The skill of 500hPa height forecasts

A.J. Simmons
European Centre for Medium-Range Weather Forecasts
Shinfield Park, Reading, UK

Summary

The properties of two objective measures of skill of 500hPa height forecasts, root-mean-square error and anomaly correlation, are discussed and illustrated. Long-term trends in skill of ECMWF forecasts are presented, showing steady improvement in the short range and an overall improvement in the medium range. Following arguments of Lorenz(1982), the evolution of differences between consecutive forecasts valid at the same time is studied to estimate intrinsic error growth rates. Diagnosis of current high-resolution (T213L31) ECMWF forecasts and comparison with forecasts from the UK Meteorological Office gives confidence in recent estimates. Short-term variability of these forecasts is also discussed. Current T213L31 forecasts are also compared with the operational forecasts of earlier years, and with the lower resolution (T63L19) control forecasts of the ECMWF Ensemble Prediction System (EPS). An apparent lack of improvement of medium-range forecasts from 1986 to 1992 is ascribed to the introduction of more active and more realistic versions of the forecast model during this period. These portrayed atmospheric evolution more accurately given accurate initial conditions, but also amplified error more rapidly. Low amplification rates of the current T63L19 version are consistent with a lack of spread in EPS forecasts. Intrinsic error growth rates deduced from the T213L31 forecasts are larger in summer than winter, in contrast with the amplification factors found in singular-vector calculations which do not include a full range of diabatic processes. Ways of smoothing forecasts to remove unpredictable scales are illustrated, and the potential for improvement of deterministic medium-range forecasts is discussed.

1. Introduction

The performance of past and present numerical weather prediction systems provides a baseline for studies of the predictability of the atmosphere. Objective verification of the 500hPa height field has for long provided a useful way of measuring the skill of large sets of synoptic-scale forecasts, and is still widely used to assess proposed changes to operational forecasting systems, or to compare the performance of different operational centres. This paper begins with a discussion of two of the most popular measures, root-mean-square error and anomaly correlation. Examples are drawn from recent operational ECMWF forecasts of the 500hPa height field. Results are then presented showing the evolution of skill of these forecasts over the period from 1980 to 1995. Month-to-month variations in skill are also illustrated.

Several years ago there were concerns expressed both at ECMWF and in the Member States over a lack of long-term improvement of the skill scores of the ECMWF medium-range forecasts. A distinct short-term variability in the quality of these forecasts was a further problem. An important practical outcome of this was the development of an ensemble

(probabilistic) prediction system (EPS; Molteni et al., 1996), a topic discussed in several contributions to these Proceedings. Improved understanding of the behaviour of the deterministic forecasting system was nevertheless needed, as a guide to improvement of the system (and thereby the control forecasts of the EPS) and as a guide to the extent to which research effort and computational resources should be allocated between the areas of modelling and data assimilation, and between deterministic and probabilistic forecasting. This prompted a number of studies. One was the sensitivity analyses (Rabier et al., 1996) discussed elsewhere in these Proceedings. Another was a re-examination of predictability estimates derived by Lorenz(1982), using more recent forecast results (Simmons et al., 1995).

Lorenz(loc. cit.) had examined the prospects for more accurate numerical weather predictions. He studied the evolution of differences between successive forecasts of the 500hPa height field produced daily by ECMWF and verifying in the 100-day period beginning 1 December 1980. He argued that if the forecast model in operational use at the time was realistic enough for small differences in initial conditions to amplify at a rate close to that at which separate but similar atmospheric states diverge, then the rate of growth of differences between consecutive forecasts valid at the same time provided a limit to the potential accuracy of the forecast which could not be surpassed without reduction of the one-day forecast error. Lorenz also calculated a "best fit" to a simple analytical model of this limiting error growth, from which it was possible to estimate the further reduction in forecast error beyond the one-day range that could be expected to result from a given reduction in the one-day forecast error. He concluded that it appeared to be possible to make predictions that would be at least as skilful at ten days ahead as the 1980/81 forecasts were at seven days ahead, and that additional improvements might be realized if the one-day forecast was capable of being improved significantly.

Many changes have been made to the ECMWF forecasting system over the years since the forecasts analyzed by Lorenz were made. The principal changes affecting the accuracy of 500hPa height forecasts have been summarized by Simmons et al.(1995) for the period up to the end of 1993. Since then, the one-dimensional variational processing of satellite radiance measurements introduced for the northern hemisphere in 1992 has been revised and extended to the tropics and southern hemisphere (McNally and Vesperini, 1995). Implementation was in December 1994. In April 1995 there were major model changes in the areas of cloud parametrization (Tiedtke, 1993), the representation of orography (Lott and Miller, 1996), and numerical formulation (Hortal, 1994). Miller et al.(1995) have summarized the impact of this important set of changes.

Here we present and extend material from the study by Simmons et al.(1995). In particular we discuss results for seasons up to the autumn of 1995, thereby extending our earlier study by 18 months. This extra period was one in which forecasts were of unprecedented accuracy. In addition, we make some comparisons with recent UK Meteorological Office forecasts and with the lower resolution control forecasts of the ECMWF Ensemble Prediction System (EPS).

2. Measures of forecast skill

Two measures of forecast skill (or forecast error) are widely used to assess numerical medium-range forecasts of height fields. These are the root-mean-square (rms) error and the anomaly correlation coefficient. In this section we set out the basic definitions and discuss some properties of these scores.

2.1 Root-mean-square error

The rms error E_j of the day j forecast is defined by

$$E_j = \sqrt{\overline{(f_j - a)^2}} \tag{1}$$

where f_j is the day j forecast of the height field at a particular pressure, here taken to be 500hPa, and a is the corresponding verifying analysis. The overbar $\overline{(\ )}$ denotes an average over area and over a set of forecasts. The rms error is initially small, and vanishes in the case (as here) in which the verifying analysis is the same as the analysis from which the forecast started. The error generally grows in time and reaches the asymptotic level indicated below.

The rms error can be written

$$\begin{aligned}(E_j)^2 &= \overline{((f_j - c) - (a - c))^2} \\ &= \overline{(f_j - c)^2} + \overline{(a - c)^2} - 2\overline{(f_j - c)(a - c)} \\ &= (A_j)^2 + (A_a)^2 - 2\ \overline{(f_j - c)(a - c)}\end{aligned} \tag{2}$$

Here c is the climatological value of the height field for the verifying day. A_j is the rms anomaly of the day j forecast and A_a is the rms anomaly of the verifying analysis. We shall generally use the word "variance" as applied to analyses or forecasts to mean the quantities $(A_a)^2$ or $(A_j)^2$. The differences between these quantities and the variances conventionally defined as the mean square deviations from the sample means (as opposed to climatology) will be discussed later, in section 7.

As the forecast range and sample size increase, the forecast and analyzed anomalies tend to become uncorrelated:

$$\overline{(f_j - c)(a - c)} \to 0$$

This assumes there to be no systematic correlation as would occur if forecasts and analyses shared a common error, or if the estimate of the climatological field itself were in error. A correlation could also occur for a limited sample (for example involving all forecasts for just one season) if upper-air forecasts and analyses shared a common anomaly, such as might

result from anomalous sea-surface temperatures or soil moistures. These effects will not be completely absent in practice, but assuming them to be negligible,

$$E_j \rightarrow \sqrt{(A_j)^2 + (A_a)^2}$$

For a perfect model $A_j \rightarrow A_a$, so $E_j \rightarrow \sqrt{2}\, A_a$. If the model loses variance about the climatological mean, E_j tends to a limit less than $\sqrt{2}\, A_a$.

2.2 Anomaly correlation coefficient

The anomaly correlation coefficient measures the correlation between forecast and analyzed deviations from climatology:

$$ACC_j = \frac{\overline{(f_j - c)(a - c)} - \overline{(f_j - c)}\,\overline{(a - c)}}{\sqrt{[\overline{(f_j - c)^2} - \overline{(f_j - c)}^2]\,[\overline{(a - c)^2} - \overline{(a - c)}^2]}} \qquad (3)$$

For presentation purposes it is common to multiply the coefficient by 100 to express the correlation as a percentage. In practice, if systematic model errors are low, if the correlation is calculated over an area such as the extratropical northern hemisphere, and if the averaging period is as long as a season, then the mean deviations of analyses and forecasts from climatology, $\overline{(a - c)}$ and $\overline{(f_j - c)}$, are sufficiently small that the anomaly correlation coefficient is well approximated by the simpler form defined by

$$AC_j = \frac{\overline{(f_j - c)(a - c)}}{\sqrt{[\overline{(f_j - c)^2}]\,[\overline{(a - c)^2}]}} \qquad (4)$$

Quantitative comparisons of seasonal-mean values of ACC_j and AC_j are given in Table 1 of Simmons et al. (1995), where it is shown that the approximate form is accurate to within a few tenths of one per cent for present values of systematic forecast error.

2.3 Normalized mean square error and anomaly correlation

The anomaly correlation coefficient is closely related to a normalized mean square forecast error. The normalizing factor is the asymptotic limit of mean square error indicated in section (2.1). The normalized error N_j is given by

$$N_j = \frac{(E_j)^2}{(A_a)^2 + (A_j)^2}$$

From equation (2),

$$N_j = 1 - \frac{2 A_j A_a}{(A_j)^2 + (A_a)^2} AC_j$$

$$= 1 - AC_j \left(1 - \frac{(A_j - A_a)^2}{(A_j)^2 + (A_a)^2} \right)$$

If the forecast model is such that A_j is close to A_a (as is typically the case for today's better models),

$$N_j \approx 1 - AC_j \qquad (5)$$

2.4 Smoothing the forecast to reduce rms error

A smoothed forecast s_j can be defined by linearly combining the numerical forecast and a simple forecast of climatology:

$$s_j = \epsilon_j f_j + (1 - \epsilon_j) c$$

ϵ_j can be determined to minimize rms error. Minimization of $\overline{(\epsilon_j f_j + (1 - \epsilon_j) c - a)^2}$ is achieved by choosing

$$\epsilon_j = \frac{\overline{(f_j - c)(a - c)}}{\overline{(f_j - c)^2}} = \left(\frac{A_a}{A_j} \right) AC_j$$

If the blending with climatology is carried out uniformly at grid points, then s_j has a lower rms error than the numerical forecast f_j, but has the same anomaly correlation. In this case the blend can be made optimal for a particular region. Alternatively, the blend can be made separately for each wavenumber component of spectrally decomposed fields, enabling the less

skilful smaller-scale components of the forecast to be more heavily filtered than the larger, more predictable scales.

3. Some examples

3.1 Root-mean-square error

Fig. 1 shows rms errors and asymptotes for 500hPa height forecasts. The domain is the extratropical northern hemisphere, and the sample comprises all operational forecasts verifying within the months of December 1994, January 1995 and February 1995, a period which we shall refer to as Winter 1995. These calculations are based on a dataset of 500hPa forecasts in which fields are represented spectrally and truncated at wavenumber 40. This "Lorenz" dataset as a whole covers the period from 1 December 1980 to the present, and was designed originally to facilitate the predictability study by Lorenz(1982). Further details of the calculations are as given by Simmons et al.(1995).

The solid curve in Fig.1 shows the rms error E_j as given by equation (1). Error grows at an increasing rate out to about day 5. The error growth rate subsequently decreases, though by day 10 the error is still growing, and is some 20% below the asymptotic limit $\sqrt{(A_j)^2 + (A_a)^2}$, which is shown by the uppermost dashed line. Note that this line slopes slightly upward to the right, indicating that the mean square forecast anomaly increases with increasing forecast range, or in other words that the forecast model tends to simulate too high a variance of the 500hPa height field. By day 10, the limit is about 3% higher than the perfect-model limit $\sqrt{2}\,A_a$, suggesting scope for a similar percentage reduction in rms forecast error at this range due to correction of this particular model deficiency. Note however that this reasoning assumes there to be no systematic underestimation of variance by the initial analyses.

The dotted curve in Fig.1 shows the rms error of the smoothed forecast s_j formed by blending the numerical forecast and climatology in the way indicated in section (2.4). The blending is done in a scale-independent way, using coefficients ϵ_j computed from data for Winter 1994. Smoothing the forecast in this way gives very little reduction in rms error out to about day 4. Rms error is reduced at longer ranges in such a way that it approaches the asymptotic level A_a, the root mean square anomaly of the analyzed fields. As mentioned earlier, this smoothing of the forecast gives no change in anomaly correlation.

Also shown in Fig. 1 is a dashed curve which lies very close to the solid curve representing E_j. This curve is computed by taking the arithmetic mean of daily rms errors, the way monthly-mean forecast scores are compiled in the operational verification system at ECMWF (Nieminen, 1983; Norris, 1994). The two methods of computing a seasonal-mean score

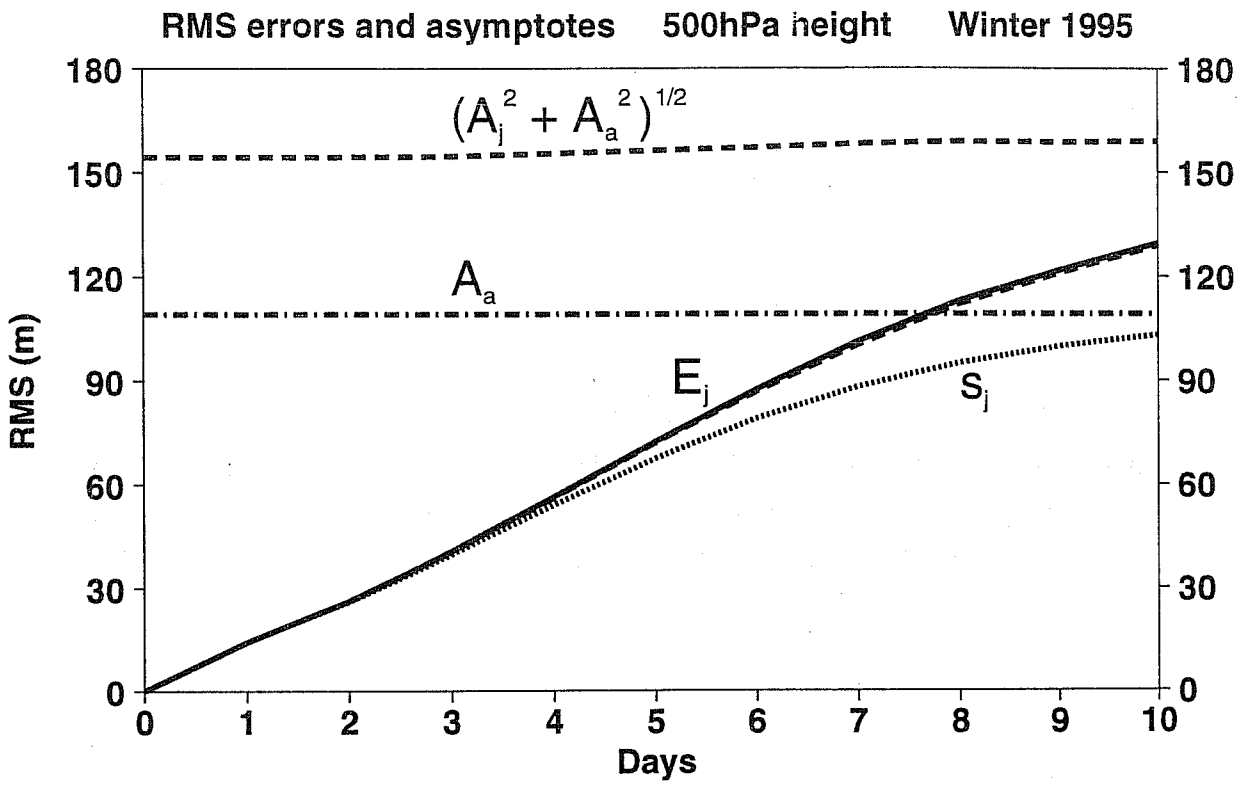

Fig. 1 Rms errors and asymptotes (m) of 500hPa height forecasts for the extratropical northern hemisphere averaged for Winter 1995. The solid curve denotes the rms error of the whole sample, E_j, and the dashed curve that lies very close to the solid curve is the mean of daily values. The dotted curve shows the rms error of the smoothed forecast s_j formed by blending the numerical forecast with climatology. The uppermost dashed curve is the asymptotic limit of the unsmoothed numerical forecast, and the dash-dotted curve the asymptotic limit of a forecast of climatology.

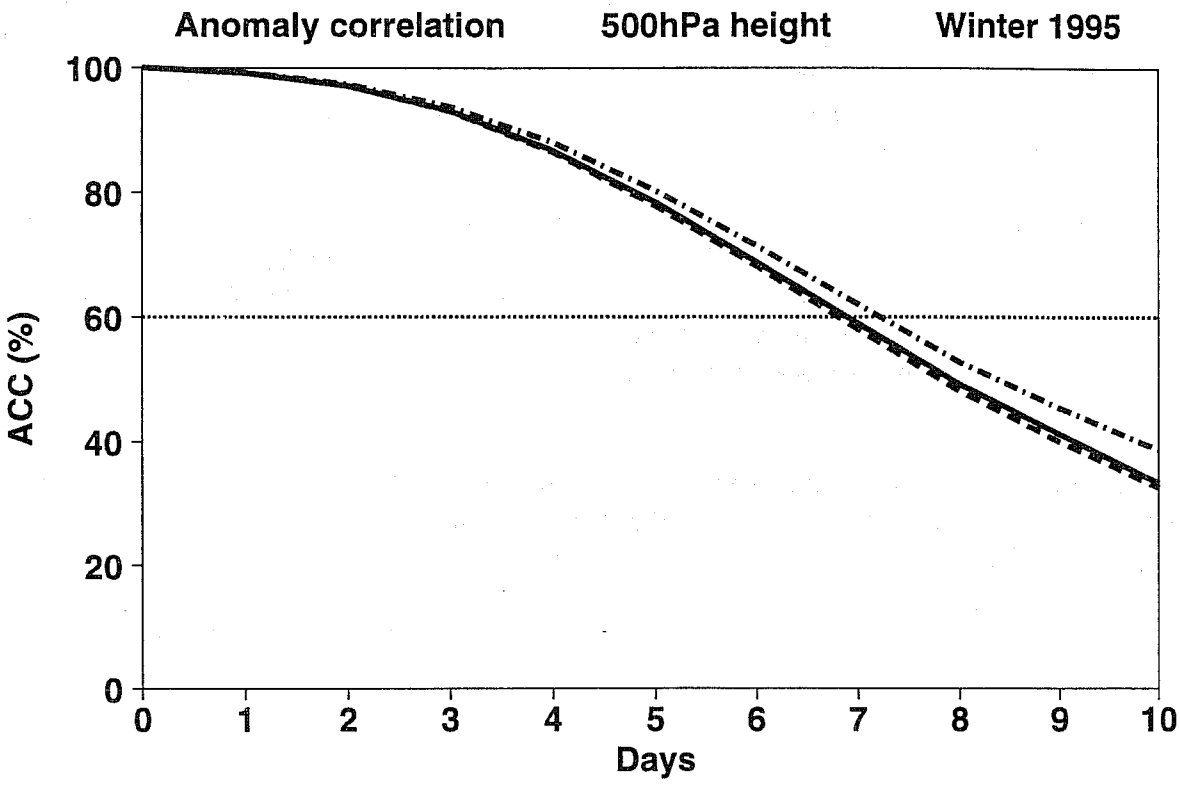

Fig. 2 Anomaly correlations (%) of 500hPa height forecasts for the extratropical northern hemisphere averaged for Winter 1995. The solid line denotes the correlation computed for the sample as a whole, and the dashed line the average of daily correlations. The dash-dotted curve is the average of daily correlations computed with a different climatology.

evidently give very similar results. To see why, we denote the set of daily day-j rms errors by e_j , and let the overbar here denote an average over the set of forecasts. Then

$$E_j = \sqrt{\overline{e_j^2}}$$

$$= \overline{e_j} \sqrt{1 + \overline{(e_j - \overline{e_j})^2} / \overline{e_j}^2}$$

$$\approx \overline{e_j} \left(1 + \tfrac{1}{2} \overline{(e_j - \overline{e_j})^2} / \overline{e_j}^2 \right)$$

At day seven, for Winter 1995, the mean daily rms error for the extratropical northern hemisphere $\overline{e_j}$ is around 100m and the standard deviation of e_j is about 16m. The alternative ways of computing the seasonal-mean score thus differ by only a little more than 1%. Scores are more variable from day to day for smaller domains, and differences between the two mean scores are of the order of 5% when computed for Europe alone. Typical variations in daily rms scores are about twice as large for Europe as for the extratropical northern hemisphere.

