relative structure) required for equality of spread and skill during other times of year.

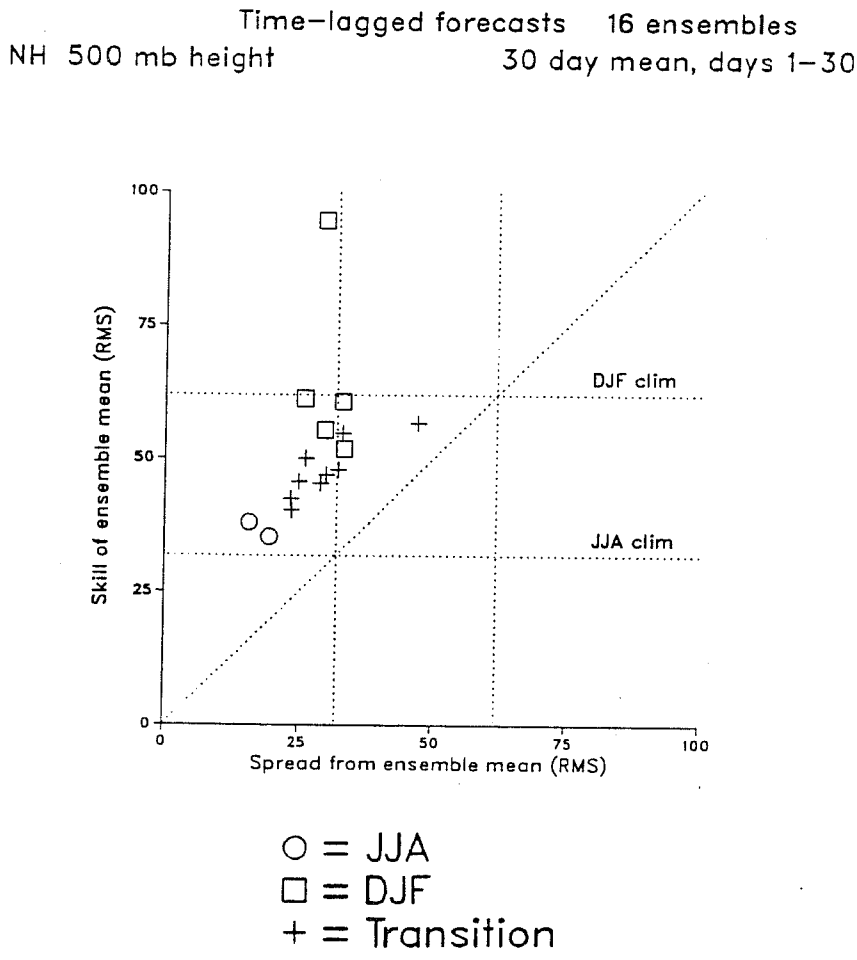


Fig. 7 30-day mean RMS error vs. spread (m) of the northern hemisphere 500 mb height.

Considering cases within a single season only, the only evidence of a nontrivial positive correlation between spread and skill appears to come during the transition period. Of course it is still possible that there is a residual effect of the seasonal cycle within the transition sample. Indeed the ensembles with largest and smallest error start from dates within one month of the winter and summer solstice respectively. For the others, the error is not well correlated with starting date. Moreover, for the transition season, there appears to be a correlation between spread and skill using the anomaly correlation coefficient (not shown), not evident for the other seasons.

4.3 Time evolution of the ensemble skill scores

We now briefly discuss the evolution of skill of the ensemble forecasts during the course of the integration. For clarity, all ensembles are separated into 'extended winter' (October to March) and ' extended summer' (April to September) seasons.

In Fig 8, the daily scores show large variability. The poorness of the January 1986 forecast is confirmed in both the RMS error and anomaly correlation coefficient. The ensemble-mean score crosses the 0.6 line by day 5, and it continues to fall rapidly, reaching zero anomaly correlation by day 10. The ensemble forecast for the next month, February 1986, is clearly the most skilful in terms of anomaly correlation coefficient, which does not fall below 0.6 until day 14 and stays relatively high at the end of the forecast.

In summer, daily anomaly correlation scores cross the 0.6 line, on average, earlier than in winter and drop towards zero correlation faster,staying quite low until the end of the forecast. The RMS error is lower, but saturation is reached earlier than in winter. Dispersion between ensembles is smaller and is relatively uniform throughout the forecast time.

The potential improvement that time averaging makes to the ensemble forecast is shown in Fig 9. The forecasts have been grouped into two sets: an extended winter (October to March) and an extended summer (April to September). Since RMS is trivially reduced through time averaging, we show in Fig 9 only the impact on anomaly correlation coefficient. If the predictability of the

