30- AND 60-DAY FORECAST EXPERIMENTS WITH THE ECMWF SPECTRAL MODELS

F. Molteni, U. Cubasch and S. Tibaldi
European Centre for Medium Range Weather Forecasts
Reading, U.K.

ABSTRACT

During the last five years, the feasibility of dynamical extended-range forecasts by means of atmospheric general circulation models (GCMs) has received support from a number of experimental studies. However, despite the encouraging results obtained in some case studies, two main factors still prevent the skill of dynamical forecasts from reaching a level acceptable for practical applications: the intrinsic instability of the atmospheric flow and the presence of systematic errors in the fields predicted by almost all the large GCMs. It has however been shown that both time-averaging and ensemble averaging can effectively reduce the impact of initial uncertainties and model-generated random errors on the quality of the forecast. On the other hand, if really systematic features appear in the error fields produced by a GCM, it should be possible to remove them from the output of a prediction experiment by subtracting the mean error computed from a (possibly) large sample of independent forecasts. The main purpose of the experiments described here has been the evaluation of the impact of i) ensemble forecasting, ii) correction of the systematic error and iii) increased resolution of the model on the forecast skill. Two versions of the ECMWF operational spectral model were used, adopting a triangular truncation at total wavenumber 21 (T21) and 42 (T42) for the expansion of the horizontal fields. The Lagged-Average Forecasting method was used to generate the ensemble of predictions described in the first part of the work. Nine forecasts, starting from initial conditions separated by 6-hour intervals, were averaged at the corresponding verification times up to 30 days after the last initial analysis; the single, purely deterministic forecasts starting from the last analysis were used, together with persistence, for a comparative assessment of the results. Statistically corrected Lagged-Average and deterministic forecasts were also obtained by subtracting from the predicted fields an estimate of the systematic error deduced for each experiment from a

sample of 10 forecasts starting from analyses separated by 10-day intervals in the same and the following month of the previous two years. The T42 corrected Lagged-Average forecast appears to have the highest skill among all the combinations of resolution/error filtering methods explored. For the 850 mb temperature the forecasts are, on average, more skilful than either climate or persistence.

Finally, a series of 38, purely deterministic, thirty and sixty day forecasts run from randomly selected winter initial conditions with the ECMWF spectral model using the same two resolutions (T21, T42) have been examined. The purpose of this further study was to assess objectively the skill of purely deterministic extended-range forecasts in a quasi-operational environment. The large ensemble of forecasts allows, furthermore, an estimate of the spread in the forecast quality. It appears again that the higher resolution model (T42) is more skilful than the lower resolution version (T21) and has a higher reliability. The T42 model is better than a persistence forecast of 10 day means up to the first half of the 30 day period. The T42 30-day mean forecast is almost consistently better than persistence. The skill of the T42 model forecast appears to be uncorrelated with the skill of the persistence forecast which undergoes large case-to-case and interannual variations.

1. INTRODUCTION

During the last twenty years, extended and long-range weather forecasting (defined as forecasts from 10-15 days to one month and from 1 month to 3 months respectively) have progressed from an empirical stage, dominated by purely statistical approaches with only qualitative physical bases, to the level of an important branch of dynamical meteorology. At present, observational studies and theoretical and numerical models allow the evaluation of the structure and physical causes of the low-frequency components of the atmospheric circulation and, at least partially, to predict their temporal evolution (see Nicholls et al., 1984).

As pointed out in an earlier work by Nicholls (1980), statistical forecasting methods have generally failed to produce reliable operational monthly or seasonal forecasts. However, when used as diagnostic tools, statistical methods have provided plenty of observational evidence on the existence of some modes of low-frequency atmospheric variability that should be dynamically predictable: see, among others, Namias (1969), Ratcliffe and Murray (1970), Horel and Wallace (1981) on the influence of sea-surface temperature anomalies, or the work by Wallace and Gutzler (1981) on the existence of preferred patterns of variability connected to 'centres of action'. It has also been demonstrated that potential long-range predictability is characterised by a strong seasonal and geographical dependence (e.g. Madden, 1976 and Walsh and Richman, 1981).