3.2 Anomaly correlation

Fig. 2 shows anomaly correlations computed for the extratropical northern hemisphere for Winter 1995, derived from the "Lorenz" dataset. The solid curve denotes the coefficient computed according to the complete formula (3). The approximation given by (4) is not shown as it is almost indistinguishable from the solid curve. The anomaly correlation decreases from its initial value of 100% as the forecast range increases, dropping below 50% at about the same time as the rms error reaches the value A_a , the rms error of a climatological forecast. This is consistent with the relationship between the normalized mean square error and approximate anomaly correlation given by equation (5). A slightly higher correlation, 60%, is usually chosen to delimit a useful forecast, based on comparisons with subjective assessments of forecast skill (e.g. Simmons, 1986). The 60% line is marked in Fig. 2, and is crossed at about day 7.

The dashed curve in Fig. 2 shows the mean of the daily anomaly correlations. As for rms error, averaging the daily scores gives much the same answer as calculating the score for the sample as a whole. If the averaging of daily scores is done through the Fisher z transform (discussed by Buizza(1996) in this volume), this average lies just above the solid curve. These three calculations produce seasonal-mean anomaly correlations that differ at most by under 2%.

Of more interest is the difference between the dashed and dash-dotted curves in Fig. 2. The dash-dotted curve shows the averages of the daily anomaly correlations from the operational ECMWF verification system, rather than from the "Lorenz" dataset. This clearly gives an impression of higher forecast accuracy, with a difference in correlation of more than 5% at day ten. There are several differences of detail between the operational calculations and those we have carried out from the "Lorenz" dataset. Eliminating them, it is found that the

difference in anomaly correlation comes mostly from the different climatologies used in the two verifications. The operational verification has been running since January 1980, and makes use of a climatology available at that time, produced at NCAR. Our present calculations use a fifteen-year climatology constructed from the ECMWF analyses for the years from 1981 to 1995. It would be desirable in principle to change the operational verification to use a more up-to-date climatology. However, it is impractical to recompute the full set of archived verification statistics, so any change would be at the expense of a loss of continuity of the statistics, which would inhibit their use for long-term monitoring of forecast quality.

3.3 Scale dependence

A spectral decomposition of the global-mean square error of 500hPa height forecasts is presented in Fig. 3a. Results are shown for days one, three, five, seven and ten, for the period from December 1994 to February 1995 (which we have referred to as Winter 1995 when discussing results for the northern hemisphere). The net contribution, S_n, to the error from all spectral components with total wavenumber n is shown for $n \leq 40$: if an error field e is represented by the spectral expansion

$$\sum_{m=-40}^{m=40} \sum_{n=|m|}^{n=40} e_n^m \, Y_n^m$$

in terms of the spherical harmonics Y_n^m, then the global mean of e^2 is given by

$$\sum_{n=0}^{n=40} S_n$$

where $S_n = |e_n^0|^2 + 2\sum_{m=1}^{m=n} |e_n^m|^2$ for an appropriate normalization of the Y_n^m.

The error spectrum is relatively flat at day one. Error grows at all scales, and tends to saturate first at small scales, though this has not happened by day three for any part of the wavenumber range illustrated. Wavenumbers higher than about 30 have saturated by day five, and wavenumbers higher than about 20 have saturated by day seven. This saturation occurs later in the time range than was the case in earlier years (Simmons et al., 1995). Results for the complete wavenumber range of the current T213 model have been presented by Boer(1994) for the month of February 1993.

A filtering of less-predictable scales has been derived by forming a blending with climatology (as in section 2.4) in which coefficients are determined separately for each total wavenumber n. Coefficients are thus determined optimally for the reduction of global-mean error. However, the additional reduction of error due to scale selectivity gives rms errors for the extratropical northern hemisphere that are in fact slightly lower than illustrated in Fig. 1 for the scale-independent grid-point smoothing that is optimal for this domain. The scale-selective smoothing also increases the anomaly correlation, though not to a large extent. This is illustrated in Fig. 3b for the extratropical northern hemisphere. Note that in this

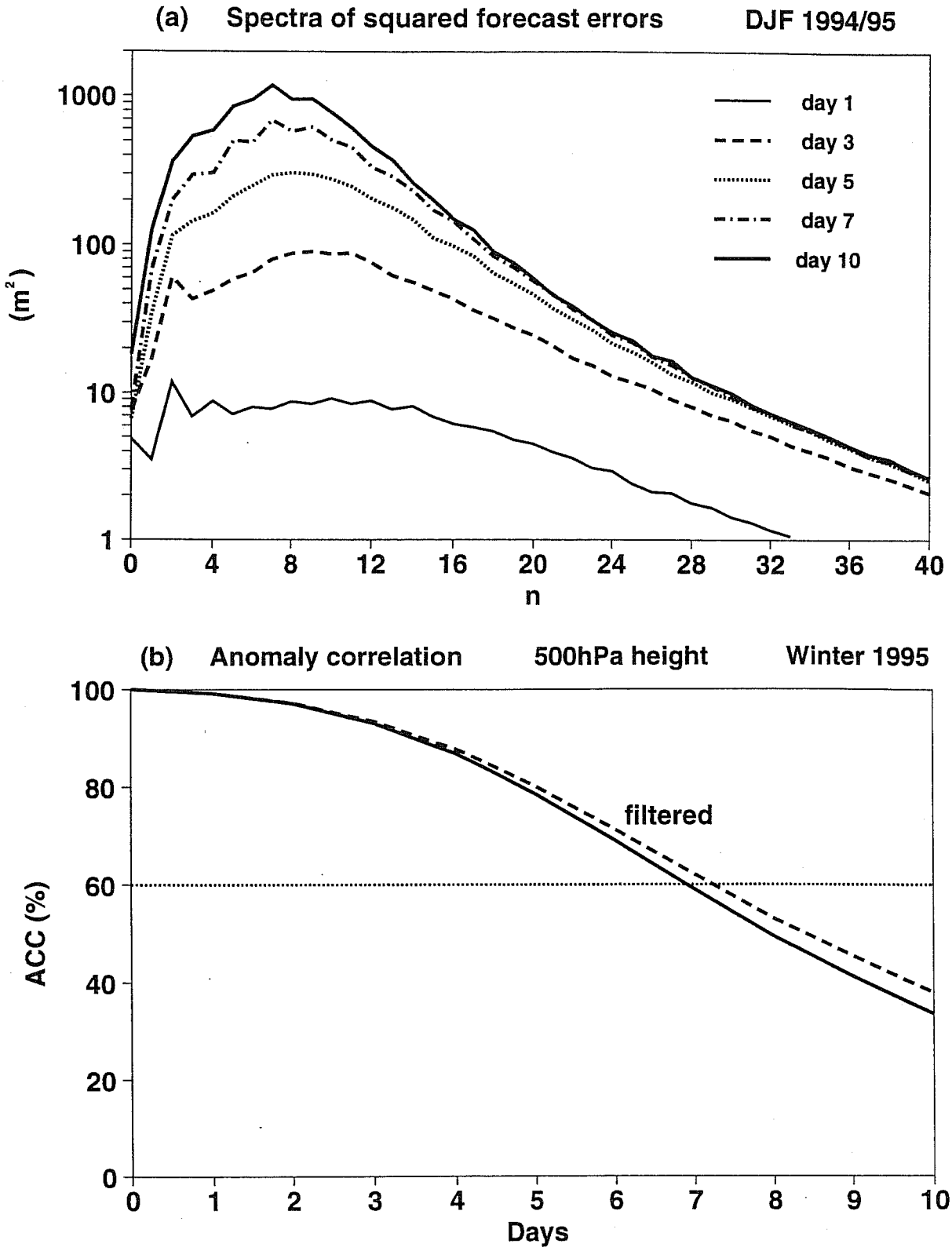

Fig. 3 (a) Spectra of global squared errors of 500hPa height forecasts (m^2) averaged for the period 1 December 1994 to 28 February 1995 ("Winter 1995").

(b) Corresponding anomaly correlations for the extratropical northern hemisphere, showing the impact of scale-selective filtering of forecast fields.

verification no smoothing was applied to the verifying analyses. Substantially higher correlations are found if the larger-scale components of the forecasts are verified against just the larger-scale components of the analyses (as illustrated, for example, by Simmons, 1986).

3.4 Seasonal dependence

The most pronounced seasonal effect on verification scores is a pronounced annual cycle in rms error. This can be seen in Fig. 4, which shows monthly-mean rms errors and anomaly correlations for the extratropical northern hemisphere averaged over the years from 1980 to 1995, for forecast days five, seven and ten. Verification data produced operationally have been used. The predominant cause of the variation in rms error of the medium-range forecasts is the annual cycle in variance of the analyses (and forecasts) of the 500hPa height. We have seen how the anomaly correlation is approximated by a simple function of the mean square error of the forecast normalized by the sum of the analyzed and forecast variances. It thus exhibits less of an annual cycle than rms error, but correlations are notably higher in winter months (especially February, on average) than in summer and early autumn (especially June, September and October). This is discussed again later in this paper.

Interannual fluctuations in the variance of the 500hPa height can make it difficult to interpret changes in rms error. The upper panel of Fig. 5 shows the rms day-seven error computed for a European domain. The plot is based on archived operational monthly-mean scores for the period from 1989 to 1995, and the annual cycle has been removed by computing a 12-month running average of the monthly-mean data. The sharp rise in rms error in 1991/92 was of concern at the time as it was (in part probably correctly) linked to problems in the performance of the higher resolution (T213L31) semi-Lagrangian model introduced operationally in September 1991 (Ritchie et al., 1995). Similarly, the fall in 1992/93 was ascribed to solution of some of the problems identified in the performance of the new model. Caution is needed, however. The lower panel of Fig. 5 shows the corresponding rms anomaly of the verifying analysis, A_a . In the period 1991-1993 it shows much the same variation as the rms error, suggesting that at least some of the variation in error was a consequence of variations in atmospheric circulation characteristics over the period. Conversely, the fall in rms error over the last year has been at a time of increasing rms anomaly in the verifying analyses, suggesting that there has been a very real benefit of the operational changes made in late 1994 and April 1995.

4. Long-term trends in skill scores

Fig. 6 gives an indication of trends in the skill of ECMWF medium-range forecasts from 1980 to 1995. The plotted points denote the forecast range at which the monthly-mean anomaly correlation (from the operational verification system) reaches the 60% value, and the solid lines denote 12-month running means. The upper panel is for the extratropical northern hemisphere as a whole, and the lower panel is for Europe.

It is evident from Fig. 6 that there can be very substantial fluctuations in skill from month to month, particularly for the smaller, European domain. There is a tendency for the highest

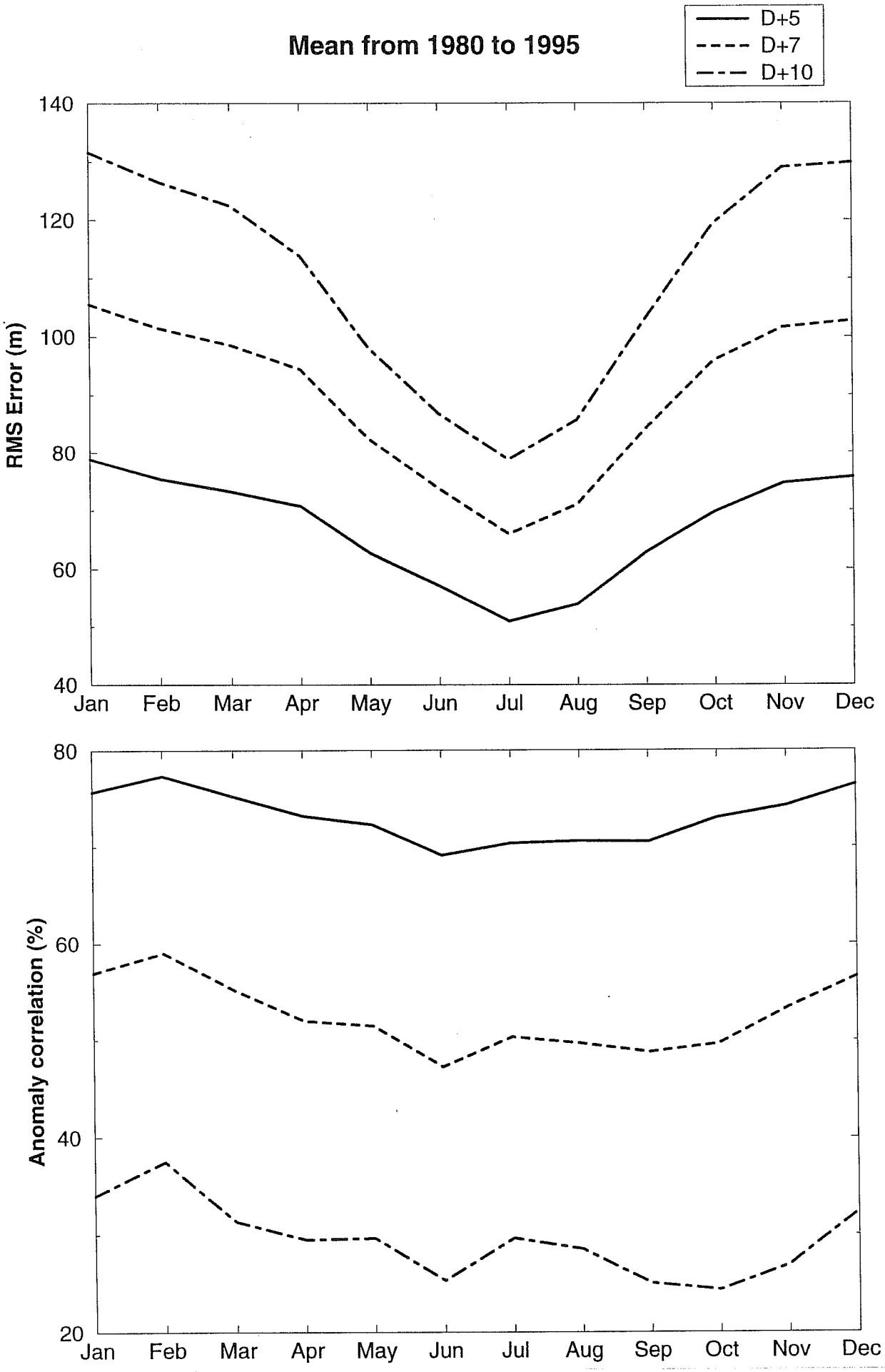

Fig. 4 Seasonal variation of rms errors (upper) and anomaly correlations (lower) for five-, seven- and ten-day 500hPa height forecasts for the extratropical northern hemisphere. Mean data for each month have been averaged over the period from 1980 to 1995.

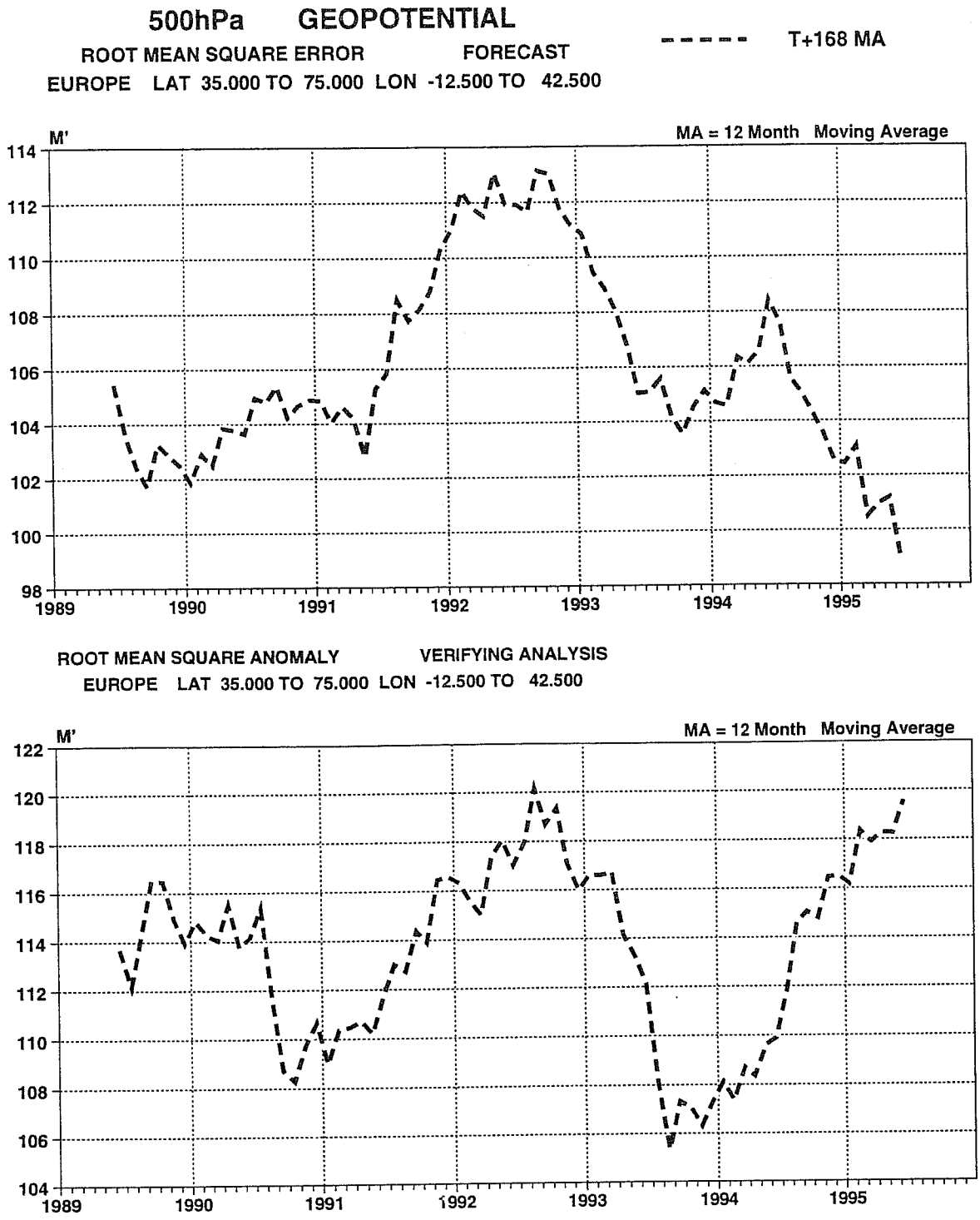

Fig. 5 Rms error of seven-day 500hPa height forecasts and rms anomaly of the verifying analyses for European-area forecasts from 1989 to 1995. Twelve-month running means of monthly-mean data are plotted. The domain is from 12.5°W to 42.5°E and from 35°N to 75°N.