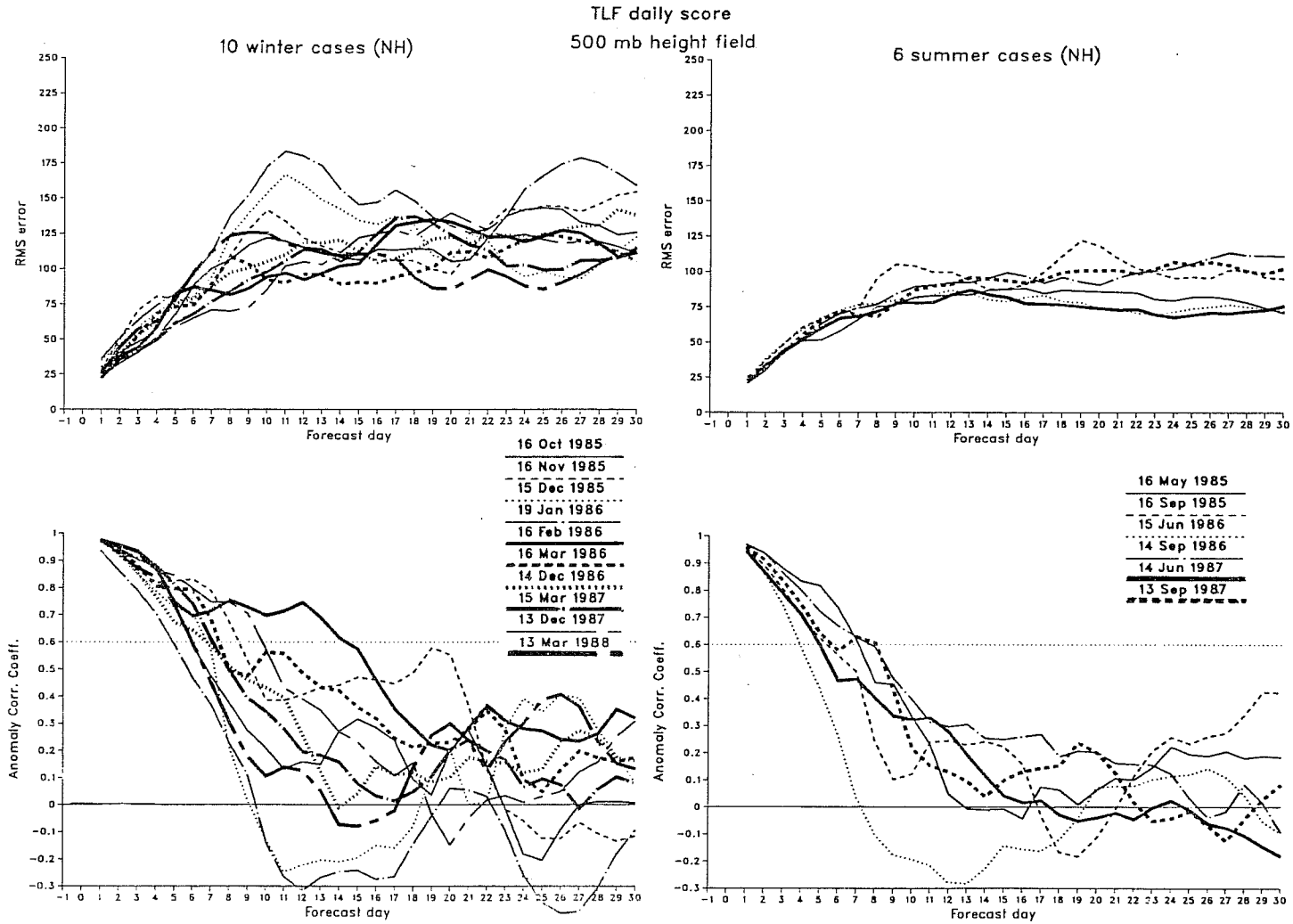


Fig. 8 Time evolution of ensemble-mean daily skill for the northern hemisphere 500 mb heights in 'extended' winter (left) and 'extended' summer (right). Top: RMS error (m); bottom: anomaly correlation coefficient.

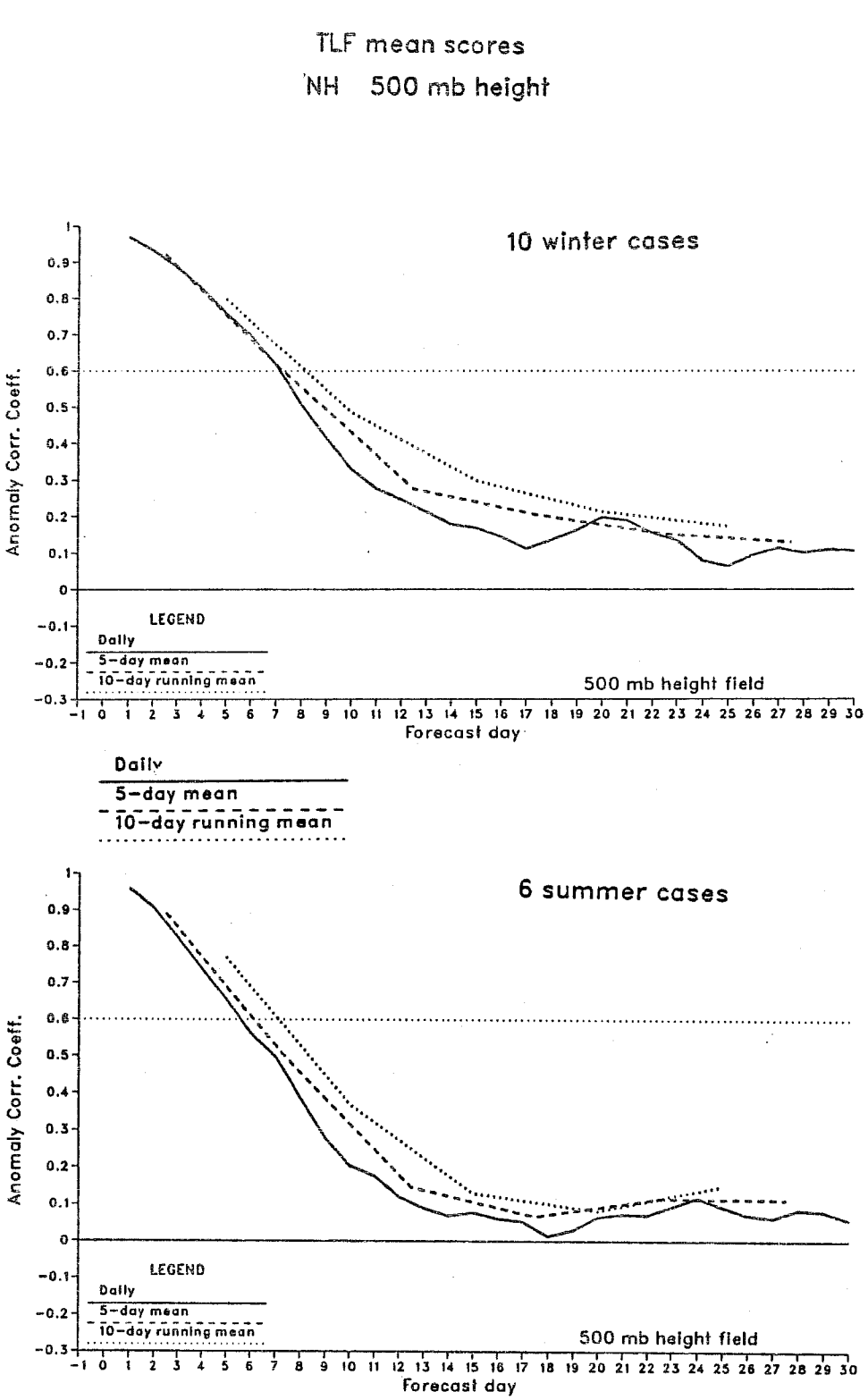


Fig. 9 Time evolution of mean daily, 5-day mean and 10-day running mean anomaly correlation coefficient for 'extended' winter (top) and 'extended' summer (bottom).