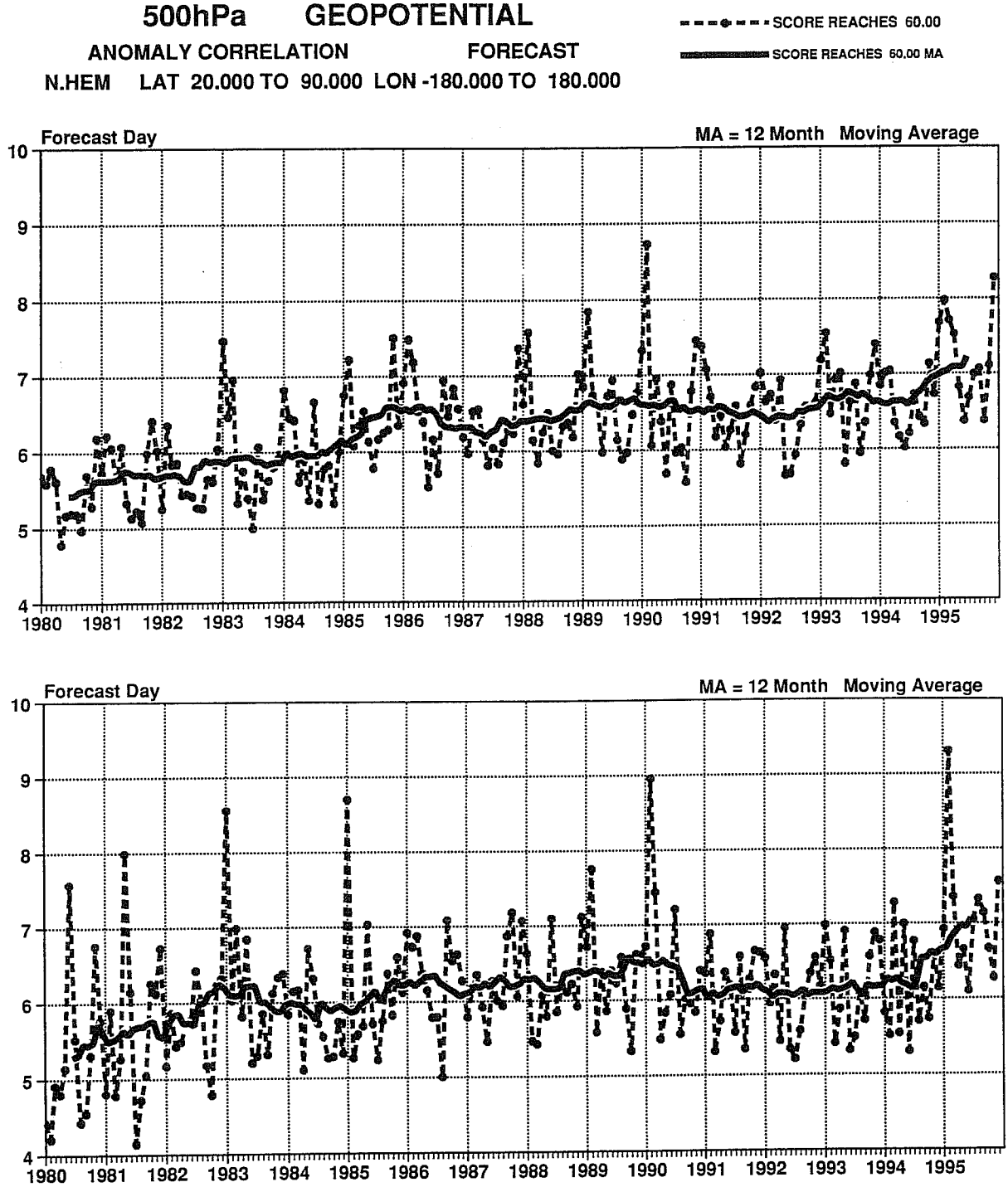

Fig. 6 Forecast range at which the monthly-mean 500hPa height anomaly correlation reaches the 60% value, plotted for each month from 1980 to 1995 (dashed) and for 12-month running means (solid). The upper panel is for the extratropical northern hemisphere and the lower panel is for Europe.

skill to occur in winter months, but the outstanding European scores in February 1990 and February 1995 have to be contrasted with the relatively poor performance over Europe in February 1988 and February 1994. Also, the mean skill of the European forecasts in June 1980 and May 1981 has yet to be surpassed in a non-winter month.

The month-to-month variations in Fig. 6 cannot be explained by changes to the forecasting or observing system, and indicate a dependence of forecast skill on the character of the prevailing atmospheric circulation. This may be in part due to a dependence of model error on circulation type, or a dependence of forecast skill on how well key development regions (whose positions may vary) are covered by the largely fixed and inhomogeneous distribution of observations. It may also be indicative of a more fundamental variability in the stability of the atmospheric circulation pattern. Whatever the cause, the variations can make it difficult to use operational scores to assess the impact of particular changes to the forecasting system. Interpretation of scores can however be helped by comparison with the scores of other operational centres.

A general trend of forecast improvement is nevertheless clearly seen in the overall upward slope of the running means in Fig. 6. In 1980, the 60% level was reached on average before day 5½. In 1995, it was reached just beyond day seven. It is noteworthy that much of the increase in forecast accuracy appears to have occurred either early or late in the 16-year period, with little evidence in these plots of improvement from 1986 to 1992. This will be discussed later.

A more uniform trend in forecast skill is seen at earlier forecast ranges. Fig. 7 shows annual running means of the forecast ranges at which the anomaly correlation passes the 98%, 92.5%, 80% and 60% levels, again for the extratropical northern hemisphere and for Europe. The improvement in the one- to three-day range is especially steady for the larger domain. In particular, there is a clear reduction in short-range forecast error over the period from 1986 to 1992.

The above assessments of forecast skill are based on time averages of daily forecast scores. An alternative assessment is provided by counting the numbers of forecasts whose scores are below or above certain thresholds. Fig. 8 presents examples based on the day-six forecast for Europe. The numbers of forecasts with anomaly correlations less than 60%, 30%, and 0% are plotted for each year from 1981[1] to 1995, and as totals per month for the period 1981-1995. After some reduction in the number of poor forecasts in the early 1980s, there was little change over the period 1985-1994, with if anything a slight increase in poor forecasts after 1990. The numbers were, however, substantially lower in 1995. The number of poor forecasts tends to be relatively large in summer, early autumn and early spring. Very poor forecasts are most common in September, and least common in May and November. The sample size is small for correlations below 0%, but the peak in September did not come entirely as a surprise when these statistics were first examined. European forecasts in late

[1]Numbers are not shown for 1980, because operational forecasting was carried out only on weekdays in the first seven months of the year.

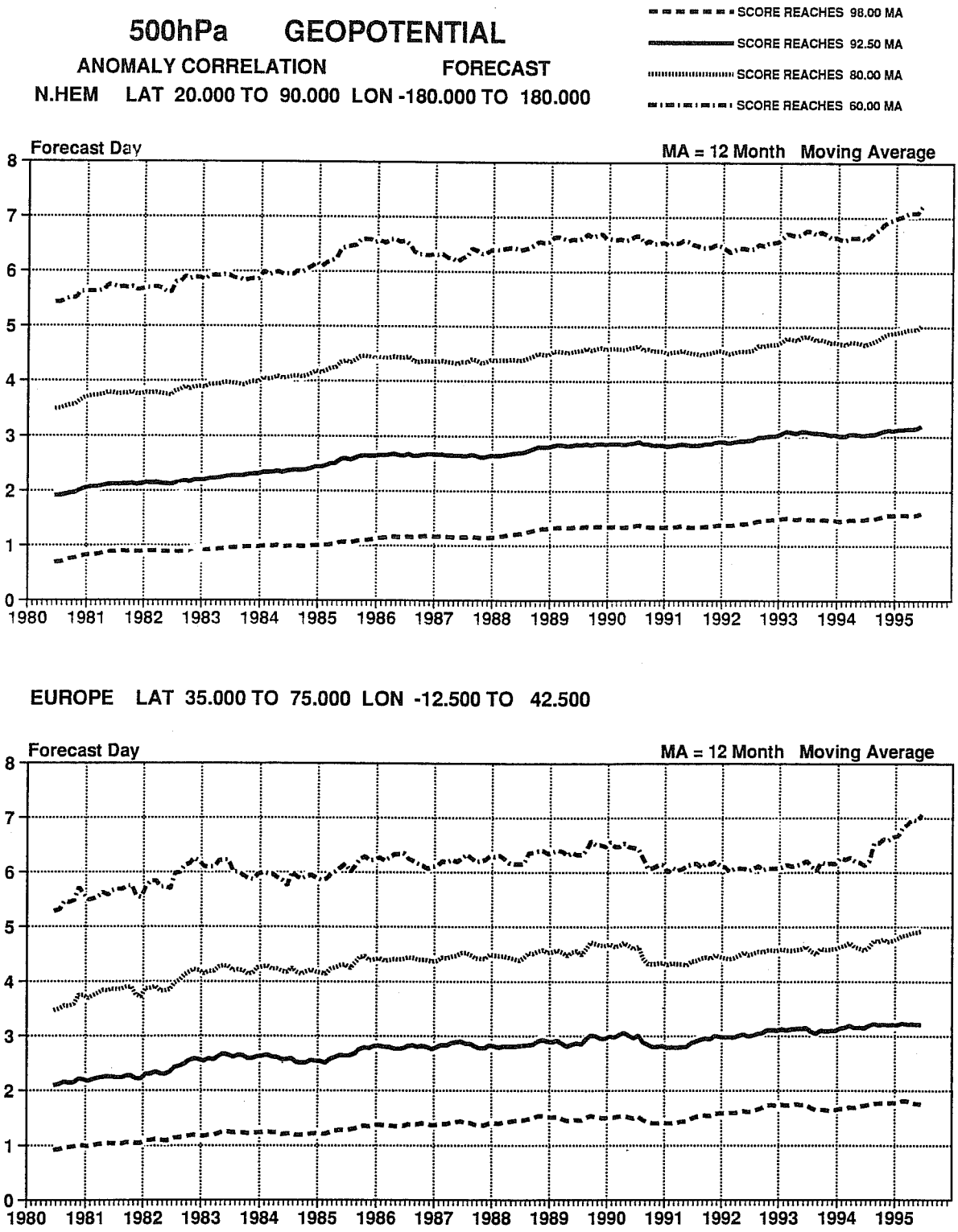

Fig. 7 Forecast ranges at which the monthly-mean 500hPa height anomaly correlation reaches the values 98%, 92.5%, 80% and 60%. Twelve-month running means are plotted for the extratropical northern hemisphere (upper) and for Europe (lower).

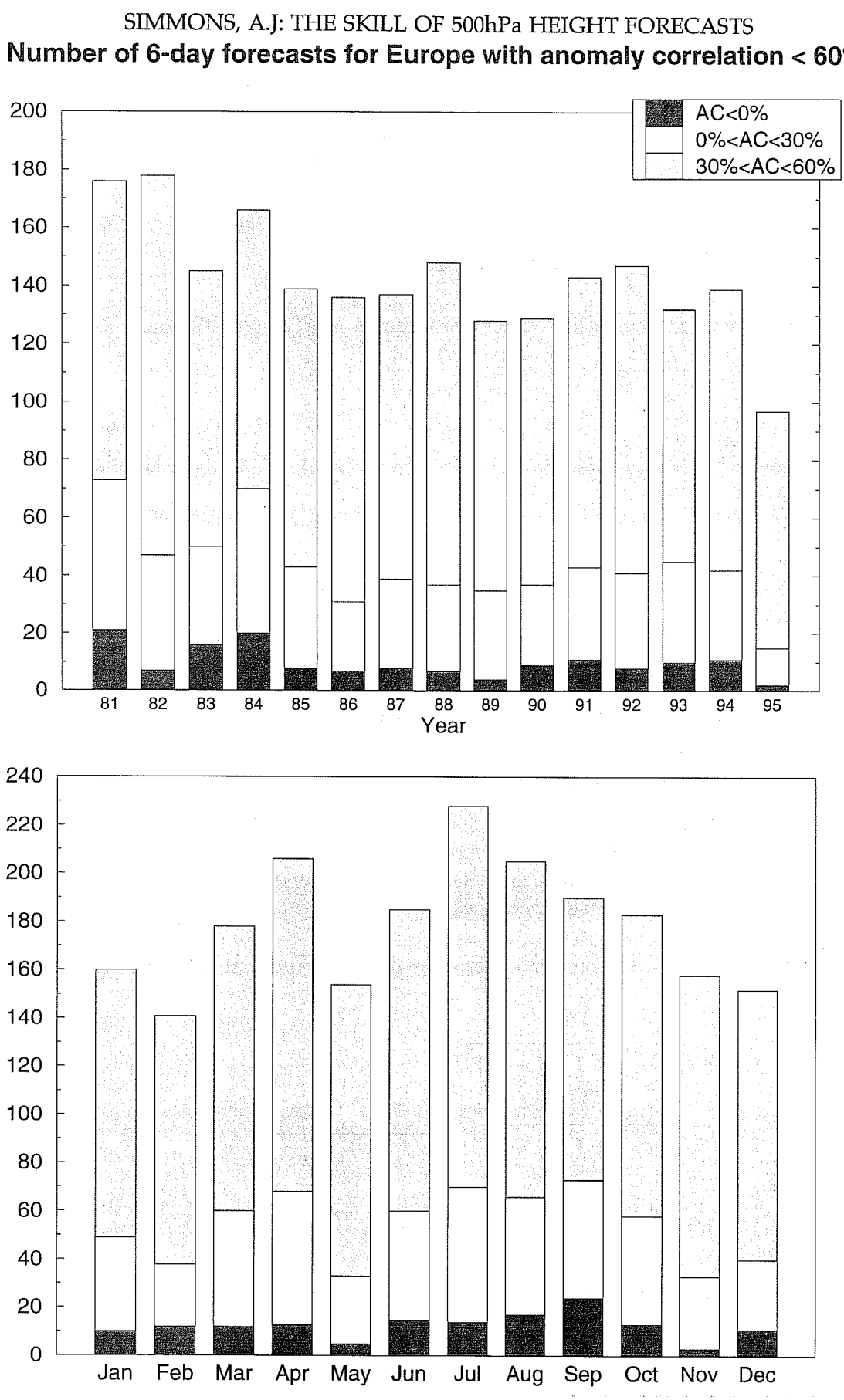

Fig. 8 Numbers of six-day forecasts for which the 500hPa height anomaly correlation for Europe lies below 60%, 30% and 0%. Results for each year from 1981 to 1995 are shown in the upper panel, and the totals for each month for the period 1981-1995 are shown in the lower panel.

summer and early autumn are prone to error due to difficulties in correctly predicting the movement of tropical storms into the extratropical westerlies and subsequent reintensifications over the eastern Atlantic. More generally, the zonality of the flow in early autumn may lead to low anomaly correlations associated with large phase errors in the forecasts of rapidly-moving systems.

5. Differences between consecutive forecasts

The rms difference, D_j, between consecutive forecasts verifying on the same day is given by

$$D_j = \sqrt{\overline{(f_j - f_{j-1})^2}}$$

Here f_{j-1} is the $(j-1)$-day forecast from the analysis made $(j-1)$ days prior to verification time, and f_j is the j-day forecast from the analysis made j days prior to verification time.

Lorenz (1982) made the observation that f_j can equally be regarded as the $(j-1)$-day forecast starting from the one-day forecast valid $(j-1)$ days previously. D_j, for $j > 1$, thus measures the divergence of two model runs which start from relatively small initial differences D_1 (or equivalently, E_1). Lorenz argued that if the forecast model, though imperfect, was realistic enough for small differences in initial conditions to amplify at a rate close to that at which separate but similar atmospheric states diverge, then D_j provides an estimate of the lower limit of rms error beyond day one (the "perfect-model" error), for an unchanged one-day forecast error. Rms error could only be reduced beyond this limit by reducing the one-day forecast error. The accuracy of this estimate evidently depends on the realism of the forecast model's representation of growth processes.

The asymptotic limit of rms forecast differences can be derived in the same way as for rms error:

$$\begin{aligned}(D_j)^2 &= \overline{((f_j - c) - (f_{j-1} - c))^2} \\ &= \overline{(f_j - c)^2} + \overline{(f_{j-1} - c)^2} - 2\overline{(f_j - c)(f_{j-1} - c)} \\ &= (A_j)^2 + (A_{j-1})^2 - 2\overline{(f_j - c)(f_{j-1} - c)} \qquad (6)\end{aligned}$$

As the forecast range and sample size increase:

$$\overline{(f_j - c)(f_{j-1} - c)} \to 0$$

assuming no systematic forecast error or error in the estimation of climatology. Then $D_j \to \sqrt{2}\, A_j$. For a perfect model $A_j \to A_a$ and D_j tends to the same limit $\sqrt{2}\, A_a$ as E_j. If the model loses variance about the climatological mean, D_j tends to a limit less than E_j. The difference between the limiting value of D_j^2 and the perfect-model limit $2 A_a^2$ is twice as large as the difference between the limiting value of E_j^2 and the perfect-model limit.

Plots of E_j and D_j are shown in the upper panel of Fig. 9 for Winter 1995 and the extratropical northern hemisphere. The lower panel shows the corresponding anomaly correlations. These results indicate a significant scope for forecast improvement beyond day one due to model improvements alone. The potential skill levels indicated by the dashed curves in Fig. 9 must be regarded as underestimates of the full benefit to be gained from model improvement. Better models should lead to a reduction of one-day error, both because of lower error growth during the first day of the forecast and because of improved initial conditions resulting from use of a better model in data assimilation. The anomaly correlation indicates a larger scope for improvement in the medium range, as measured for example by the forecast range at which the difference score attains the value that the error score reaches at day seven. The correlation of consecutive forecast anomalies may be related approximately to the mean square difference, D_j^2, normalized by $(A_j)^2 + (A_{j-1})^2$, in a way similar to that demonstrated in section 2.3 for forecast error. It may thus be less affected by the model's overestimation of variance in Winter 1995, which causes D_j to increase by more than E_j.

D_j is a measure of the consistency of successive forecasts. Numerical forecasts which are consistent from day to day (low values of D_j) are understandably valued by the forecaster as they make it easier for him or her to convey a consistent message to the end-user. However, inconsistency is inevitable if the numerical forecasts are produced from a model which provides a realistic simulation of atmospheric variability. A forecast for a particular day can only be more accurate than an earlier forecast for the same day if the two forecasts are to some degree different and thus inconsistent. A model which underestimates variance or the growth-rate of small perturbations will give more consistent forecasts than a perfect model (a daily forecast of climatology is perfectly consistent), but may give less accurate initial analyses and short-range forecasts, and will be less suitable for use in ensemble prediction where the aim is to provide reliable probabilistic forecasts. Given a realistic model, the sound way to improve the consistency of daily deterministic medium-range forecasts is to reduce the one-day forecast error.

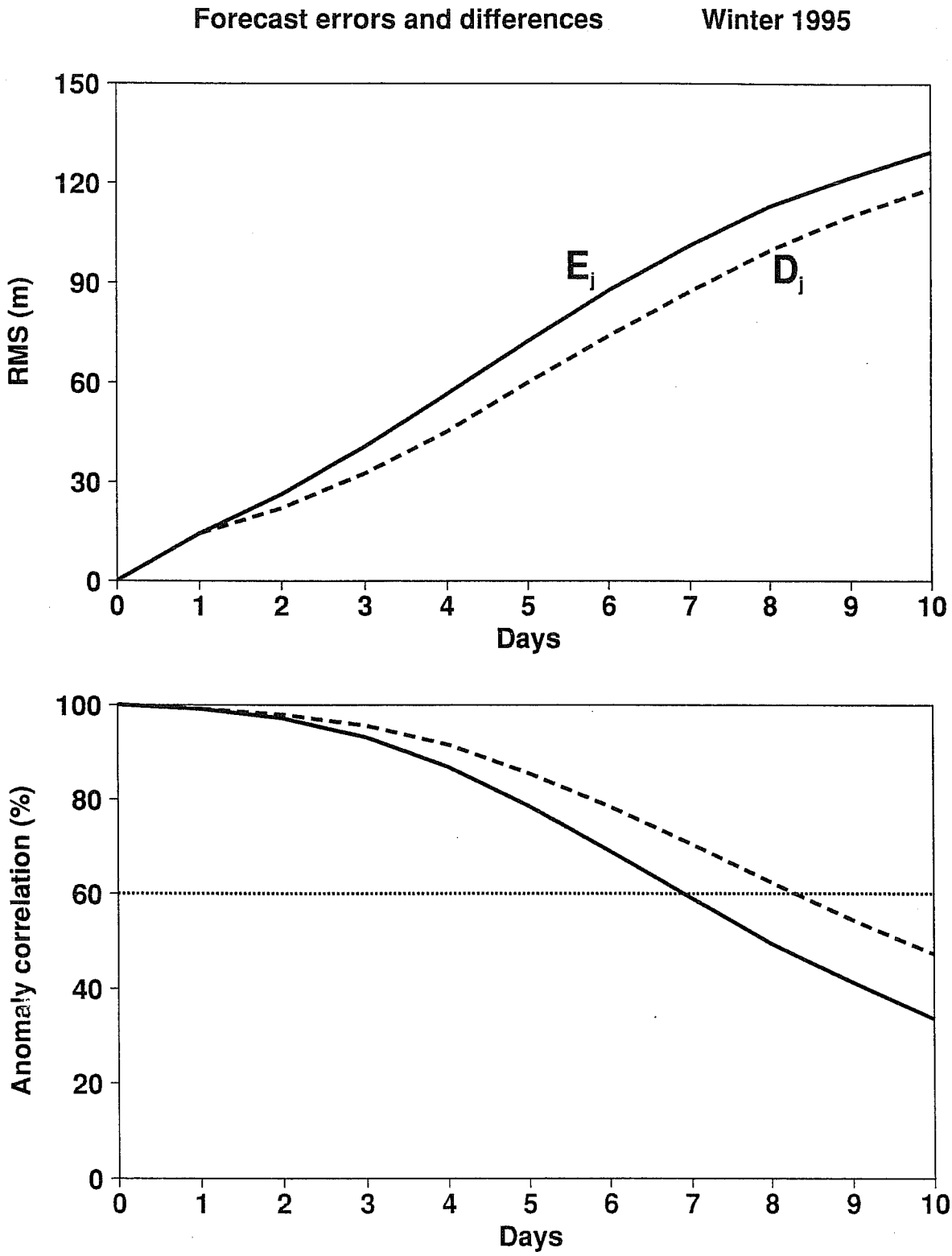

Fig. 9 Objective measures of forecast error and of the difference between consecutive forecasts valid at the same time, for 500hPa height forecasts for the extratropical northern hemisphere averaged for Winter 1995.

Upper: Rms errors (solid) and differences (dashed);
Lower: Correlations of forecast and analysed anomalies (solid) and of consecutive forecast anomalies (dashed).

6. Comparison of forecasting models or systems.

In this section we examine rms errors and differences, and some related diagnostics, from different forecasting models or systems. Specifically, we compare the recent operational ECMWF T213L31 forecasts with the operational forecasts of the United Kingdom Meteorological Office (UKMO), with the T63L19 control forecasts from the ECMWF EPS, and with the ECMWF forecasts produced operationally in 1986. The aims are to gain confidence in the predictability estimates derived from the current performance of the T213L31 forecasting system, to quantify differences between the deterministic and EPS forecast models, and to understand better the way operational skill scores have evolved over the years.

Two types of map will be presented to provide guidance as to interpretation of the rms error and difference curves. The first of these shows the mean square anomalies of analyses and forecasts, the quantities $(A_a)^2$ and $(A_j)^2$ discussed earlier when considering asymptotic limits. Similarity between these variances about climatology gives confidence that rms errors and differences are not evolving towards asymptotic limits that are too low (or high) because of an underestimation (or overestimation) of variance by the forecast model.

The second diagnostic is the mean square change from one day to the next of the analyses and forecasts. For the forecasts, this is the average over all forecasts of the squared change from one day of the forecast range to the next for each individual forecast. This is not the same as the change in consecutive forecasts valid for a particular day, the area-average of which is D_j^2 . This diagnostic measures the capability of the forecast model to simulate realistic day-to-day changes of the height field. These changes may result either from the movement or from the growth or decay of weather systems. Good agreement between analyses and forecasts gives some confidence in model estimates of the rate of divergence of small initial differences.

(i) Comparison of ECMWF and UKMO forecasts

Fig. 10 shows rms errors and differences between consecutive daily forecasts, E_j and D_j , from the ECMWF and UKMO forecasting systems. All calculations here are carried out on the 2.5^ox5^o grid on which UKMO data are provided to ECMWF. The domain is the extratropical northern hemisphere and results are shown for Winter and Summer[2] 1995. The forecast range is out to day six, the range for which UKMO data are available.

The ECMWF forecasts have distinctly lower rms errors for the winter period, but the growth of the differences between consecutive forecasts is remarkably similar for the two systems in this season. An exception occurs between days one and two, when the difference grows more slowly in the UKMO system (though from a higher day-one value), suggesting that initial adjustment ("spin-up") processes differ between the two systems. Rms errors are also lower for the ECMWF system in summer, but rms differences are higher then. The summer

[2] June, July and August

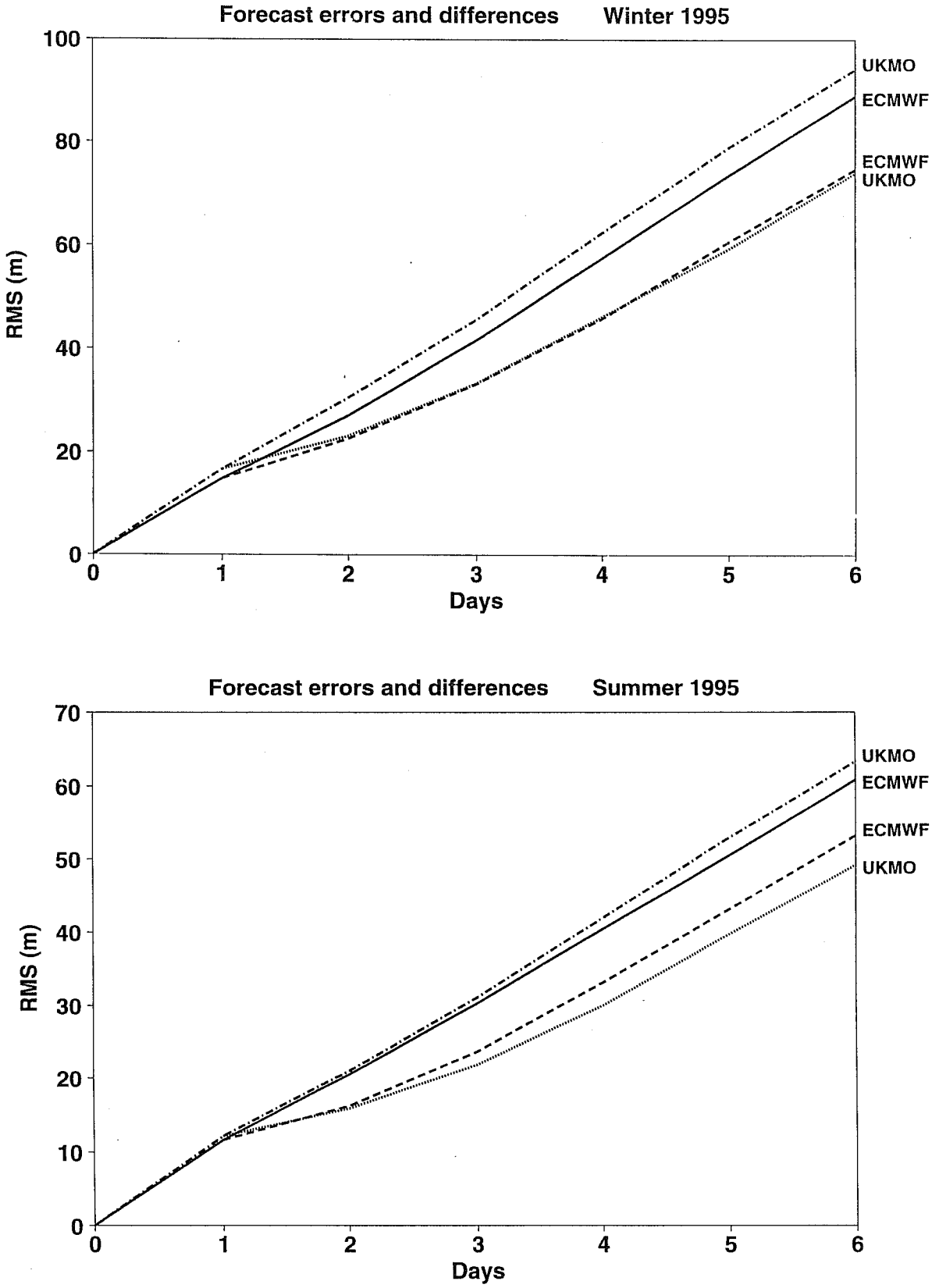

Fig. 10 Rms errors and differences between consecutive forecasts from the ECMWF (solid, dashed) and UK Meteorological Office (dash-dotted, dotted) systems. Results are for 500hPa height over the extratropical northern hemisphere for Winter 1995 (upper) and Summer 1995 (lower).

ECMWF forecasts are, according to these measures, more accurate but more inconsistent from day to day. The rms ECMWF and UKMO differences are nevertheless much closer than are the other pairs of differences compared later.

Maps of the variance of analyses and day-6 forecasts are presented in Fig. 11 for Winter 1995. The ECMWF and UKMO analyses are evidently in quite good agreement, with slightly higher variances from the ECMWF system. The two sets of forecasts give variances that are generally similar both to each other and to the analyses. Close inspection reveals some points in favour of one forecasting system and some in favour of the other. Perhaps the most striking feature is the similarity in the two models' overestimation of variance around the dateline at high latitudes and over eastern Canada.

Corresponding maps for Summer 1995 are presented in Fig. 12. As for winter, the two analyses give similar variances, with slightly higher values from the ECMWF system. Both ECMWF and UKMO forecasts underestimate variance over the Atlantic and Pacific in this season, though the discrepancy is less for the ECMWF forecasts. The ECMWF forecasts appear to overestimate variance near the pole, though there is some uncertainty as to the analyzed values in this data-sparse region.

Mean square daily changes from the analyses and between days five and six of the forecasts are shown in Fig. 13 for Winter 1995. Analyzed changes are somewhat larger in the ECMWF than in the UKMO analyses. The changes in the ECMWF forecasts match the analyzed changes particularly well. For Summer 1995 (Fig. 14), change is generally underestimated in both sets of forecasts, though by less in the ECMWF system.

These diagnostics suggest that interpretation of the rms forecast errors and differences is not seriously impeded by deficiencies of the forecast models. This is particularly so in winter, and for the ECMWF system. In summer, the larger growth of rms differences (or larger inconsistency) of the ECMWF forecasts appears to be associated with the higher (and generally more realistic) variance and rates of daily change of the ECMWF model. The ECMWF model is nevertheless less active than the atmosphere in summer, and may thus give too optimistic a picture of the scope for future gains in predictive skill.

It is also of interest to compare the day-to-day performance of the ECMWF and UKMO forecasting systems. This is illustrated in Fig. 15, which plots for each day of 1995 the anomaly correlation for Europe of the operational day-six forecasts. ECMWF results are joined by solid lines and UKMO results by dashed lines. Over the year, the mean day-six anomaly correlations were 69.2% for ECMWF and 65.8% for UKMO. The mean difference of less than 3.5% is clearly small compared to day-to-day fluctuations in the accuracy of one or other of the sets of forecasts. On some occasions both forecasting systems produced a very poor forecast (11 July is an example) while on others just one of the systems did poorly (19 March, 9 September and 5 November are the dates with the largest differences). This variability of forecast quality has for long been a major concern to forecasters as it limits the confidence they have in the medium-range forecast. This was a major factor behind the development of ensemble prediction.

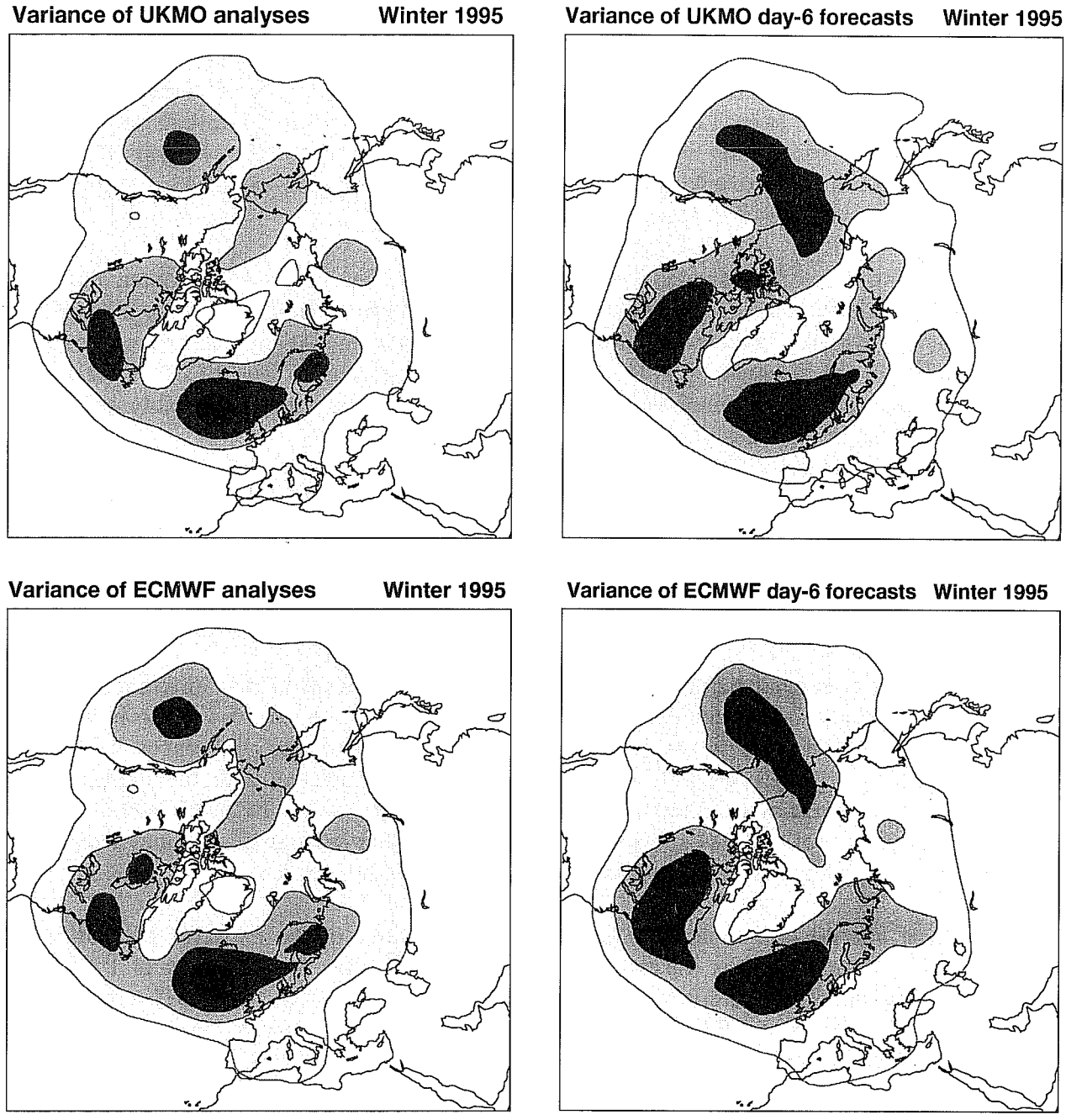

Fig. 11 Mean square anomaly of UK Meteorological Office (upper) and ECMWF (lower) analyses (left) and six-day forecasts (right) of the 500hPa height field for Winter 1995. The contour interval is 10000m^2, and shading denotes values above 10000m^2.

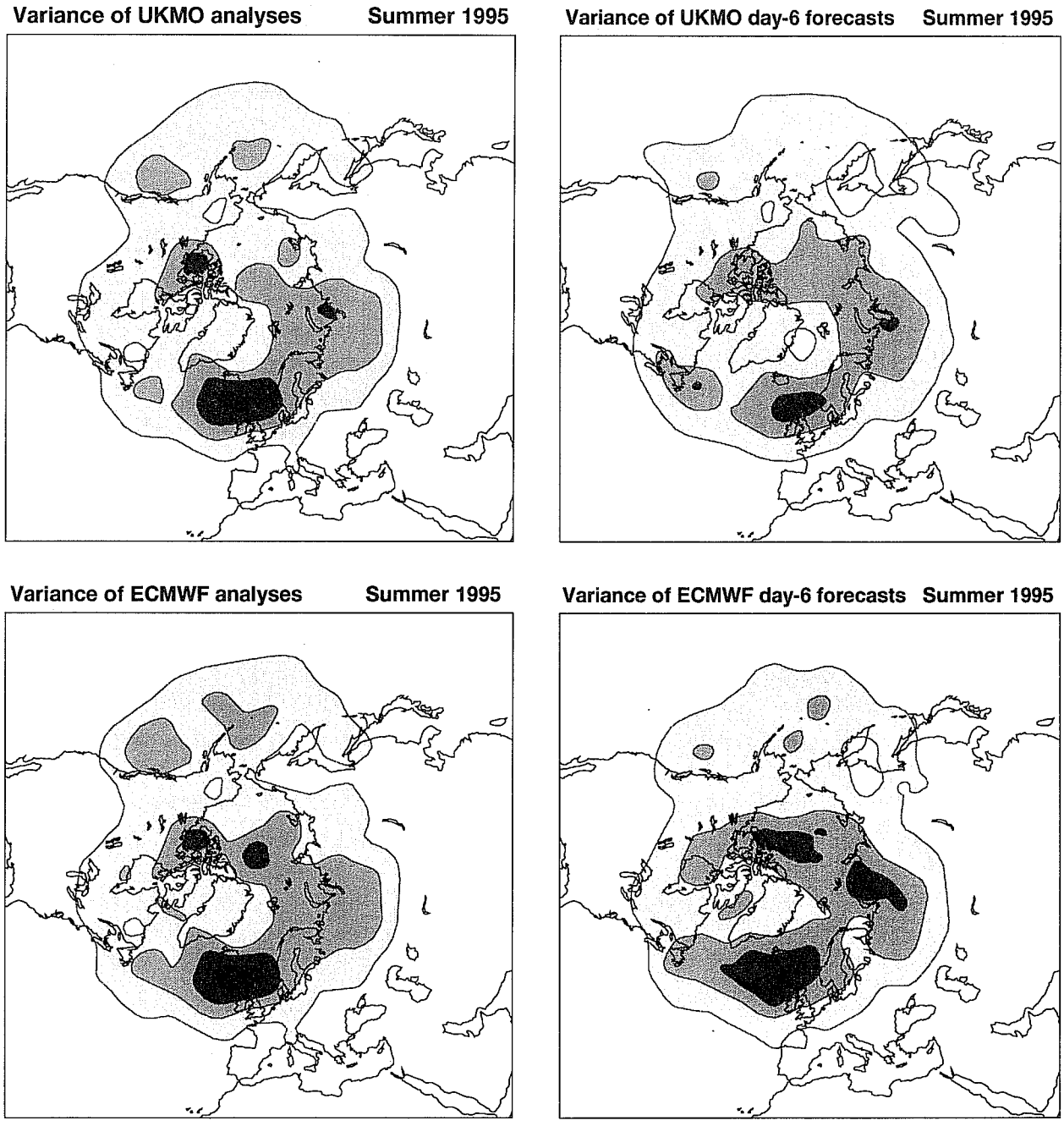

Fig. 12 Mean square anomaly of UK Meteorological Office (upper) and ECMWF (lower) analyses (left) and six-day forecasts (right) of the 500hPa height field for Summer 1995. The contour interval is 5000m^2, and shading denotes values above 5000m^2.