ensemble mean forecast was the same for all time scales, the time-mean operation should merely result in a smoothing of the daily skill scores. In fact, as Fig 9 shows, there are some improvements for both 5 and 10 day mean forecasts compared with daily values, albeit rather modest ones. The largest improvement occurs around day 10-15, and 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 10-15 could be very significant. At the present, the upper limit for medium-range forecasting is normally considered to be about 10 days. Having in mind steady improvements in numerical models and observing/analysis systems, the improvement in skill shown in Fig. 9 give rise to the hope that the upper limit for medium-range forecasts might extend beyond day 10 in future years, at least for time averaged fields.

5. PROBABILISTIC APPROACH TO ENSEMBLE FORECASTING

A simple approach to extracting additional useful information from individual forecasts of an ensemble is by defining probabilities that a forecast variable falls within different predefined categories. Though the probabilistic approach is well established in statistical long range forecasting in some meteorological centres (for example in the UKMO and NMC), an application of this method to numerical ensemble forecasting has only been made recently (Déqué, 1988). As mentioned in the introduction, spread from the ensemble mean forecast can offer only an estimate on how well that mean forecast represents the elements of the ensemble, giving no indication on possible alternatives. With the probabilistic approach a predicted extreme event will be captured even when only one single member of the ensemble gives such a prediction, no matter how small probability might be. This would be of practical value to the forecaster.

To demonstrate the potential of the probabilistic method, in Fig. 10 we look at predicted 5-day averages of 850 mb temperature anomalies over the European region from December 1986 ensemble individual forecasts. 850 mb temperature is used, firstly because of its practical importance, and secondly, as discussed in section 4, many of the ensemble-mean forecasts were particularly skilful (relative to a climate forecast) in this variable compared with, say, 500 mb height. We define three classes for anomalies: above normal, normal and below normal. The limits of the classes are set beforehand to +2 and -2 K

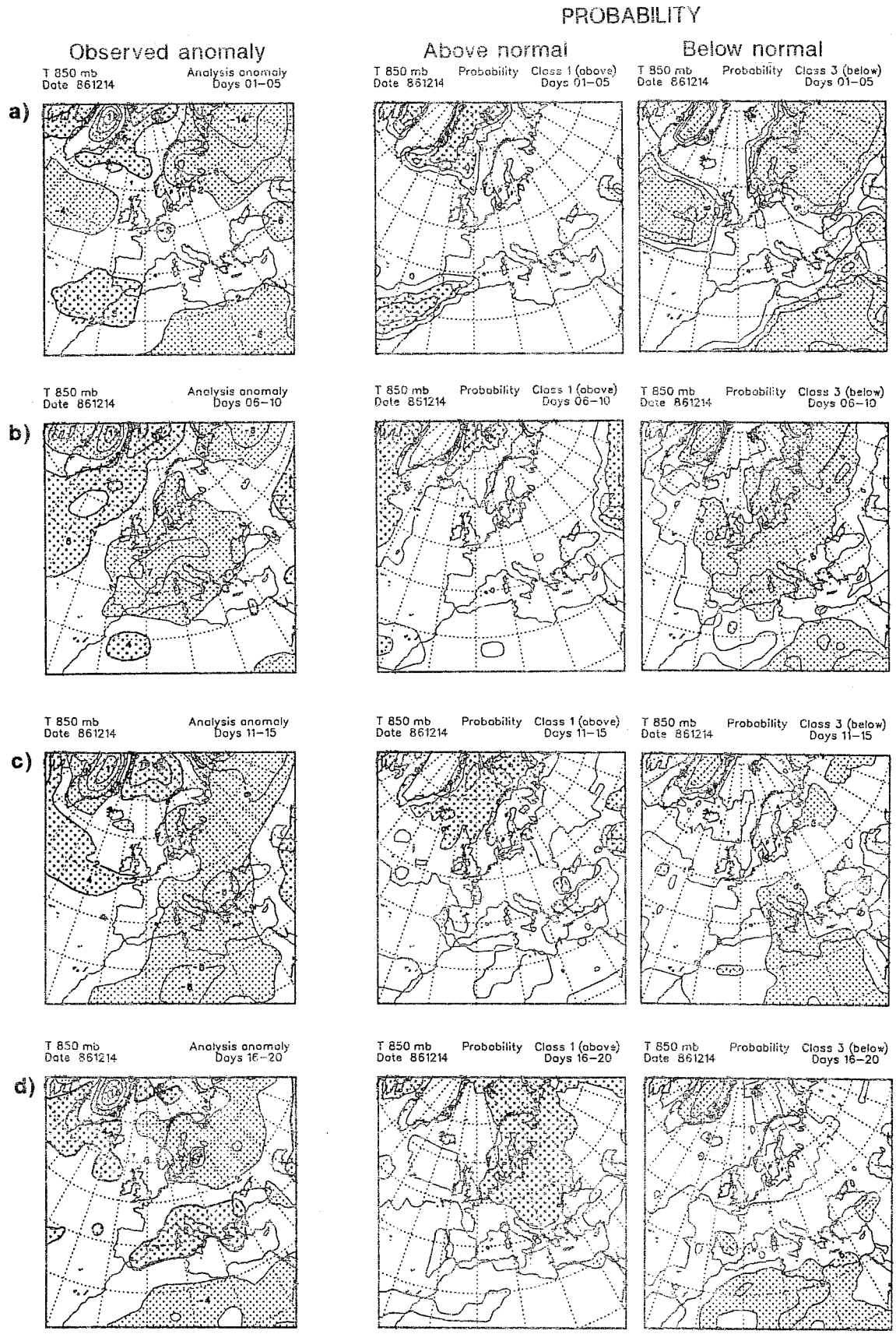


Fig. 10 Observed 5-day mean 850 mb temperature anomalies (left; in deg K) and probabilities for the 'above normal' (middle) and 'below normal' (right) classes for the first four pentads of the December 1986 ensemble.

respectively. Of course, these limits should in principle be defined more objectively, thus to avoid that various climatic zones are treated identically. One possible way of defining class limits might be from the climate variances for each considered grid-point. The probability that a certain temperature anomaly falls into a prescribed class is then computed by simply counting the number of occurrences within a given ensemble. If at a given grid-point each individual forecast lies within the same class, the probability is 1.

Fig. 10a shows the observed 850 mb temperature anomaly and probabilities for the 'above normal' and 'below normal' classes for the period corresponding to the first 5 forecast days of the December 1986 ensemble. The observed anomaly over the most of Europe was between -2 and +2 K, i.e. normal according to our definition. A large negative (cold) anomaly is found in the north-eastern part of the continent, whilst a somewhat less intense cold anomaly is seen over northern Africa and to the west of the British Isles. The difference in amplitude between cold anomalies in the northern Europe and northern Africa clearly illustrates the need to define limits of classes from climatology, because 'cold' in Africa is 'normal' in the north and vice versa. The probability for these cold anomalies is .9 or more suggesting that most or all members of the ensemble predicted cold spells. Another indicator of consistency in predictions of individual forecasts is a relative closeness of probability contours.

In the next 5 days (Fig. 10b) the cold anomaly is found in the two areas over Europe stretching in the SW-NE direction. It is well captured in the 'below normal' class with probabilities above .5. The probability for a cold anomaly higher than .9 is now found only in the northern Soviet Union. Also worthy of note is a high probability for a cold anomaly in the Sahara, which is shown as a shaded area in the map of analysed anomalies. The North Atlantic warm anomaly is reasonably well predicted in the Arctic region and to the south of Greenland. The .1 contour between Europe and Greenland indicates that this warm anomaly is captured by some individual members but not all. The probability contours have now become less densely packed, thus pointing to more dispersion of the forecasts within the ensemble.

In days 11-15 the cold anomaly is extending in the north-south direction over the central and eastern Europe and over north Africa. The maximum of this

anomaly over the south-east Europe is predicted by the .5 probability or more and so is the maximum over the Sahara. Now even larger dispersion between forecasts is seen as the .1 contour encompasses almost the whole continent in the map of the 'above normal' probabilities. A higher probability of cold anomaly may have decisive preponderance in any decision making; however, because of systematic errors high probability does not necessarily imply a correct forecast. This is illustrated for a small area of cold anomaly off the coast of Iceland which is wrongly predicted as a warm anomaly with probability of more than .5. Such an ambiguity cannot be avoided by the probabilistic approach.

The observed cold anomaly maximum over Europe in Fig. 10c was in the next pentad (days 16-20) replaced by warm anomaly and a complex pattern developed at the time corresponding to a well advanced forecast (Fig. 10d). Nevertheless, a high probability for the cold anomaly over the north of Africa is quite realistic. On the contrary, a relatively high probability for warm anomaly over the northern Europe suggests that many individual forecast predicted a wrong development in that particular region.

The probabilistic approach to ensemble forecasting shown in the above example is still rather crude; however, its practical usefulness is undoubtful. The maps of probabilities should be combined with other products of the time-lagged ensemble technique, such as the mean forecast and standard deviation from the mean. The skill of such a probabilistic approach could be verified by the ranked probability scoring method (see for example Palmer and Molteni, 1987; Déqué, 1988).

6. CLUSTER ANALYSIS OF ENSEMBLES

A cluster analysis of the large scale flow patterns has mainly been employed for diagnosing the observed general circulation (e.g. Molteni et al., 1988). Murphy and Palmer (1986) performed a simple cluster analysis on an ensemble of 7 extended range forecasts, based on a subjective assessment of the large scale flow pattern. Grouping of individual members of an ensemble into clusters may help to identify those forecasts which all predict a certain type of development.

Various meteorological parameters and various statistical properties of these parameters may serve as the basis for cluster analysis. Clusters obtained

from, say, 500 mb heights may differ from those obtained from, say, MSL pressure. However, we shall assume that time averaging of geopotential height fields is sufficient to make dominant the barotropic component of the flow and we select 5-day means of the 500 mb heights as the basis for our clustering. The next problem relates to the choice of time (or period) of the forecast we want to apply cluster analysis for. In the perfect model assumption grouping of forecasts indicate instability of initial data. In the case when the model is not free from systematic biases, the grouping may be different, in terms of both the size of clusters and distances, at one forecast time compared to grouping at some other forecast time, even if we look at the same meteorological parameter.

A convenient way to describe clustering within an ensemble is by plotting trajectories of some (statistical) property of individual forecasts in the phase space. Fig. 11 shows trajectories for the first two EOFs of the northern hemisphere 500 mb height for December 1986. The first two EOFs are dominant since they explain 40.5% of the total variance. The starting points of the arrows correspond to the mean EOF values in days 1-5, and the arrow heads correspond to subsequent 5-day averages. Thus the arrows describe time evolution from the first pentad onwards. The figures next to the arrow heads designate individual forecasts as numbered in Fig. 1, and zero and open arrow stand for the verifying analysis.