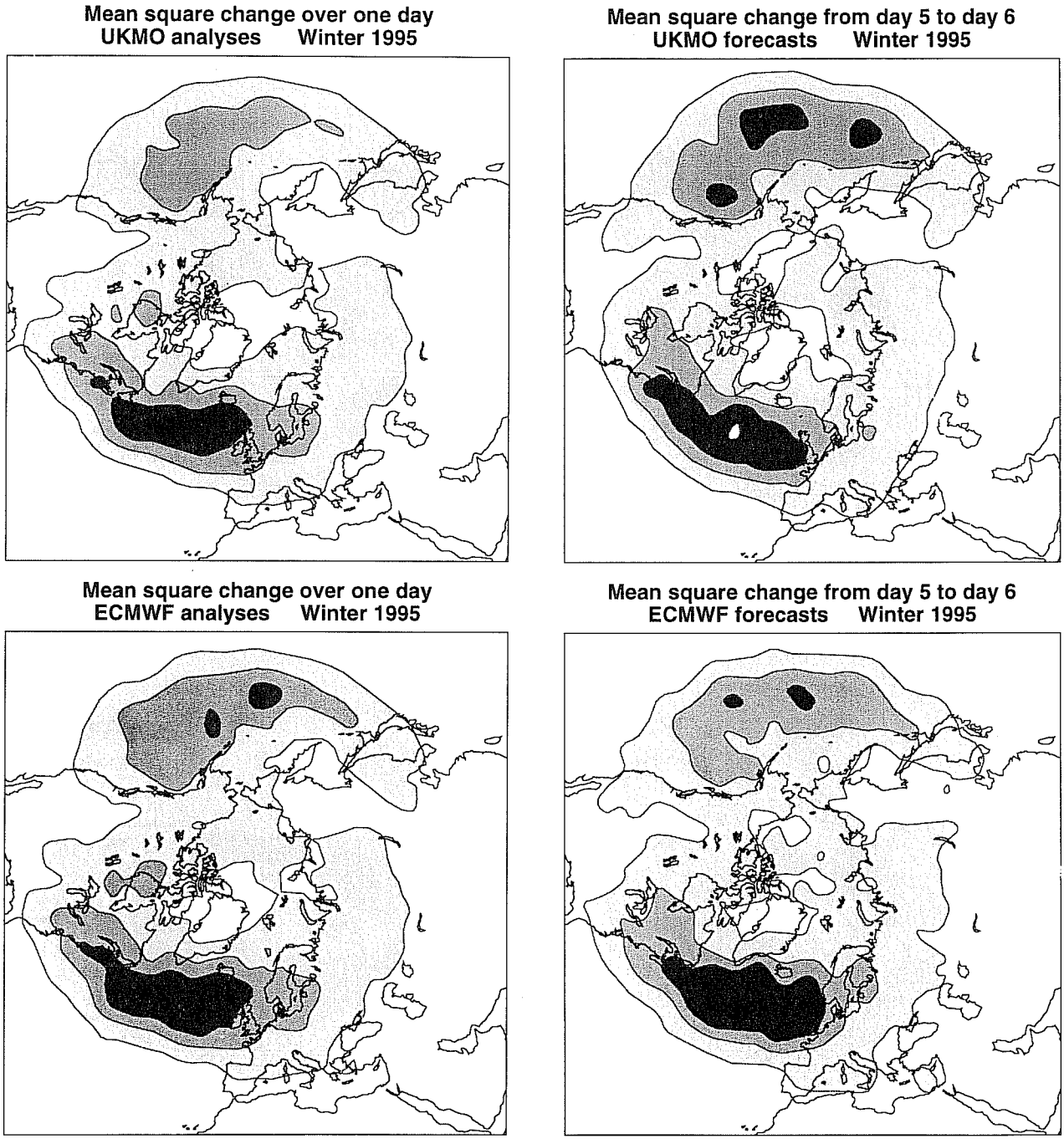

Fig. 13 Mean square change from one day to the next of UK Meteorological Office (upper) and ECMWF (lower) analyses (left) and from day five to day six of the corresponding forecasts (right) for the 500hPa height field for Winter 1995. The contour interval is 5000m^2, and shading denotes values above 5000m^2.

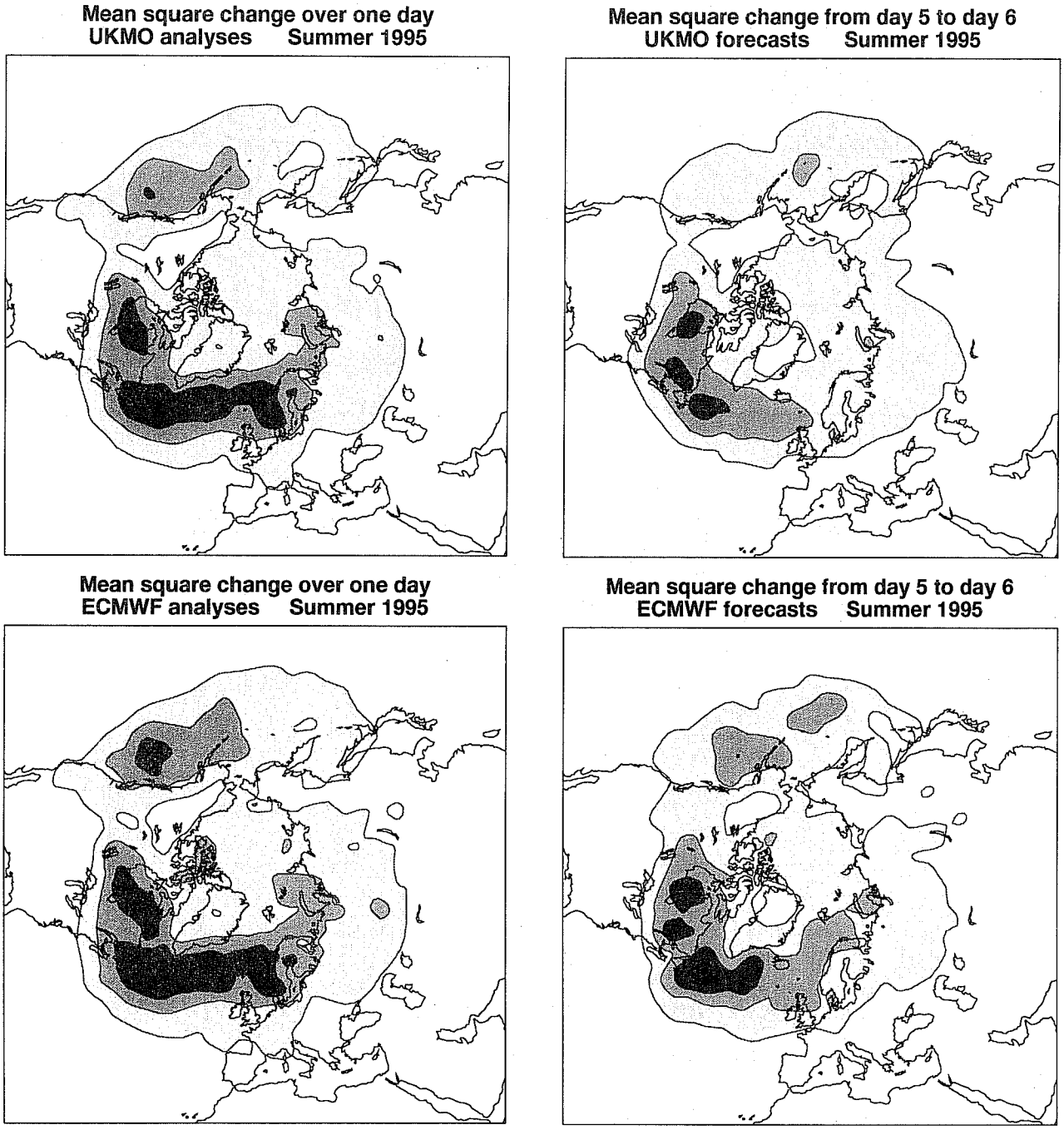

Fig. 14 Mean square change from one day to the next of UK Meteorological Office (upper) and ECMWF (lower) analyses (left) and from day five to day six of the corresponding forecasts (right) for the 500hPa height field for Summer 1995. The contour interval is 2000m^2, and shading denotes values above 2000m^2.

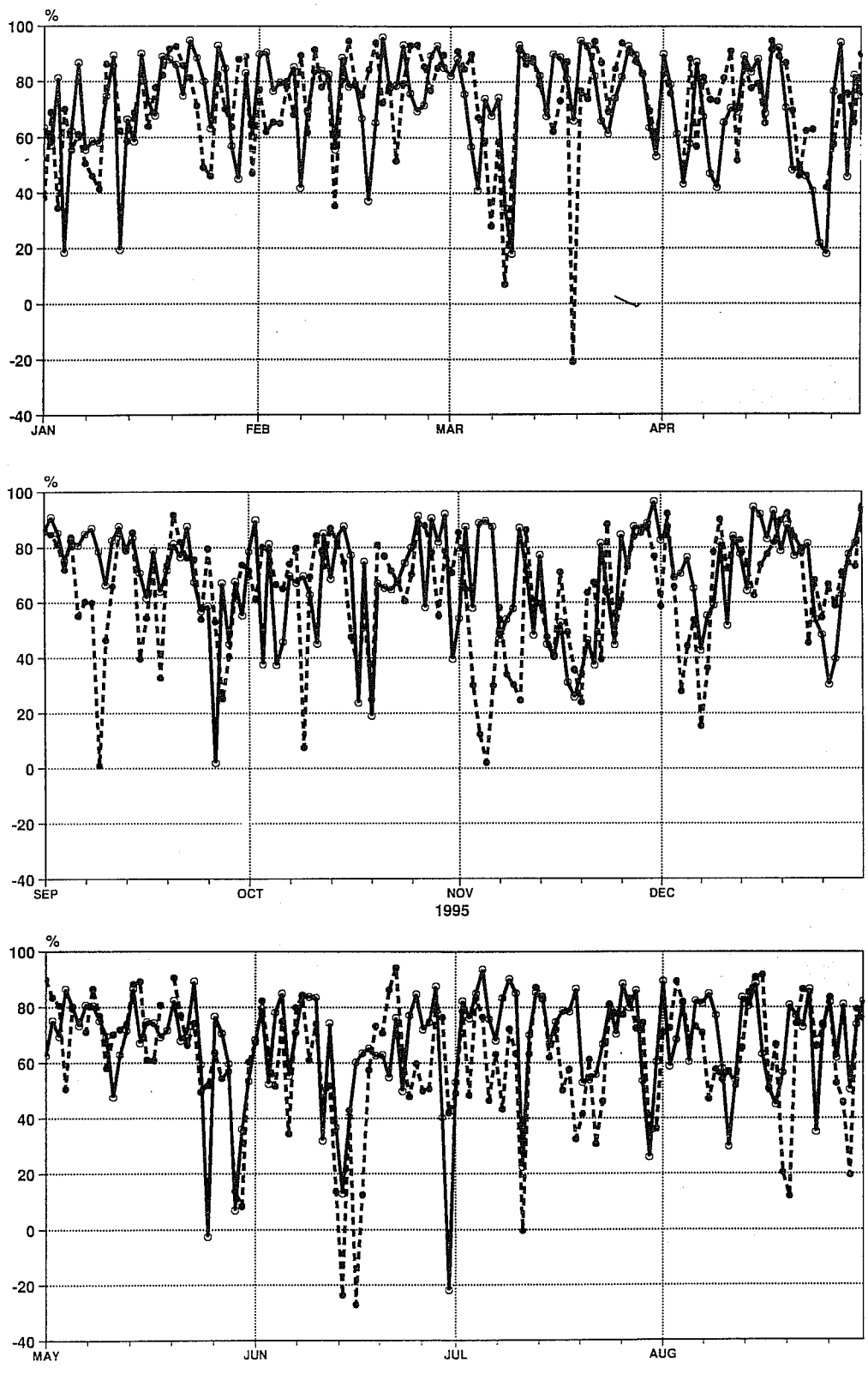

Fig. 15 Daily values of the 500hPa height anomaly correlation of six-day forecasts for Europe throughout 1995. Solid lines denote ECMWF forecasts and dashed lines forecasts from the UK Meteorological Office.

(ii) Comparison of T213L31 and T63L19 forecasts

Rms errors and differences of the T213L31 and T63L19 forecasts are compared in Fig. 16 for Winter and Summer 1995. The T63L19 forecasts start from truncated and interpolated versions of the initial analyses used for the T213L31 forecasts. The rms errors of the higher resolution forecasts are lower in the first half of the forecast range, more so in summer than in winter, but higher in the second half of the range. Rms differences grow substantially more slowly in the T63L19 forecasts, particularly in summer.

Diagnostics show that the smaller growth of differences (higher consistency) of the T63L19 forecasts arises because of a substantial underestimation of atmospheric activity by this version of the ECMWF model. This is seen in plots of both variance and daily rates of change. Fig. 17 illustrates this for the mean square change from day five to day six of T213L31 and T63L19 forecasts for the winter and summer periods.

Underprediction of wave amplitudes gives larger rms errors if phases are accurately predicted, but smaller rms errors if phase errors are significant. This can be illustrated by considering the contribution to global mean square error from a particular spherical harmonic component. If the relationship between the forecast value of the spectral coefficient, f_n^m, and the analyzed value, a_n^m, is written in the form $f_n^m = (1-\alpha)\exp(i\beta)\,a_n^m$, for real α and β, then the contribution to the global mean square error can be written in the form

$$\left(\alpha^2 + 4(1-\alpha)\sin^2(\beta/2)\right)|a_n^m|^2$$

For sufficiently small β (small phase error), this quantity increases as α (amplitude underprediction) increases. However, it decreases as α increases from zero if β is sufficiently large.

The rms errors shown in Fig. 16 are consistent with the above results. In the early part of the forecast range, when flow patterns are predicted with relatively high accuracy, the more-active T213L31 model produces the forecasts with the lower rms errors. The forecasts from the less-active T63L19 model have on average the lower rms errors later in the range, by which time phase errors will usually have become large, and more fundamental errors in the flow pattern may have developed. The lack of activity of the T63L19 model also means that errors in the initial analysis will typically amplify less rapidly than in the T213L31 model. The much lower growth rate of differences between consecutive T63L19 forecasts is indicative of this.

Of immediate concern for the ECMWF EPS is the implication that this deficiency of the T63L19 model will result in an inadequate spread of the ensemble forecasts, and an underestimation of probabilities of extreme events. Use of initial perturbations that are larger than would otherwise be needed can provide some compensation, but cannot be entirely satisfactory. This provided one reason for the acquisition of more powerful computing facilities at ECMWF, which will enable a significant increase in the resolution of the model used for ensemble forecasting.

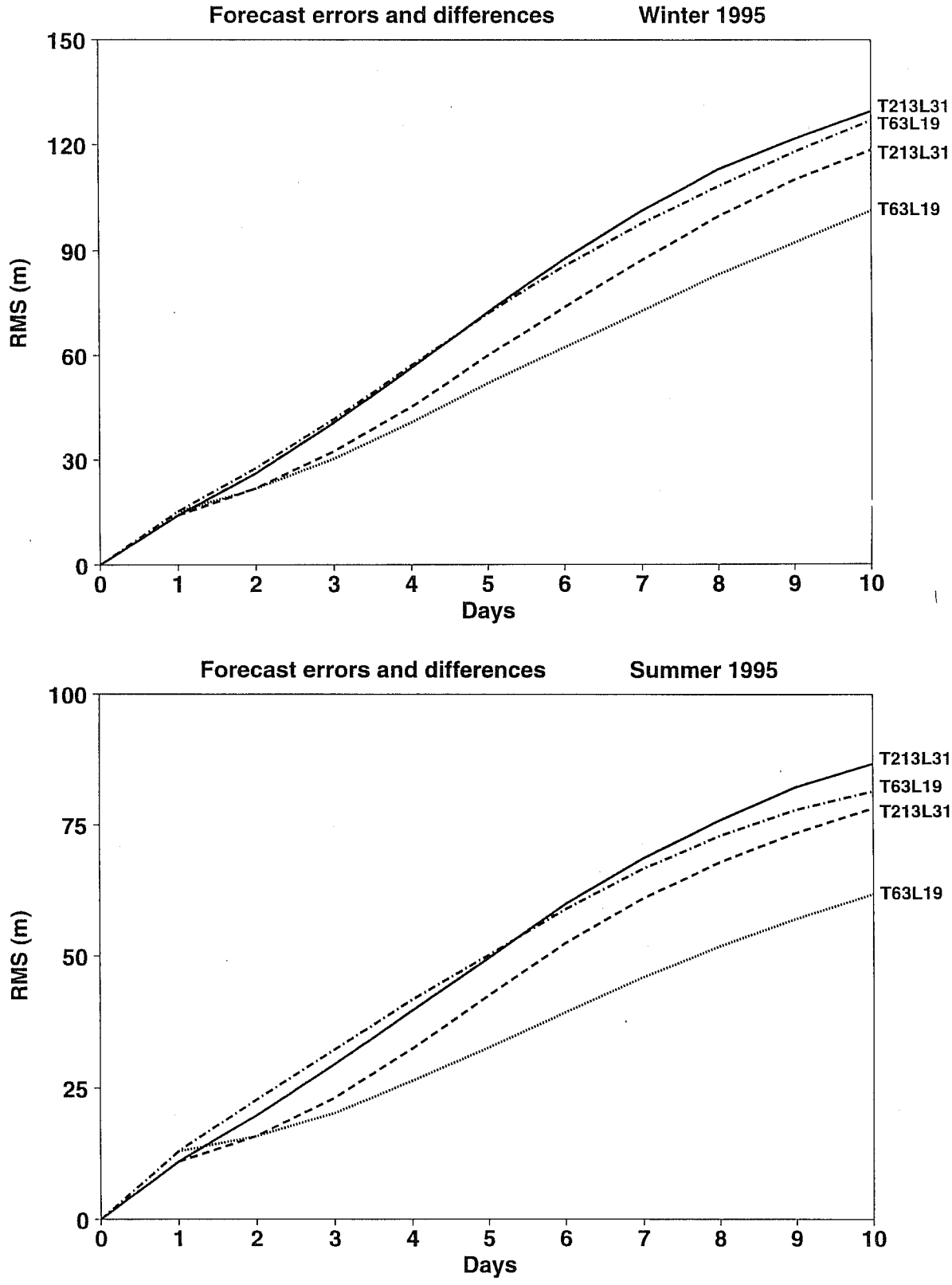

Fig. 16 Rms errors and differences between consecutive forecasts from T213L31 (solid, dashed) and T63L19 (dash-dotted, dotted) versions of the ECMWF model. Results are for 500hPa height over the extratropical northern hemisphere for Winter 1995 (upper) and Summer 1995 (lower).

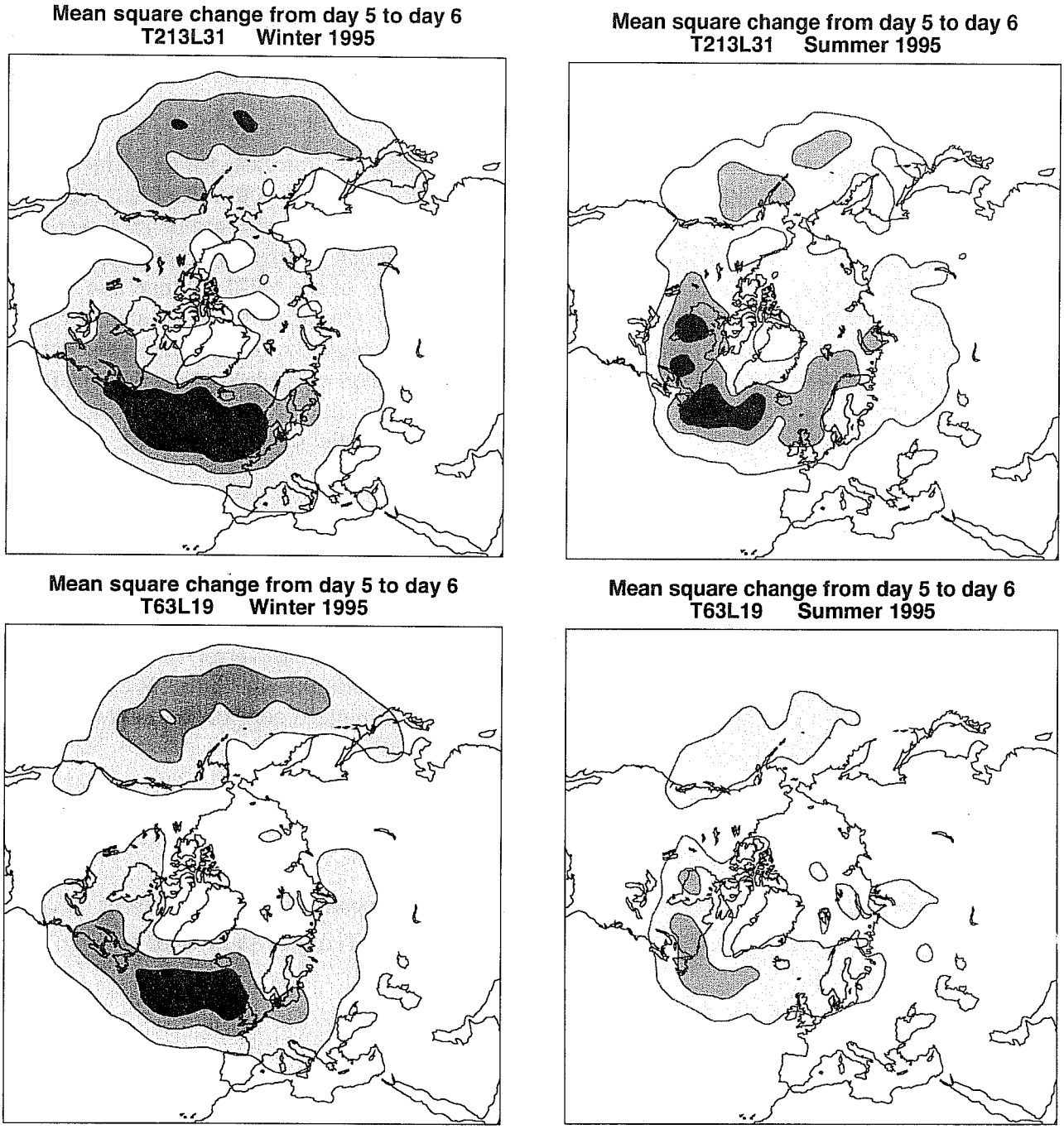

Fig. 17 Mean square change from day five to day six of T213L31 (upper) and T63L19 (lower) 500hPa height forecasts for Winter 1995 (left, contour interval 5000m^2) and Summer 1995 (right, contour interval 2000m^2).

(iii) Comparison of 1995 and 1986 forecasts

We have already seen how anomaly correlation scores on average showed little or no improvement in the medium range from 1986 to 1992. The operational forecast model used T106L16 resolution in Winter 1986 and T106L19 in Summer 1986. Changes in resolution and physical parametrization introduced since that time have significantly increased the activity of the model.

Fig. 18 compares rms errors and differences between consecutive forecasts for the winters and summers of 1986 and 1995. The error curves here differ because of both model and analysis differences, and because of differences in synoptic situation. The winter errors are lower for 1995 than for 1986 at all forecast ranges, although the reduction in error is largest in the short range. The summer errors for 1986 are lower near the end of the ten-day range and much more clearly appear to be approaching saturation by day ten, as is the case also for the differences between consecutive forecasts. Earlier in the range, the differences grow much more slowly in 1986 than in 1995, especially in summer. Maps of mean square daily changes for 1986 are presented in Fig. 19. They clearly indicate a lack of variability in the 1986 versions of the forecast model. This is particularly marked for the summer forecasts, as might be expected from Fig. 18.

The large reduction in short-range rms forecast errors between 1986 and 1995 is presumably due to both lower analysis errors in 1995 and an operational model which could simulate change more realistically in 1995. This appears not to have been fully reflected in reduced medium-range forecast errors because analysis error amplified less rapidly in 1986 due to the model's unrealistic representation of growth processes, and because error saturated at an artificially low level in 1986 due to the model's underestimation of variance. We examine this further in the following section in the context of a simple model of the evolution of forecast error.

7. Error-growth modelling

Lorenz(1982) proposed the use of a simple model of the dependence of the rms error E of a sequence of "perfect-model" forecasts on their range t . The growth of error is given by

$$\frac{\partial E}{\partial t} = \alpha E \left(1 - \frac{E}{E_\infty} \right) \qquad (7)$$

Parameters of the model are the rate α at which small error grows early in the range, and the asymptotic level E_∞ at which error saturates. The doubling time of small errors t_d is $(\ln 2)/\alpha$.

A corresponding anomaly correlation C can be defined from equation (5):

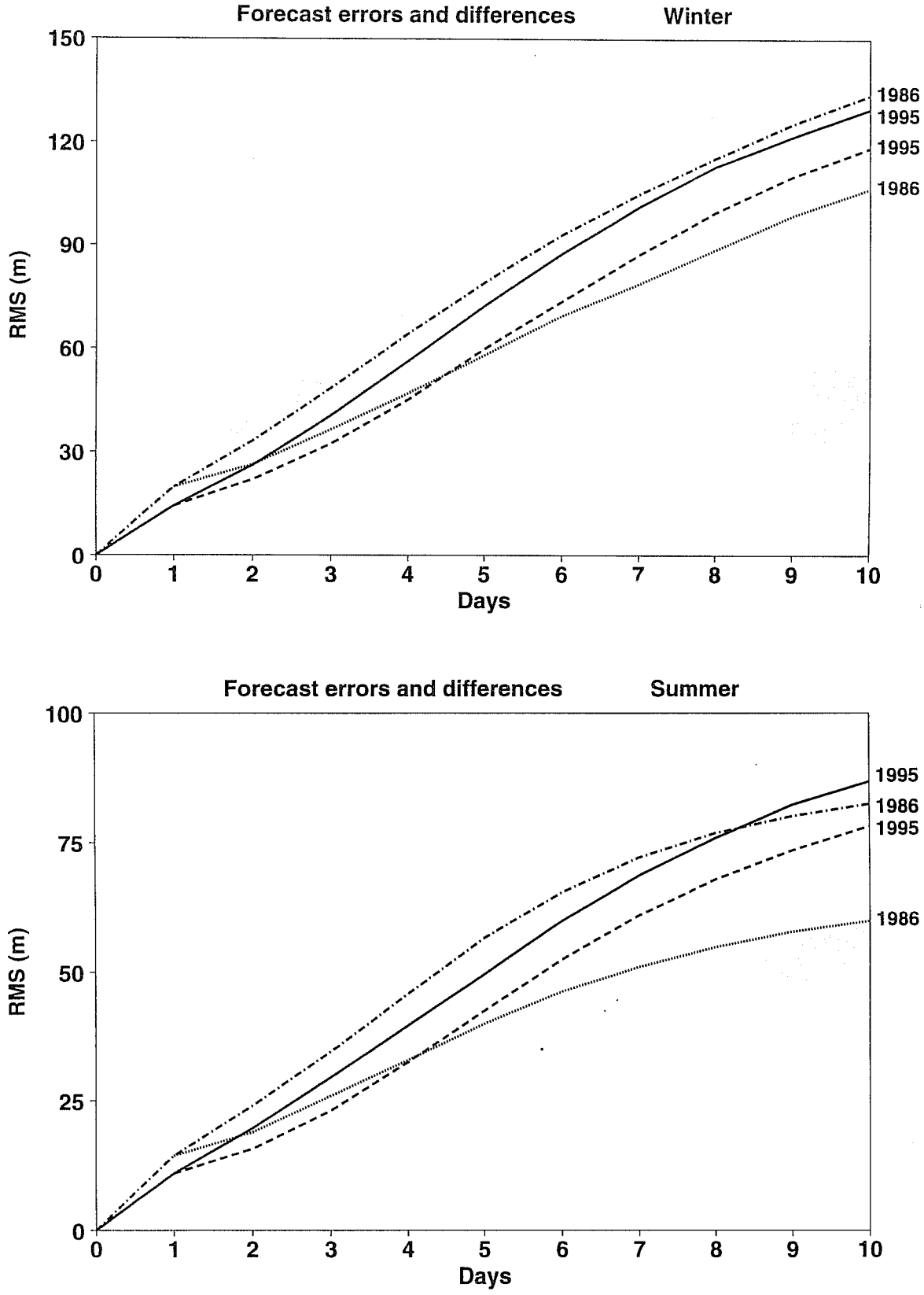

Fig. 18 Rms errors and differences between consecutive forecasts from the 1995 (solid, dashed) and 1986 (dash-dotted, dotted) versions of the ECMWF forecasting system. Results are for 500hPa height over the extratropical northern hemisphere for Winter (upper) and Summer (lower).

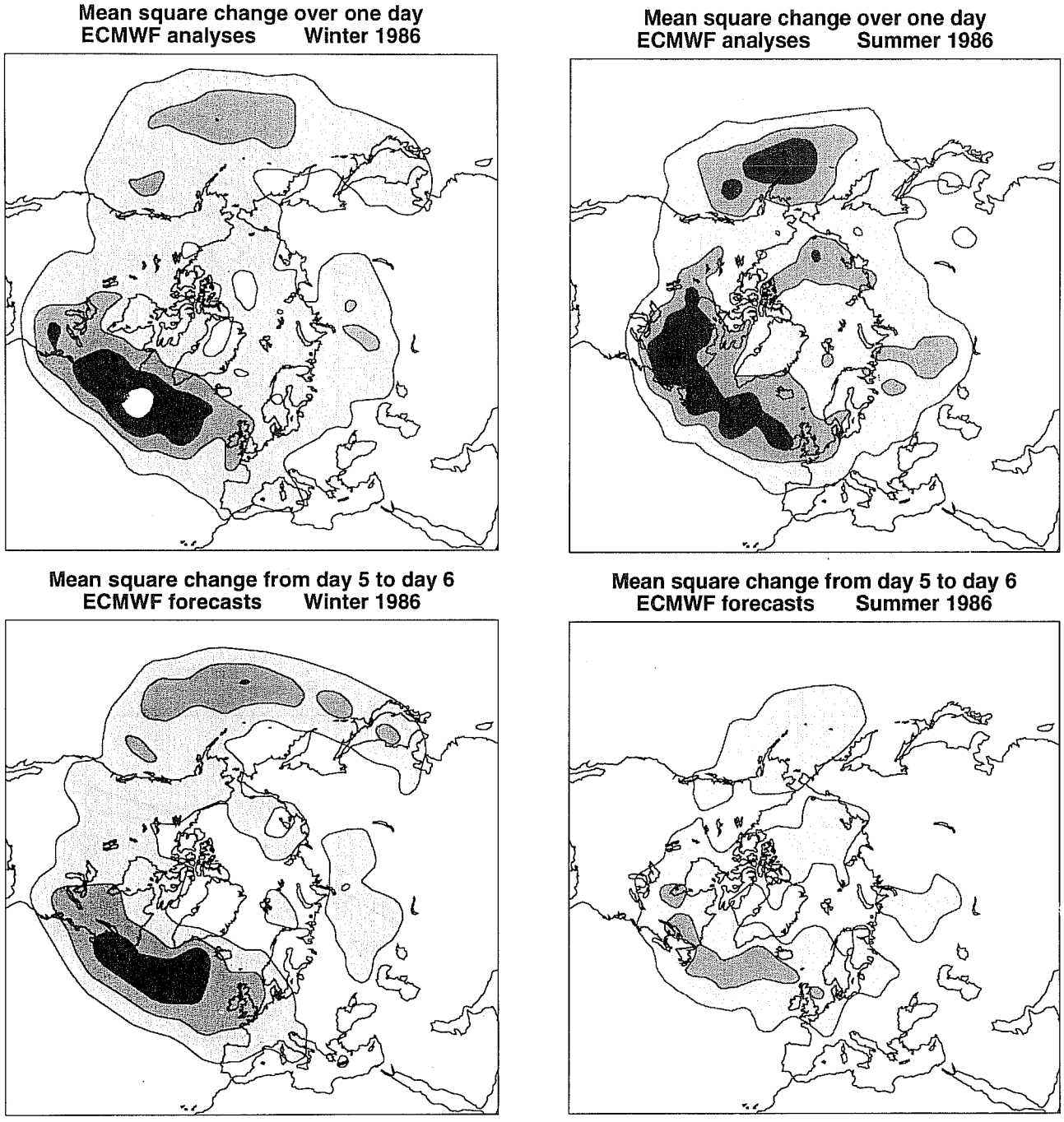

Fig. 19 Mean square change from one day to the next of ECMWF 500hPa height analyses for Winter 1986 (upper left, contour interval 5000m^2, contours up to a maximum value of 25000m^2) and Summer 1986 (upper right, contour interval 2000m^2). Corresponding plots for the changes from day five to day six of the 1986 operational forecasts are shown in the lower panels.

$$C = 1 - \frac{E^2}{E_\infty^2} \tag{8}$$

It satisfies the equation

$$\frac{\partial C}{\partial t} = -2\alpha(1-C)\left(1-\sqrt{(1-C)}\right) \tag{9}$$

This "perfect-model" anomaly correlation depends only on the growth rate of small errors α and on the initial value of normalized rms error E/E_∞.

The parameters α and E_∞ have been determined as in Lorenz' study by a weighted least-square fit of the rms differences, D_j, between consecutive real-model forecasts. Equation (7) is written in the finite-difference form

$$\frac{\Delta E}{\Delta t} = \alpha \overline{E}\left(1 - \frac{\overline{E}}{E_\infty}\right)$$

Δt is one day, and the values ΔE_j and $\overline{E}_j$ to be fitted are given by $\Delta E_j = D_{j+1} - D_j$ and $\overline{E}_j = \frac{1}{2}(D_{j+1} + D_j)$, for j=1, 2, 3, ... 9.

	t_d (days)	E_∞ (m)	t_d (days)	E_∞ (m)	t_d (days)	E_∞ (m)	t_d (days)	E_∞ (m)
					Summer 94		Autumn 94	
T213 L31					1.5	86	1.4	127
T63 L19					2.0	77	1.7	112
	Winter 95		Spring 95		Summer 95		Autumn 95	
T213 L31	1.5	132	1.5	125	1.5	86	1.5	128
T63 L19	1.8	120	1.9	117	2.1	89	1.8	109

Table 1 Doubling times (t_d) and asymptotic limits (E_∞) of "perfect-model" error growth, from the Lorenz model, for the 500hPa height field computed over the extratropical Northern Hemisphere for Summer and Autumn 1994, and for each season of 1995, from T213L31 and T63L19 forecasts.

Table 1 presents values of t_d and E_∞ for T213L31 and T63L19 forecasts for the seasons for which results are available from the daily control runs of the ECMWF EPS. The small-error doubling times are significantly longer for the T63L19 forecasts than for the T213L31 forecasts. This is especially so in summer, but in the other seasons the T63L19 doubling times are at least 20% longer. Error-saturation levels are generally lower for

T63L19; Summer 1995 is the exception. The particularly low growth rates of T63L19 in summer are consistent with diagnosis of EPS performance, which shows that spread of the ensemble is particularly weak in summer, as illustrated by Buizza(1996) in this volume.

The evolution since 1981 of various parameters relating to error growth can be seen in Table 2. There has clearly been a significant reduction in both small-error doubling times and one-day rms errors over the past fifteen years. Doubling times have come down from about two days or a little under to about one and a half days, and the one-day rms errors have been reduced by about 40%. The situation as regards asymptotic limits is less simple.

Two "perfect-model" estimates of error saturation levels are presented in Table 2. One is the asymptotic limit E_∞ from the Lorenz model of error growth. As indicated above, this is determined from the differences between consecutive forecasts, and is likely to be an underestimate if the forecast model underestimates the variance of the atmosphere. The second is the "perfect-model" limit determined from the rms anomaly of the analyses, $\sqrt{2}A_a$. The analyzed level of the variance about climatology may be weakly dependent on errors in the level of variance simulated by the model used in data assimilation, but the variations in $\sqrt{2}A_a$ from year to year seen in Table 2 are not obviously linked to variations in model characteristics, and presumably represent natural interannual fluctuations in the level of variance. The limit $\sqrt{2}A_a$ shows a much weaker interannual variability in summer than in other seasons, even when allowance is made for the annual cycle in the level of variance. Also shown in Table 2 is the quantity $\sqrt{2}A_{10}$ derived from the variance of the day-ten forecasts. These values are included to indicate the extent to which the forecast model at the time overestimated or underestimated variance in the mean over the extratropical hemisphere.

Values of E_∞ and $\sqrt{2}A_{10}$ are generally lowest for the years 1985 to 1987. Model changes prior to this time had increased error growth rates, but reduced variance. Since then, both error growth rates and variance have been increased. Comparing $\sqrt{2}A_a$ and $\sqrt{2}A_{10}$ indicates that the forecast model currently overestimates variance in all seasons, although for summer the difference between $\sqrt{2}A_a$ and $\sqrt{2}A_{10}$ is marginal. We have already seen in Fig. 12 that variance at day 6 was overestimated in polar regions and slightly underestimated in middle latitudes in summer 1995. The same is true at day 10, but quite different patterns were found in 1993 and 1994, with underestimation of variance in polar regions. Considerable interannual variations in the regions of overestimation are also found for other seasons, making generalization difficult. It is, however, almost certain that the relatively large overestimation of variance in Autumn 1991 and in 1992 was due to previously mentioned (and since corrected) problems with the performance of the T213L31 semi-Lagrangian model introduced operationally in September 1991.

	t_d (days)	E_1 (m)	E_∞ (m)	$\sqrt{2}A_a$ (m)	$\sqrt{2}A_{10}$ (m)	t_d (days)	E_1 (m)	E_∞ (m)	$\sqrt{2}A_a$ (m)	$\sqrt{2}A_{10}$ (m)
		Winter					Spring			
1981	1.8	25	135	158	174	1.8	21	122	138	155
1982	1.9	23	139	164	176	1.9	21	129	132	153
1983	2.0	23	138	154	167	1.9	21	124	134	159
1984	1.8	22	117	161	164	1.9	19	117	135	152
1985	1.8	21	112	171	164	1.8	19	100	134	140
1986	1.9	20	126	158	159	1.8	17	106	140	137
1987	1.7	19	117	154	161	1.7	16	109	133	138
1988	1.6	18	121	149	163	1.6	16	120	132	145
1989	1.6	17	134	161	177	1.5	15	116	139	143
1990	1.7	16	131	159	167	1.6	15	125	142	150
1991	1.7	16	135	160	170	1.6	14	128	137	148
1992	1.5	16	148	150	177	1.5	14	138	135	158
1993	1.6	15	148	159	177	1.5	13	126	139	150
1994	1.6	14	151	150	168	1.5	13	125	128	144
1995	1.5	14	132	154	163	1.5	13	125	145	154
		Summer					Autumn			
1981	1.9	18	81	98	104	1.7	20	123	133	143
1982	2.1	19	92	96	116	1.7	20	119	124	141
1983	1.9	17	82	98	107	1.8	18	107	126	135
1984	1.9	17	79	95	107	1.9	19	109	130	138
1985	1.7	15	65	94	86	1.6	17	117	139	138
1986	1.6	14	64	95	86	1.6	16	97	132	126
1987	1.8	14	70	98	87	1.6	16	98	129	126
1988	1.7	13	75	98	94	1.5	14	108	126	134
1989	1.7	13	92	96	107	1.5	14	116	126	134
1990	1.7	12	89	95	102	1.5	14	120	126	139
1991	1.7	12	76	98	93	1.4	14	133	129	154
1992	1.4	12	98	97	108	1.4	13	133	130	143
1993	1.5	11	81	98	97	1.4	13	125	131	141
1994	1.5	11	86	98	101	1.4	12	127	129	144
1995	1.5	11	86	96	98	1.5	13	128	129	139

Table 2 Doubling times t_d from the Lorenz model, the day-one rms error E_1, two "perfect-model" estimates of asymptotic error limits, E_∞ from the Lorenz model and $\sqrt{2}A_a$ from the analyzed rms anomaly, and $\sqrt{2}A_{10}$ from the day-ten rms anomaly, for the 500hPa height field computed over the extratropical Northern Hemisphere for each season since the winter of 1981.