In the second pentad (days 6-10) all forecasts behave similarly and no clustering could be seen. The analysis displays a somewhat different phase shift compared to the forecasts. In the third pentad (days 11-15) there is a hint of a weak clustering: forecasts 1, 6 and 7 turn into other direction at a larger angle than the rest of the ensemble. However, more impressive is the phase reversing of the verifying analysis, unlike in any forecast. In the next pentad (days 16-20) three apparent clusters are found: A(1,6,7), B(4,5,9) and C(2,3,8). This is an interesting stage in the development because in each cluster the two subsequent forecasts are accompanied by one member from the other end of the initial time sequence. The centroids of the clusters belong to a different quadrants of the phase space. When superimposed to the mean state, derived from all individual members, the EOF coefficients of clusters from three different quadrants will yield three different fields. In days 21-25 we note a tendency of the cluster B to approach cluster C and form a new

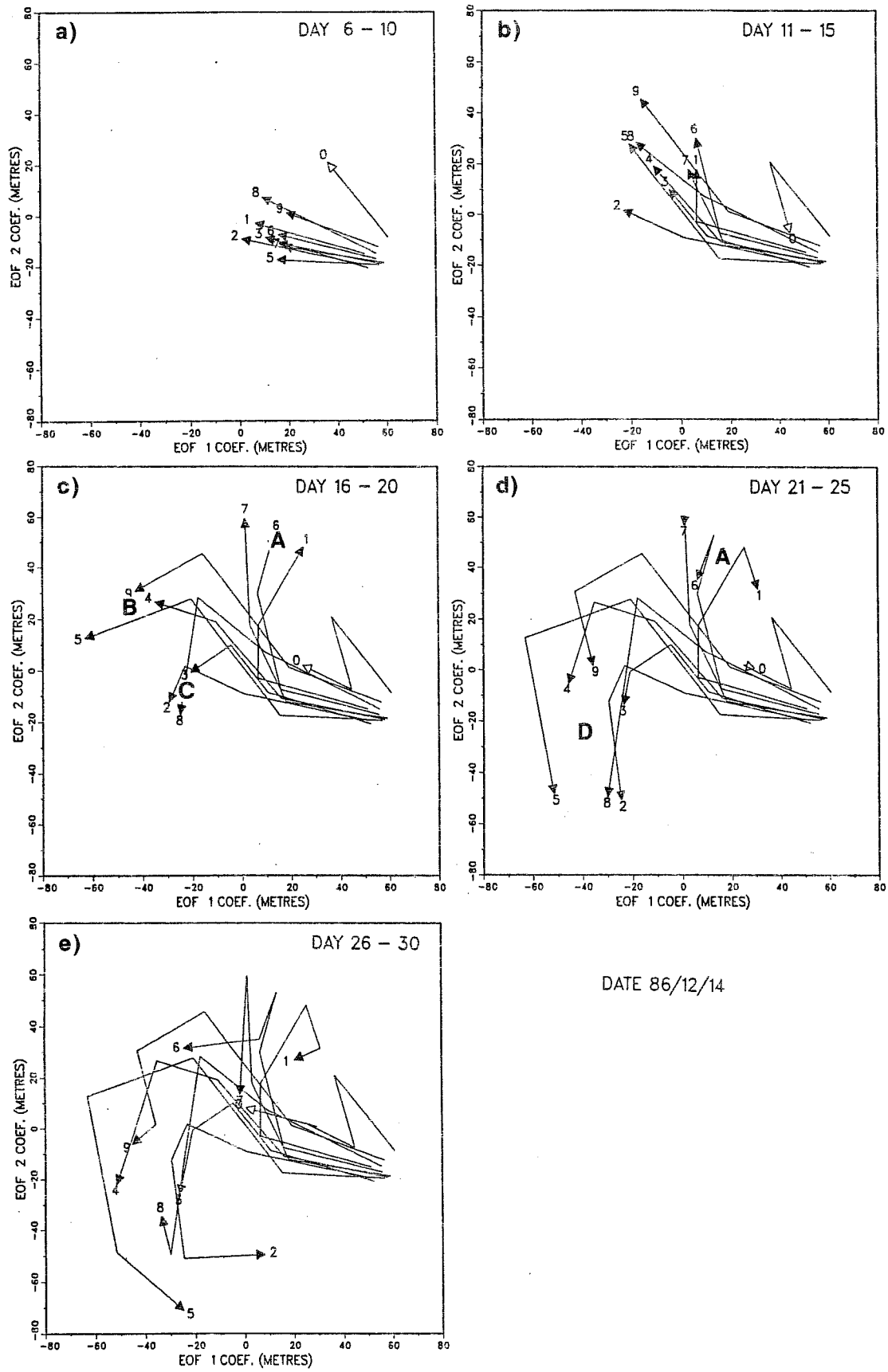


Fig. 11 Trajectories of the 5-day average northern hemisphere 500 mb heights in the EOF 1/EOF 2 plane of the phase space for December 1986 ensemble. 1 to 9 denote individual forecasts as depicted in Fig.1, zero and open arrow denote verifying analysis.

cluster D, though one may argue that three clusters can still be recognized but with a different composition compared to the previous pentad. At the end of the forecast period (Fig. 11e) we can identify two relatively less distinct clusters: one cluster includes forecasts (1,6,7), i.e. those members who differ from the others from the early time of the forecast and the other one composed from the rest of the ensemble. Though the forecast no. 7 is in the end the closest to the verifying analysis, its development was quite different.

It is clear from the above example that in the plane of the phase space defined by EOF 1 and EOF 2 members of the cluster A(1,6,7) are showing distinctly different trajectory when compared with other individual forecasts. Though generally this may not be true for the coefficients of some higher EOFs, cluster A(1,6,7) can still be well distinguished even when EOF 3 is combined with the first two EOFs (not shown), however the improvement over the cluster D is then less obvious. Closer inspection of Fig. 11 c-e reveals that forecasts from cluster A have the shortest distances from the verifying analysis in the direction of EOF 1 but not very much in the direction of EOF 2. When EOF 3 is included this advantage of cluster A seems to vanish completely. It should be pointed out that, since the EOFs have been computed from the forecast fields only, there is no guarantee that most of the spatial variance of the observed anomaly is also explained by the first two or three EOFs. Usually, higher-order EOFs represent higher-frequency modes of variability; therefore, the forecast error may have strong projections on the EOFs. However, the distance between analysis and forecast in the EOF 1-EOF 2 plane is likely to be a good approximation of the forecast RMS error when time averages over 15 or 30 days are considered.

If we look at the skill scores of clusters A and D (Fig. 12) we note that the 5-day mean anomaly correlation is slightly lower for the cluster A when compared with the cluster D or with the complete ensemble. This is because contributions from all scales are included when calculating the skill scores; cluster A includes a smaller number of forecasts, so there is less chance that contribution from the smaller scales will cancel each other. However, 30-day mean anomaly correlation is higher for the cluster A than for cluster D and for the complete ensemble, the reason being that time averaging retains only the large scale flow pattern whose variance is well described by the first two EOFs.

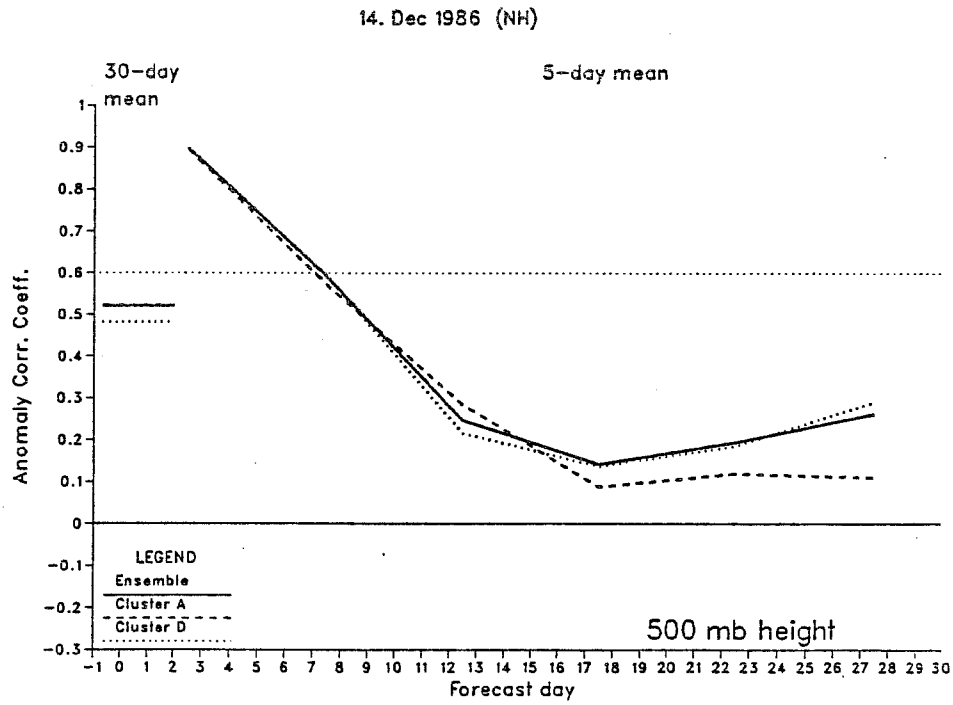


Fig. 12 Anomaly correlation of the northern hemisphere 500 mb height 5-day and 30-day (left) averages for the full December 1986 ensemble and clusters A and D described in the text and Fig.11.

Of more practical importance than skill scores is the predicted pattern of the large scale flow in the two clusters, shown in Fig. 13. Though neither of the two clusters nor the full ensemble predicted correct development, it is clear that cluster A displays the most structure, whilst cluster D has a rather smooth field very much alike to the field described by the full ensemble. This is most obvious in the polar regions where the observed pattern looks fairly complex. In the cluster A the Canadian and north Siberian lows are better captured and a stronger ridge north of Scandinavia can be seen. The anomaly pattern in cluster D is more characteristic for the model systematic error (c.f. Tibaldi et al., 1988). This suggests that the forecasts defining cluster A have fallen into a different circulation regime than the rest of the forecasts from this ensemble in the second half of the integration period. The above result, though still lacking sufficient accuracy, indicates the potential in using this clustering technique on the ensemble forecasting.

Fig. 14 displays trajectories up to days 16-20 for January and February 1986 cases. The January forecast had very poor but very consistent skill. However, the EOF coefficients show much less of that consistency in Fig. 14a, because some forecast have moved faster along the phase space than the others. There is no clear clustering and the forecasts remain scattered in the phase space until the end of the forecast range. In February 1986 almost all forecasts are tightly packed and very closed to the verifying analysis. There is some dispersion later on but not as much as in the January case. Also in this case there is no apparent clustering, unless we consider the whole ensemble as a single cluster.

In ensemble forecasting, clustering may be employed in conjuction with the standard deviation from the ensemble mean forecast as additional useful information, which in a practical sense defines more precisely the probability for a certain event. Clusters in Fig. 11 have been identified by a simple visual means and an objective criterion, for example calculating the maximum phase correlation between individual members, is desirable (Murphy and Palmer, 1986). Though no attempt has been made here, the clustering technique may be used as the first step in the more complex task of defining those features in initial conditions which cause individual forecasts to diverge from each other. No unique approach to this problem has been established yet, and an example of possible dependency of the forecast development on the sign of the Pacific/North American mode is given in the study by Palmer (1988).