It should be recalled that in the above discussion we have used the word "variance" in referring to the mean square deviation of each seasonal sample from a daily climatology[3]. The forecast variance thus calculated will include a contribution from systematic forecast error, the difference between the climatology of the forecasts and the climatology derived from the analyses. It does not provide simply a measure of the magnitude of typical fluctuations in forecast values.

	$\sqrt{2}S_a$	$\sqrt{2}S_{10}$	$\sqrt{2}S_a$	$\sqrt{2}S_{10}$	$\sqrt{2}S_a$	$\sqrt{2}S_{10}$	$\sqrt{2}S_a$	$\sqrt{2}S_{10}$
	Winter		Spring		Summer		Autumn	
1981	155	164	164	171	107	102	161	170
1982	169	167	164	171	107	105	161	159
1983	153	156	163	169	108	100	164	164
1984	159	152	161	165	105	95	164	160
1985	171	153	164	161	104	87	169	163
1986	155	148	174	166	106	88	161	153
1987	149	146	163	162	104	89	160	152
1988	155	155	165	172	108	97	164	170
1989	150	159	170	172	103	104	163	166
1990	154	151	162	165	104	101	158	168
1991	164	167	164	168	103	94	163	184
1992	147	171	160	180	105	116	167	176
1993	160	174	172	175	106	102	160	170
1994	156	165	162	175	104	98	163	173
1995	157	159	171	177	105	101	158	168

Table 3 $\sqrt{2}$ times the standard deviations of analyses S_a and ten-day forecasts S_{10} (m) of the 500hPa height field computed over the extratropical Northern Hemisphere for each season since 1981.

We cannot compute a reliable daily climatology of forecast error because of the changes made to the forecasting system over the years. We have, however, computed mean square deviations about sample- (seasonal-) mean fields, averaged these spatially, and taken square roots. Table 3 presents the resulting standard deviations of analyses and ten-day forecasts, multiplied by $\sqrt{2}$ so that comparison can be made with the asymptotic limits presented in Table 2. These standard deviations are substantially larger than the rms anomalies in Spring

[3] The daily climatology was calculated by averaging the fifteen analyses for each day of the year from 1981 to 1995. This period might be thought to be too short to derive reliable daily climatological values, but if these values are smoothed with a seven-day running mean, the winter values of $\sqrt{2}A_a$ increase by at most 3m.

and Autumn because the pronounced seasonal trend in analyzed and forecast values contributes to the deviation about the seasonal mean. A similar, though weaker, effect is seen in the analyses for Summer, but not in the forecasts. This is presumably because in this season the relatively weak effect of the seasonal trend on the standard deviation is broadly matched by the effect of the forecast model's systematic climate error on the rms anomaly. Table 3 generally indicates less of an overestimation of variance by the forecast model than Table 2, and in summer points to a general slight underestimation, although this has been small in recent years.

Returning to Table 2, we see that currently the limit E_∞ from the Lorenz model is generally less than $\sqrt{2}A_{10}$. E_∞ should be the same as the limit of $\sqrt{2}A_j$ for large forecast range j, under the idealized conditions discussed in sections 2 and 5. There are, however, several reasons why the Lorenz model might give a lower asymptotic limit.

One is simply that the error growth model itself is inappropriate. Simmons et al. (1995) illustrate how the alternative model proposed by Stroe and Royer (1993) (which has an extra adjustable parameter) gives a comparable fit to the differences out to ten days, but an asymptotic level which for Winter 1994 is around 15m higher than that given by the Lorenz model. Conversely, a slightly lower limit was found for Summer 1993. Similar results have been obtained for Summer 1994 and Winter 1995. The best fit to data for Summer 1995 gives an error saturation level very similar to that from the Lorenz model.

A long-term correlation of consecutive forecast anomalies would cause D_j to tend to a limit lower than $\sqrt{2}A_j$ (see equation (6)). This could occur either because of systematic model error or because of a similar response of consecutive forecasts to the model's fixed analyzed sea surface temperatures. A common response to slowly-varying anomalous land-surface conditions, such as soil moisture, could also cause some correlation between consecutive forecast anomalies computed from forecast data available to ten days ahead.

Comment can also be made on why the anomaly correlations of operational forecasts might tend to be lower in summer and autumn than in winter. Equation (9) shows how an estimate of the "perfect-model" anomaly correlation depends only on the growth rate of small errors and the initial value of rms error normalized by the asymptotic limit E_∞. The magnitude of the initial error depends in part on observational error, and may thus exhibit less of a seasonal cycle than does the asymptotic limit, which depends (for the perfect model) purely on the seasonal cycle of the atmosphere. The initial normalized rms error may thus be larger in summer than winter, and if error growth rates are no smaller in summer than winter, anomaly correlations will tend to be lower in summer. From the values in Table 2 we find the normalized one-day rms error, E_1 / E_∞, to be 0.13 averaged over the past three summers, and 0.10 averaged over the past three winters. Moreover, error growth rates have been slightly higher in summer than winter over this period. Growth rates are higher still in autumn, for which E_1 / E_∞ has almost the same value as for winter. The difference in doubling time of 0.1 days is sufficient to account for a difference of more than 5% in anomaly correlation at day seven.

The finding that the growth rates of small errors are generally lowest in winter is of some interest. The basic dynamical instability theories would lead one to expect, as did Lorenz(1982), that error growth rates would be larger in winter than summer. Moreover, the "singular vectors" that give maximum growth of energy over a time interval of order one to three days have been found to amplify more in winter than summer, when calculated using a forecast model which included only the adiabatic processes plus a simple surface drag and diffusion (Buizza and Palmer, 1995). These singular vectors are used to construct initial perturbations for the ECMWF EPS. If the explanation for the faster error growth in summer is that growth-enhancing diabatic processes play a less important role in winter, then the EPS perturbations may be less optimal in summer than winter. This is another factor which could contribute to a poorer spread of EPS forecasts in summer than winter.

8. Elimination of unpredictable scales

It has been argued above that it is desirable for a forecast model to simulate growth processes accurately, even though analysis error will amplify more rapidly in such a model than in a model which exhibits a low level of eddy activity. Ensemble prediction provides a comprehensive way of quantifying the uncertainty in forecasts that results from analysis error. It is nevertheless of some interest to consider ways of filtering the output from deterministic forecast runs to reduce the amplitude of unpredictable scales. This approach has been applied in some medium-range forecast offices for a number of years to produce a clearer picture of the evolution of large-scale circulation features. Such filtered forecasts also provide a control against which smoothed forecasts from the ensemble system, such as the ensemble mean or cluster means, can be compared.

The effect of a simple spectral filtering of the forecast on anomaly correlation was illustrated in Fig. 3. Improvement was relatively modest, although this tends be true also of the improvement of ensemble-mean over control forecasts from the ECMWF EPS, as can be seen in Buizza's paper in this volume.

Lagged-average forecasting was proposed by Hoffman and Kalnay (1983) as an alternative to ensemble prediction. In this approach the latest forecast is merged with earlier forecasts for the same verifying time in a statistically optimal way. Specifically, if forecasts are available out to a ten-day range, the filtered forecast for i days ahead $\bar{f}_i$ is given in terms of the available numerical forecasts f_j ($i \le j \le 10$) by

$$\bar{f}_i = \sum_{j=i}^{j=10} F_{ij} f_j \ .$$

The coefficients F_{ij} are determined from past forecast data to minimize some measure of forecast error, for example the rms error.

The coefficients F_{ij} that minimize rms error over the extratropical northern hemisphere have been determined using data for Winter 1994. They are presented below; values have been

smoothed subjectively and adjusted slightly to ensure row values sum to unity. Coefficients computed from data for Summer 1994 are quite similar. Diagonal elements are slightly larger for days two to six, and smaller for days seven and eight. The largest difference is .04.

$$\boldsymbol{F} = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ & .94 & .03 & .02 & .01 & 0 & 0 & 0 & 0 & 0 \\ & & .85 & .08 & .04 & .02 & .01 & 0 & 0 & 0 \\ & & & .75 & .13 & .05 & .03 & .02 & .01 & .01 \\ & & & & .65 & .15 & .08 & .05 & .04 & .03 \\ & & & & & .55 & .19 & .13 & .08 & .05 \\ & & & & & & .50 & .22 & .14 & .14 \\ & & & & & & & .50 & .25 & .25 \\ & & & & & & & & .55 & .45 \\ & & & & & & & & & 1 \end{pmatrix}$$

Forecast scores for lagged-average forecasts for Winter 1995 (using the coefficients based on 1994 data) are compared with those of the unfiltered forecasts in Fig. 20. Rms errors and differences of consecutive forecast are shown, together with corresponding anomaly correlations. Lagged-averaging improves skill scores by amounts comparable with those from other filtering methods for most of the forecast range. The increased consistency it introduces into consecutive forecasts shows up strongly in these objective measures.

A synoptic example is shown in Fig. 21. This case, that of six-day forecasts from 22 November 1995, was chosen because the operational T213L31 forecast gave both accurate and inaccurate synoptic features in the European-Atlantic domain. The verifying analysis is shown, along with the operational T213L31 forecast, the lagged-average forecast produced by combining T213L31 forecasts, and three products of the T63L19 EPS. These are the control forecast, the mean of all forecasts contained in the most-populated set produced by a cluster analysis (Molteni et al., 1996), and the ensemble-mean forecast. The inherent smoothness of the T63L19 model and the additional smoothness of the cluster- and ensemble-means are evident. The lagged-average T213L31 forecast maintains much of the structure of the cut-off low close to Ireland and of the ridge near 30°W, features well predicted in the unfiltered forecast. It filters the inaccurate shorter-wavelength features near 60°E and west of Greenland.

Other cases may be found in which the T213L31 forecast is in error on a larger scale. Typically the lagged average forecast reduces the intensity but fails to eliminate the erroneous feature. In some of these cases the controls and means from the EPS are even smoother versions of the same wrong forecast, as illustrated in Fig. 22 for forecasts from 14 November

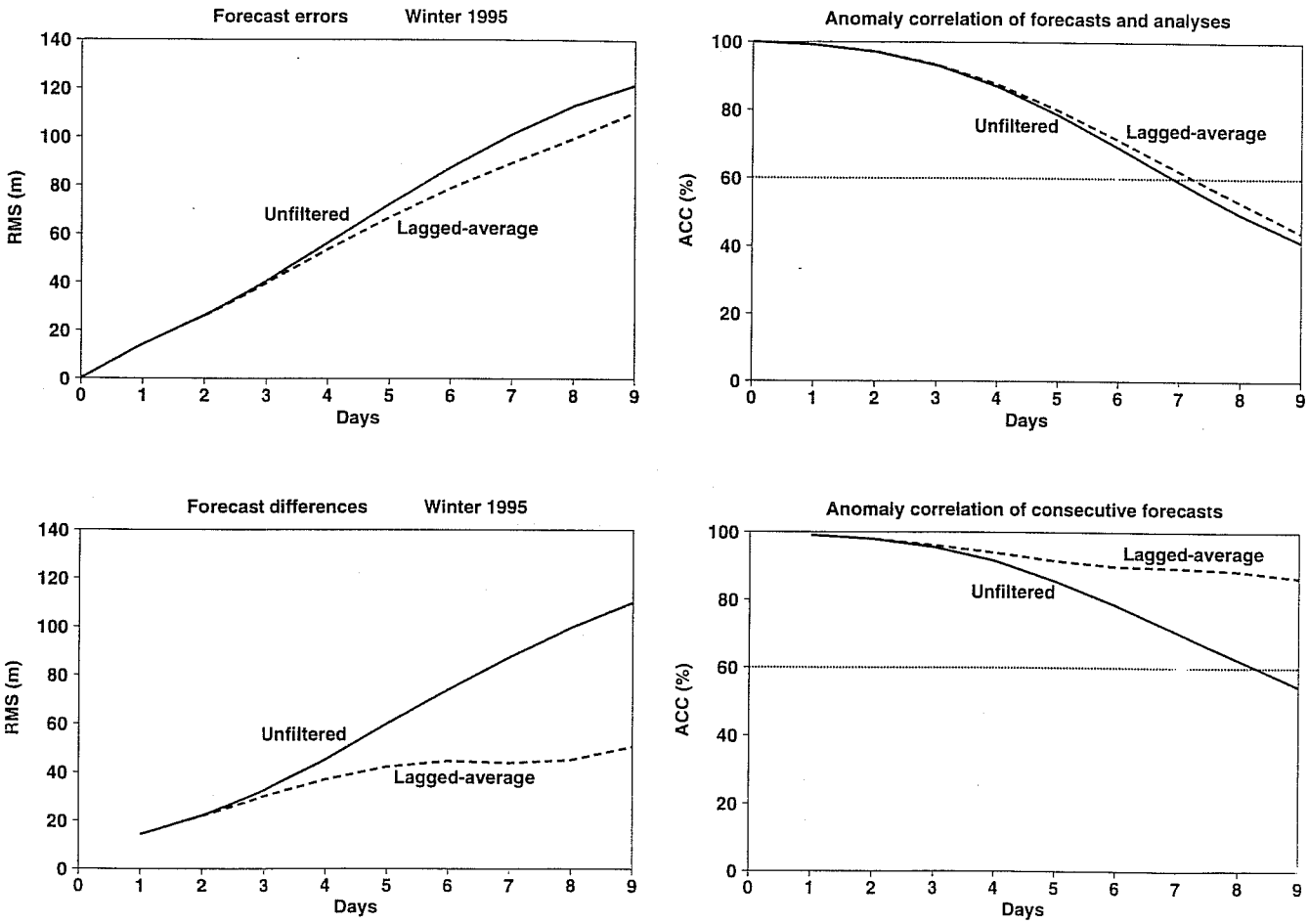

Fig. 20 Rms forecast errors (upper left) and differences between consecutive forecasts (lower left), and corresponding anomaly correlations (right), showing the impact of lagged averaging on verifications of northern hemisphere 500hPa height forecasts for Winter 1995.

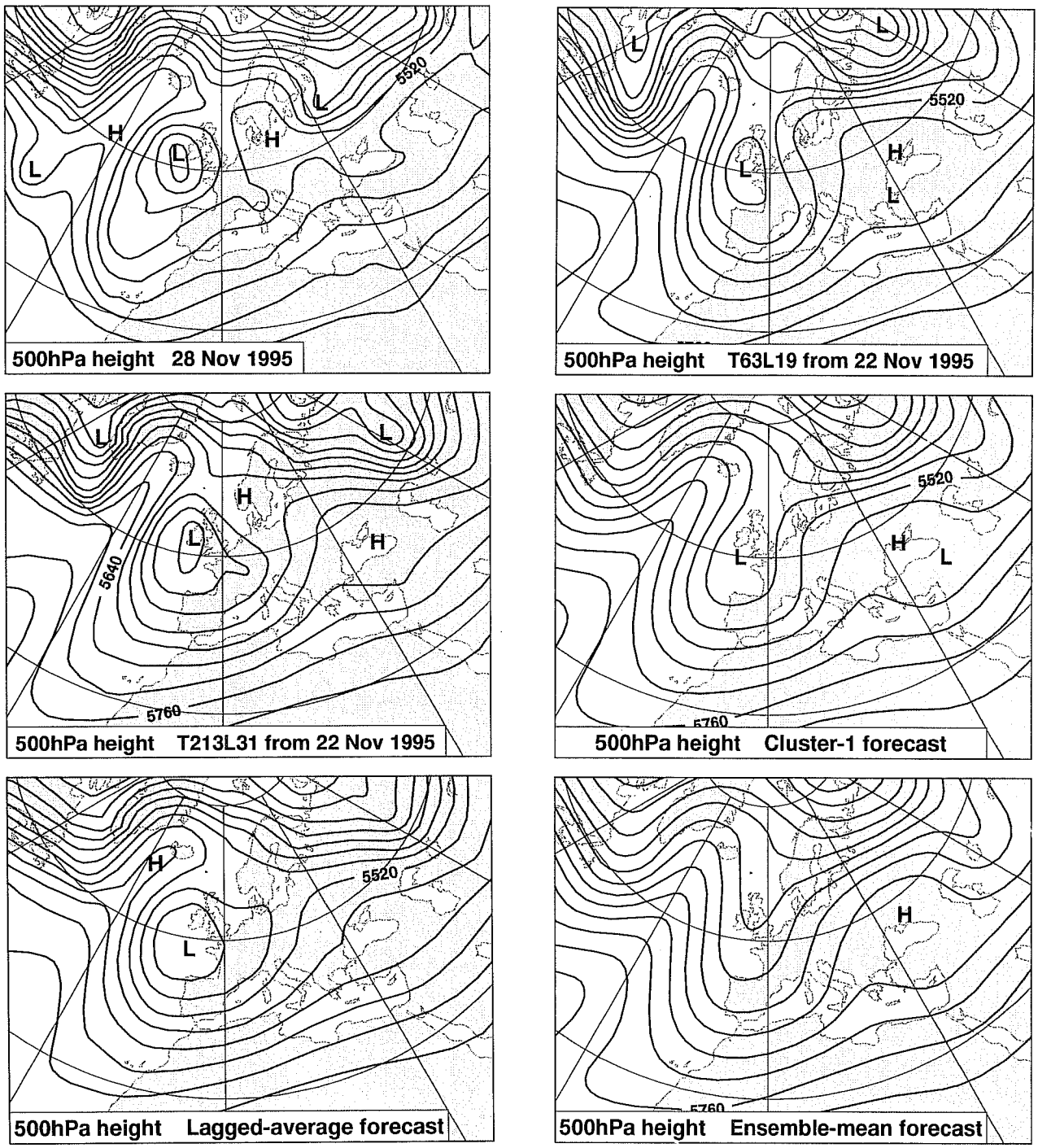

Fig. 21 500hPa height analysis for 12UTC 28 November 1995 (upper left, contour interval 60m) and six-day forecasts from 12UTC 22 November 1995:
Middle left: T213L31 operational forecast;
Lower left: Lagged-average forecast based on T213L31 forecasts from 18-22 November;
Upper right: T63L19 control forecast from EPS;
Middle right: Mean of most populated cluster of T63L19 forecasts from EPS;
Lower right: Mean of all T63L19 forecasts from EPS.