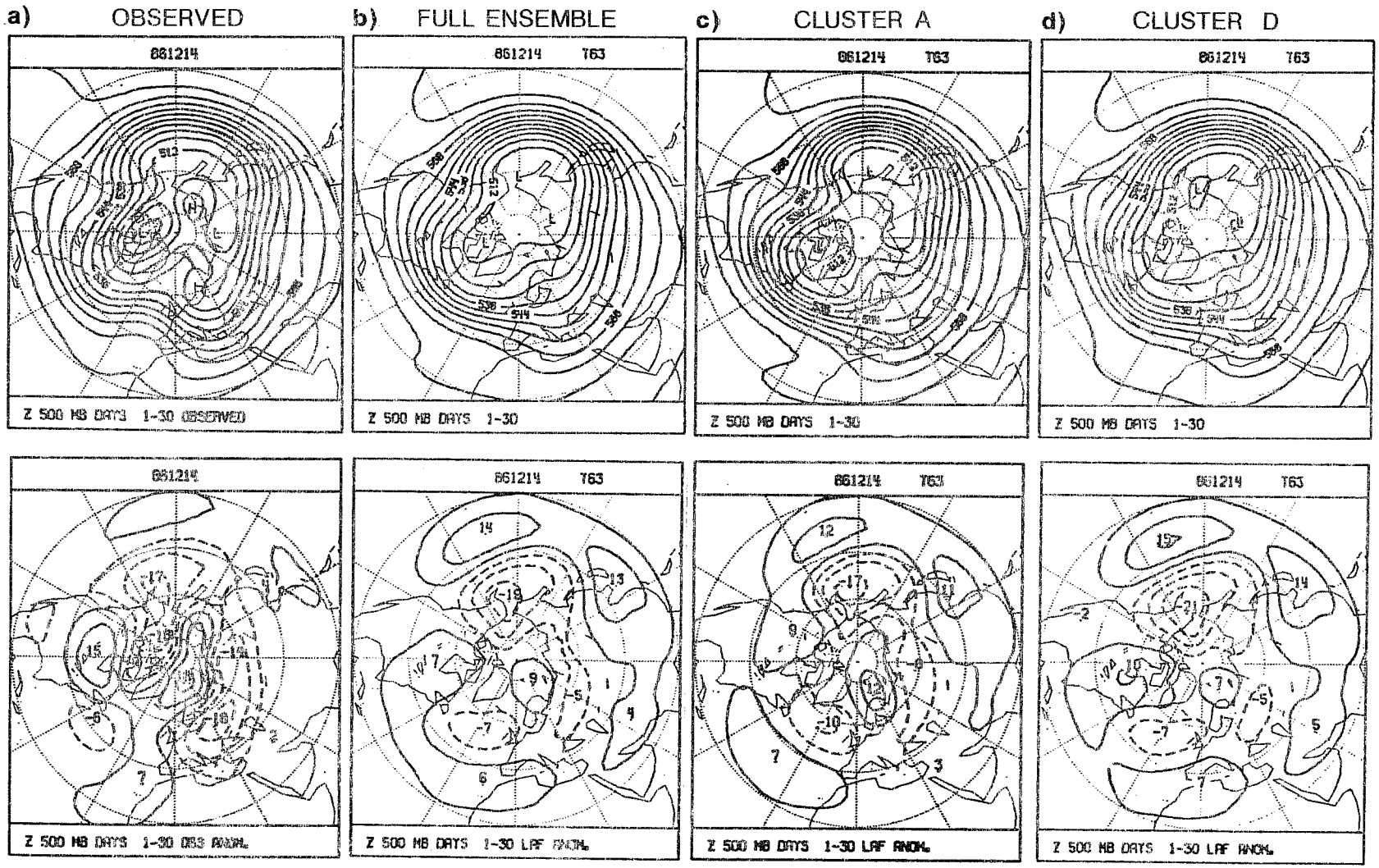


Fig. 13 30-day mean maps of the 500 mb height full fields (top) and anomalies (bottom) for the December 1986 ensemble: a) analysis, b) full ensemble, c) cluster A, and d) cluster D.

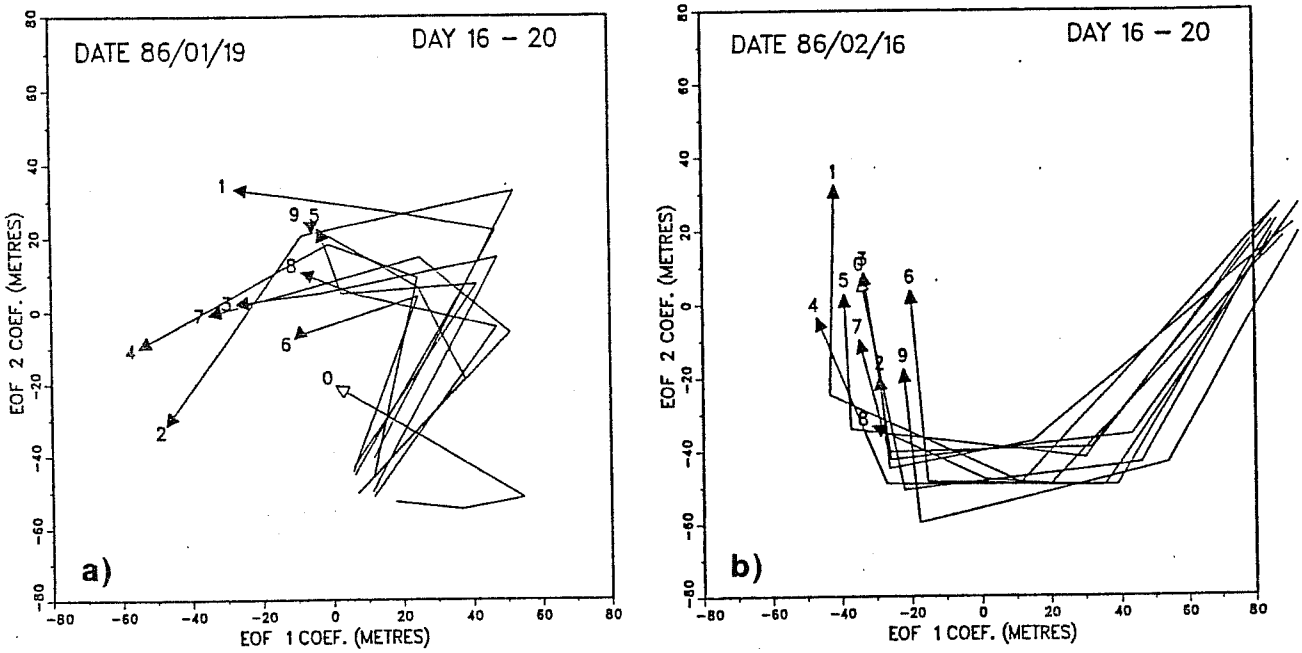


Fig. 14 As Fig.11 but up to mean days 16-20 and for January 1986 (left) and February 1986 (right) ensembles.

7. CONCLUSIONS AND DISCUSSION

In this paper, we have discussed the performance of a number of time-lagged extended-range ensembles of forecasts using the ECMWF operational model at T63 resolution, integrated over a period of three years. The forecasts have been made during all seasons of the year.

We first discussed the properties of the ensembles that would apply in a perfect model environment. According to theory, ensemble averaging invariably reduces the RMS error of the forecast in comparison with the mean skill of individual forecasts comprising the ensemble. With weak restrictions, the magnitude of the anomaly correlation is also increased by ensemble averaging. In addition, the spread of the ensemble about its mean is equal to the skill of the ensemble mean forecast for a sufficiently large ensemble size.