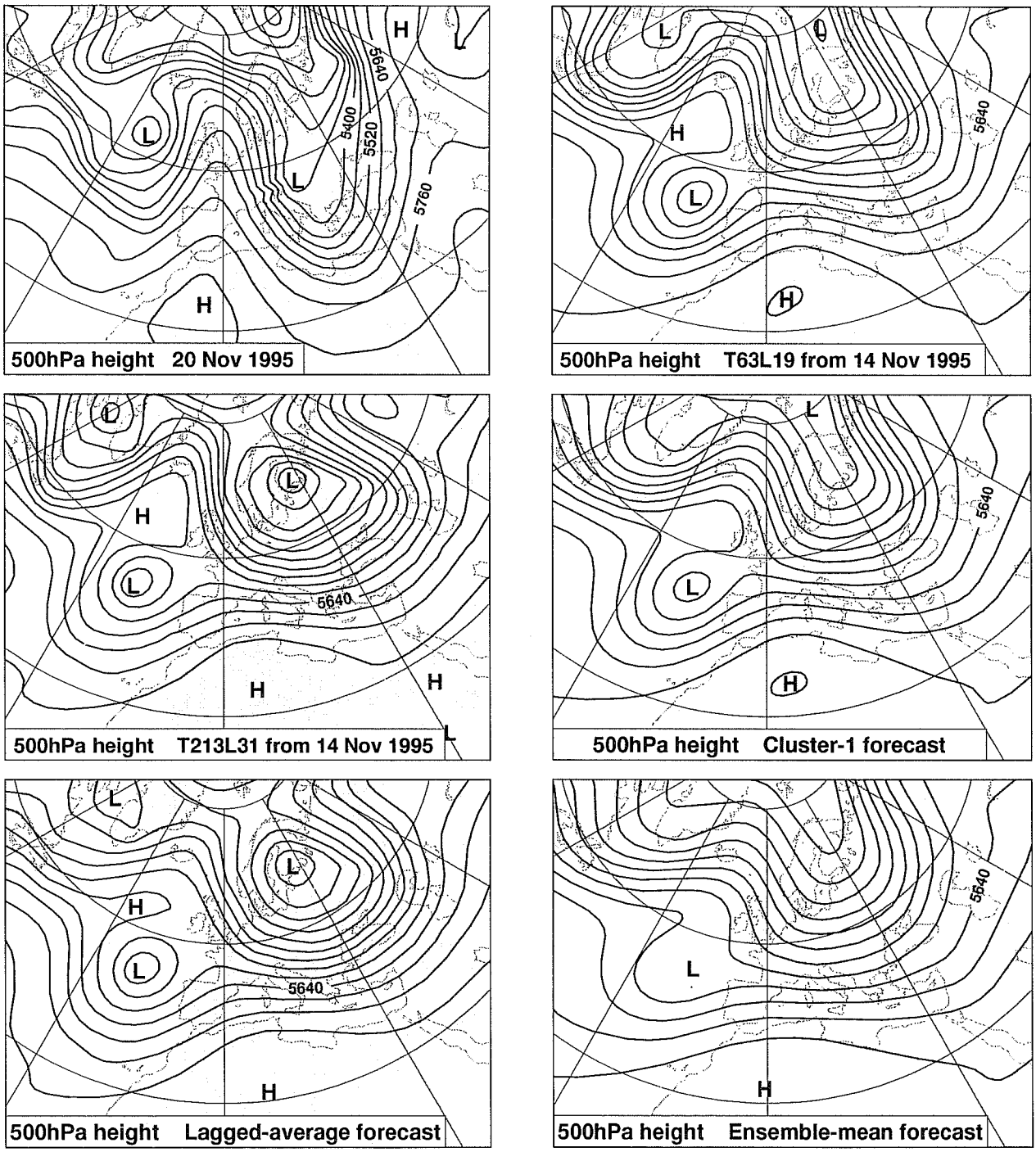

Fig. 22 500hPa height analysis for 12UTC 20 November 1995 (upper left, contour interval 60m) and six-day forecasts from 12UTC 14 November 1995:
Middle left: T213L31 operational forecast;
Lower left: Lagged-average forecast based on T213L31 forecasts from 10-14 November;
Upper right: T63L19 control forecast from EPS;
Middle right: Mean of most populated cluster of T63L19 forecasts from EPS;
Lower right: Mean of all T63L19 forecasts from EPS.

1995. In others the cluster- and ensemble-means offer smooth but synoptically preferable forecasts (Fig. 23; forecasts from 18 November 1995). Lagged-averaging high-resolution forecasts may provide a helpful supplementary product to be used together with lower resolution ensemble forecasts, though its usefulness should diminish as the resolution of the ensemble system is improved. However, formation of EPS products may benefit from incorporating time-lagged information from ensemble forecasts produced earlier, as indeed is done in the NCEP system described elsewhere in this volume.

9. Potential for improvement

We have seen in section 5 how the growth of differences between consecutive forecasts provides an indication of the potential for more accurate forecasts due to model improvement, given the current level of one-day forecast error. Lorenz' model of error growth, presented in section 7, can be used to indicate the further improvement that might result from reduction of the one-day error.

Results based on forecast performance for Winter and Summer 1995 over the extratropical northern hemisphere are presented in Fig. 24. The solid curves denote rms forecast errors, E_j , and the dashed curves the rms differences between consecutive forecasts valid at the same time, D_j . The dotted curves that lie close to the dashed curves show the good fit to rms difference curves that is provided by the Lorenz model of error growth. Finally, the dash-dotted curves show the results of the Lorenz model for a lower value of the one-day forecast error.

The reductions in one-day forecast error assumed in Fig. 24 are not arbitrary. Rather, for each season the one-day error has been reduced by the factor by which the non-systematic component of the one-day forecast error was actually reduced between 1981 and 1995. If it is indeed possible to achieve a further such reduction in the future, then the Lorenz model indicates that the reduction in rms error would be largest in the range from day four to day seven of the forecast. The range of predictability, as measured by the forecast day at which a certain level of error is reached, would increase in the medium range by two days or a little more from the combination of improvements in the model and in the initial conditions. The estimated contribution from reduction in the one-day forecast error is in broad agreement with an estimate of the impact of reduced analysis errors made by Buizza(1995) based on the sensitivity of EPS spread to the amplitude of the initial perturbations. The latter estimate does not rely on a specific model of error growth. Both estimates assume that the forecast model provides a reasonable indication of small-error doubling times, and that the reduction in initial error is not concentrated in scales or regions where error growth is substantially lower than average. Some further discussion is given by Simmons et al.(1995).

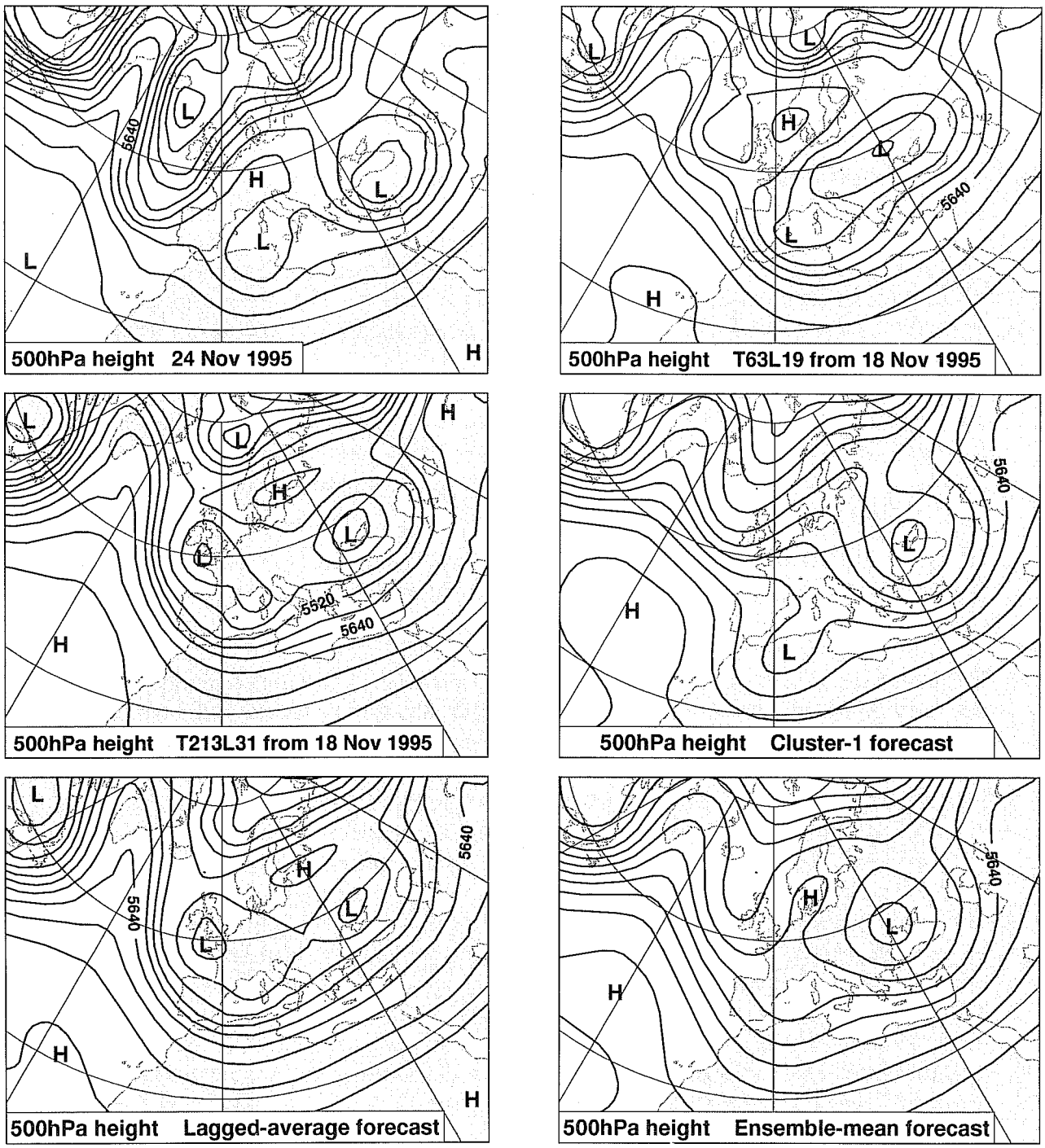

Fig. 23 500hPa height analysis for 12UTC 24 November 1995 (upper left, contour interval 60m) and six-day forecasts from 12UTC 18 November 1995:
Middle left: T213L31 operational forecast;
Lower left: Lagged-average forecast based on T213L31 forecasts from 14-18 November;
Upper right: T63L19 control forecast from EPS;
Middle right: Mean of most populated cluster of T63L19 forecasts from EPS;
Lower right: Mean of all T63L19 forecasts from EPS.

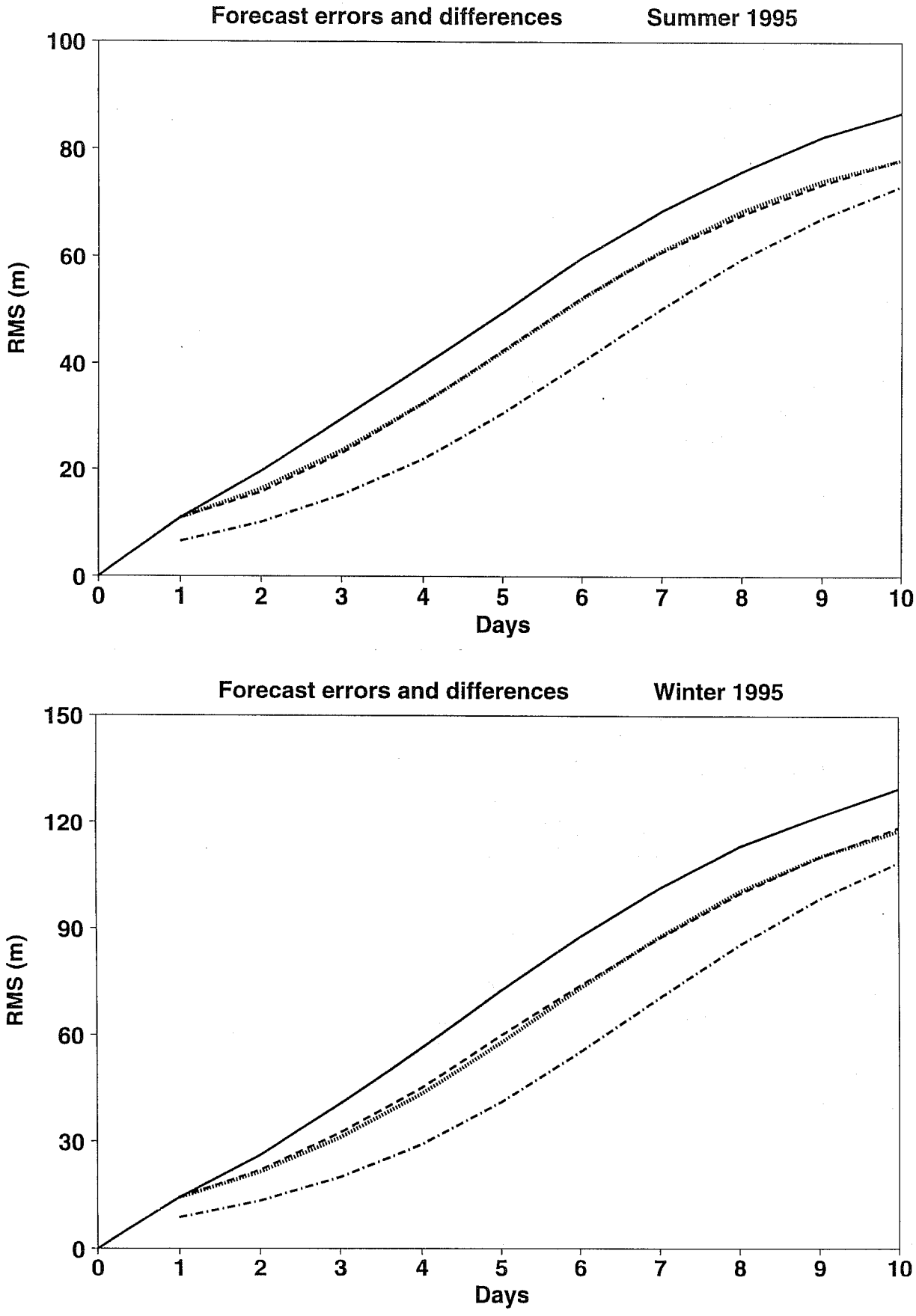

Fig. 24 Current (solid lines) and potential future rms forecast errors based on 500hPa height forecasts for the extratropical northern hemisphere for Winter 1995 (upper) and Summer 1995 (lower). Rms differences between consecutive forecasts are denoted by the dashed lines, the fits of the Lorenz error-growth model to the rms differences by the dotted lines, and the results of the Lorenz model for smaller day-one errors by the dash-dotted lines.

10. Conclusions

Objective skill scores provide a basic means of assessing trial changes to a forecasting system, of comparing the performance of operational forecast centres, and of quantifying the progress made over the years at a particular centre. Their use is, however, not without pitfalls. It has been illustrated how in the medium range they can tend to favour smooth forecasts and underactive models, and signals from particular forecasting-system changes may be masked in operational verification statistics by a dependence of skill (or at least skill scores) on the prevailing atmospheric circulation type. Subjective interpretation may be needed to draw reliable conclusions from objective verification.

A healthy overall trend in the skill scores of the operational ECMWF forecasting system has nevertheless been illustrated. This is true also for other forecasting centres. The comparisons of the 1995 levels of skill of the ECMWF and UK Meteorological Office forecasts are perhaps more noteworthy for the similarity in forecast quality than for the somewhat better performance of the ECMWF system (which anyway should be qualified in the light of the different operational constraints and commitments of the two organizations). The average UKMO forecast scores for 1995 were in fact comparable or better than the ECMWF scores of any earlier year.

The improvement in skill of ECMWF forecasts has occurred despite faster intrinsic error growth rates, as deduced from the rate of spread of consecutive forecasts. These faster growth rates appear to be an unavoidable consequence of the development of a more active and more realistic forecast model. This development, plus improvements in data assimilation and perhaps observational coverage and quality, have resulted in a substantial decline in short-range forecast error. Improvement in the medium range has been limited by the increase in error growth rate. Broadly similar conclusions have been drawn by Savijärvi (1995) for the NMC forecasting system. Diagnostics indicate that the current ECMWF model is still underactive in summer, suggesting that in this season there may still be some increase in error growth rate to come as the model is improved. In other seasons forecast accuracy may suffer from the model's overprediction of variance. These model deficiencies introduce some uncertainty into the estimates of possible future reductions in forecast error, which point to a potential for a quite substantial further improvement in medium-range forecasts.

Lack of spread of the T63L19 forecasts of the ECMWF EPS can be linked with the weaker intrinsic error growth rates of the T63L19 forecasts, a deficiency which is particularly marked in summer. Moreover, the singular vectors used to construct initial perturbations for the EPS grow more rapidly in winter than summer. This is opposite to what is found for intrinsic error growth in the T213L31 model, and may be a consequence of lack of diabatic processes (and perhaps resolution) in the model used to calculate the singular vectors. Future use of a higher resolution EPS model and incorporation of more comprehensive physics into the singular vector calculation may be of particular benefit to EPS performance in summer.

The day-to-day variations in forecast quality and in particular the occasional very poor forecasts remain a major concern. This has provided a prime motivation for development of ensemble prediction and for predictability studies in general. The average error growth

discussed above has been determined from the rms difference between consecutive forecasts, which is computed from the sum of squared differences which vary substantially in space and time. The extent of any reduction in the projection of analysis error onto rapidly-growing (singular-vector) structures will be critical in determining the extent to which the potential improvements derived from the Lorenz model of mean error growth can actually be realized. The sharp fall in the number of poor forecasts over Europe recorded in 1995 is encouraging, but time will be needed to see the extent to which this reflects genuine improvements of the forecasting system rather than a preponderance of more-predictable circulation types.

Acknowledgements

The UK Meteorological Office is thanked for making its analyses and forecasts available to ECMWF for diagnostic purposes. Thomas Petroliagis helped with the preparation of datasets. Tony Hollingsworth, Anders Persson and Roberto Buizza provided helpful comments on the text.

References

Boer, G.J. 1994: Predictability regimes in atmospheric flow. *Mon. Wea. Rev.*, **122**, 2285-2295.

Buizza, R. 1995: Optimal perturbation time evolution and sensitivity of ensemble prediction to perturbation amplitude. *Quart. J. Roy. Meteor. Soc.*, **121**, 1705-1738.

Buizza, R. 1996: Potential forecast skill of ensemble prediction, and spread and skill distributions of the ECMWF Ensemble Prediction System. *In these Proceedings.*

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

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

Hortal, M. 1994: Recent studies of semi-Lagrangian advection at ECMWF. *ECMWF Tech. Memo.*, **204**, 37pp.

Lott, F., and M.J. Miller 1996: A new sub-grid scale orographic drag parametrization: Its formulation and testing. *Submitted to Quart. J. Roy. Meteor. Soc.*

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

McNally, A.P., and M. Vesperini 1995: Variational analysis of humidity information from TOVS radiances. *Res. Rep. EUMETSAT/ECMWF Fellowship Programme*, **1**, 29pp.

Miller, M.J., M. Hortal and C. Jakob 1995: A major operational forecast model change. *ECMWF Newsletter*, **70**, 2-8.

Molteni, F., R. Buizza, T.N. Palmer and T. Petroliagis 1996: The ECMWF Ensemble Prediction System: Methodology and validation. *Quart. J. Roy. Meteor. Soc.*, **122**, 73-119.

Nieminen, R. 1983: Operational verification of ECMWF forecast fields and results for 1980-1981. *ECMWF Tech. Rep.*, **36**, 48pp.

Norris, B. 1994: METVIEW/BATCH User's Guide, *ECMWF Met. Bull.*, **M1.10/1(3)**.

Rabier, F., E. Klinker, P. Courtier and A. Hollingsworth 1996: Sensitivity of forecast errors to initial conditions. *Quart. J. Roy. Meteor. Soc.*, **122**, 121-150.

Ritchie, H., C. Temperton, A.J. Simmons, M. Hortal, T. Davies, D. Dent and M. Hamrud 1995: Implementation of the semi-Lagrangian method in a high resolution version of the ECMWF forecast model. *Mon. Wea. Rev.*, **123**, 489-514.

Savijärvi, H. 1995: Error growth in a large numerical forecast system. *Mon. Wea. Rev.*, **123**, 212-221.

Simmons, A.J. 1986: Numerical prediction: Some results from operational forecasting at ECMWF. *Advances in Geophysics*, **29**, 305-338.

Simmons, A.J., R. Mureau and T. Petroliagis 1995: Error growth and estimates of predictability from the ECMWF forecasting system. *Quart. J. Roy. Meteor. Soc.*, **121**, 1739-1771.

Stroe, R. and J.F. Royer 1993: Comparison of different error growth formulas and predictability estimation in numerical extended-range forecasts. *Ann. Geophysicae*, **11**, 296-316.

Tiedtke, M. 1993: Representation of clouds in large-scale models. *Mon. Wea. Rev.*, **121**, 3040-3061.