For the set of ensemble forecasts we were able to verify that the ensemble mean forecast had both smaller RMS error, and larger magnitude of anomaly correlation coefficient than the individual mean forecasts. When compared with the latest forecast in the ensemble, the ensemble-mean was occasionally poorer, particularly using the anomaly correlation coefficient. It was shown that, for anomaly correlation coefficient, the latest forecast appeared to be, in general, systematically better than the individual mean skill of the ensemble.

Over the full set, there was a positive correlation between RMS spread and skill. However, much of this could be explained by the annual cycle. That is, both RMS spread and RMS skill tended to be small during the summer than at other times of year. During the winter season, the spread and skill tended to be anticorrelated. It is likely that this reflects the existence of serious systematic errors in the model. It is well-known that systematic deficiencies in the model can influence both the time mean error and the variability in the model. During the forecast from January 1986 a very substantial Euro/Atlantic block occurred (Hoskins and Sardeshmukh, 1987). The forecasts not only failed to predict the occurrence of the block, they all drifted into a unrealistic zonal state with relatively small low frequency variability. Hoskins and Sardeshmukh speculated that an important causal precursor of the Euro/Atlantic block was an increase in anomalous convergence in the upper troposphere over the Caribbean area. As shown in Tibaldi et al

(1988), the model's performance in forecasting large scale tropical divergence patterns is very poor. It is quite plausible that the exceptionally poor forecast from January 1986 was associated with the model's inability to capture the change in Caribbean convergence at the end of the month, together with the drift towards a zonal midlatitude flow with little low frequency variability. The best winter ensemble forecast (in terms of 30-day mean RMS error) was from December 1987. Here the observed flow did not develop into an exceptionally blocked state, and, in addition, improvements to the model (19 levels, orographic gravity wave drag) improved its low frequency variability (Palmer, 1987). The behaviour of these two forecasts, representing extremes in skill and spread in the winter sample, indicates how a negative spread/skill correlation could arise. They show in particular, the difficulty in using ensemble spread as an a priori indicator of ensemble skill.

The only possible evidence of a positive spread/skill correlation, not related to annual cycle, came from the 'transition season' forecasts (that is to say, those forecast that were neither summer nor winter). In the transition season, predictability is greater than during the summer, yet the systematic error of the model is not so large as in winter. Indeed during the transition seasons the drift towards a zonal climate with reduced low frequency variability is not as severe as it is in winter. The analysis of spread versus skill during the transition season did indicate some effect of systematic error, but this appeared to result in a constant bias to the spread skill relation, independent of case. In this sense, it would appear that systematic error is much more 'benign' in the transition season than during the winter period.

We discussed some unconventional diagnostics of model performance. Firstly, we discussed the potential of the probabilistic approach to ensemble forecasting. We showed maps of probability of occurrence of 850 mb temperature lying within prescribed bounds. We argued that this would be the best way to summarise information from the ensemble in a practical forecasting environment. Maps showing the probability of occurrence of predefined classes of some variable offers alternatives to the forecaster otherwise unattainable from the conventional ensemble diagnostics. Secondly, we showed trajectories of some selected forecasts in a phase space spanned by the two dominant EOFs of the forecast ensemble. This gave some indication as to whether members of

the ensemble were showing evidence of clustering. In practice clustering can be used to identify specific synoptic alternatives which have the most likely probability of occurrence.

Further details of the analysis of these time-lagged ensembles will appear elsewhere (Brankovic et al, 1988).

REFERENCES

Brankovic, C., F. Molteni, T.N. Palmer, S. Tibaldi and U. Cubasch, 1988: Extended range forecasting at ECMWF. Part II: Time lagged ensemble forecasting. (in preparation)

Cubasch, U., and A.C. Wiin-Nielsen, 1986: Predictability studies with the ECMWF spectral model for the extended range: The impact of horizontal resolution and sea surface temperature. Tellus, 38A, 25-41.

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

Déqué, M., 1988: The probabilistic formulation: A way to deal with ensemble forecasts. Annales Geophys., 6, 217-224.

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

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. Proceedings of the ECMWF Workshop on stochastic dynamic forecasting, 65-86.

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: 'Atmospheric and Oceanic Variability', Ed. H. Cattle, Royal Meteorological Society, Bracknell, 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.

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., T.N. Palmer and R. Swinbank, 1988: Orographic gravity wave drag: Its parametrization and influence in general circulation and numerical prediction models. Met. and Atmos.Phys. (to be published)

Molteni, F., U. Cubasch and S. Tibaldi, 1986: Monthly forecast experiments with the ECMWF spectral models. Proceedings of the ECMWF Workshop on predictability in the medium and extended range, 51-108.

Molteni, F., S. Tibaldi and T.N. Palmer, 1988: Regimes in the wintertime extratropical circulation. Part I: Observational evidence. (submitted to Quart.J.R.Meteor.Soc.)

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

Murphy, J.M., and T.N. Palmer, 1986: Experimental monthly long-range forecasts for the United Kingdom. Part II: A real-time long-range forecast by an ensemble of numerical integrations. Meteor.Mag., 115, 337-349.

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., C. Brankovic and S. Tibaldi, 1988: Extended range forecasting at ECMWF. Part III: Impact of interannual variability. (in preparation)

Palmer, T.N., and F. Molteni, 1987: An experimental scheme to predict forecast skill. ECMWF Tech. Memo. No. 141, 13 pp.

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. (submitted to Mon.Wea.Rev.)

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, 1988: The ECMWF medium-range predicition models: Development of the numerical formulations and the impact of increased resolution. Met. and Atmos.Phys. (to be published).

Tibaldi, S., C. Brankovic and U. Cubasch, 1987: 30 day integrations using the operational ECMWF spectral model. ECMWF Tech. Memo. No 138, 53 pp.

Tibaldi, S., C. Brankovic, T.N. Palmer and U. Cubasch, 1988: Extended range forecasting at ECMWF. Part I: Impact of horizontal resolution. (in preparation)

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.