Meanwhile, the continuous progress in the numerical simulation of the atmospheric general circulation and in medium-range weather forecasting has shown that, for some situations, it is possible to obtain skillful dynamical monthly forecasts of the larger scale features (Miyakoda and Chao, 1982; Miyakoda et al., 1983).

However, the skill of a deterministic prediction carried out with a numerical model is bounded by two factors. The first factor, the intrinsic instability of the atmosphere with respect to small perturbations, has been discussed in the fundamental works of Lorenz (1963, 1969a, 1969b, 1982). Even with a perfect model, analysis errors as small as the current 24-hour forecast error of a state-of-the-art General Circulation Model (GCM) would double in about two days. However, the smaller the characteristic spatial scale of the

initial error, the smaller is its doubling time; so, even if analysis errors are increasingly confined to smaller scales, the time at which the largest scales become unpredictable (that is when a prediction of their instantaneous state becomes as good as a random forecast) cannot be increased indefinitely. This time limit was determined by Lorenz (1969a) to be about 15 days and this value is now commonly accepted as a good estimate of the limit of deterministic predictability.

The second factor which sets a limit to the forecasting of instantaneous weather patterns is the imperfections and inadequacies of even the best GCMs currently available. This adds further sources of error that, in practice, limit the usefulness of numerical weather predictions (NWPs) to about one week. An important part of this error is not random and causes a gradual drift with time of the mean state of the model atmosphere from the actual climate towards the model's own climate (as deduced from very long integrations). Also von Storch et al. (1985) have shown that many features of the systematic errors are common to different GCMs: examples are the excessive westerlies (Arpe et al., 1985, Arpe and Klinker, 1986 and Palmer et al., 1986) and the tendency in winter to produce a lower ratio of eddy to zonal kinetic energy than that observed, and in particular to underestimate the amplitude of the planetary-scale stationary eddies in the northern extratropics (Hollingsworth et al., 1980 and Wallace et al., 1983).

Despite these limitations, the hope for useful dynamical predictions up to one month or more is supported by the following considerations:

(a) Even though the instantaneous state of the atmosphere cannot be predicted after about 15 days, some statistical properties of the atmospheric variables (for example, the mean fields over periods from 10 days to one month) may 'remember' the influence of the initial conditions beyond this limit. This was illustrated by Shukla (1981) who showed that the differences between predicted monthly means computed from initial fields observed on the same date of different years are significantly greater than those obtained by superimposing random perturbations to the same initial fields (all forecasts used climatological boundary conditions). Shukla calls 'dynamical predictability' that which derives from the influence of the initial conditions.

(b) As suggested by observational and theoretical studies and confirmed by a number of numerical experiments (e.g. Rasmusson, 1983; Walsh, 1983; Shukla and Wallace, 1983; Shukla, 1984), the forcing generated by the surface boundary conditions (namely the sea surface temperature (SST), soil moisture, snow cover and sea ice) can significantly affect the general circulation and play an important role in determining the large scale anomalies of the atmospheric fields. Since most of the anomalies in the surface conditions have a typical time of variation of weeks or months, a correct initial analysis of these components of the climatic system and a good model representation of their influence upon atmospheric motions can increase the predictability of the atmospheric anomalies up to forecast times of comparable duration.

(c) Even if a deterministic prediction is subject to the continuous growth of the initial error, which ultimately masks the signal in the forecast, the mostly random nature of this error due to imperfections in the initial conditions suggests that the valuable signal may be more clearly detected by averaging the results of a number of forecasts that start from slightly different initial states. As shown by Epstein (1969), this procedure of 'ensemble forecasting' can be seen as the practical realisation of the theoretical approach of Gleeson (1968, 1970), who suggested that the observed initial state and its uncertainty can be characterised by a probability density function in phase space, and that the temporal evolution of this function gives a probabilistic estimate of the future state of the atmosphere. The theoretical and practical developments of these ideas are known as stochastic-dynamic prediction methods and the Lagged-Average Forecasting (LAF) technique proposed by Miyakoda and Talagrand (1971) and Hoffmann and Kalnay (1983) is based upon such ideas.

(d) A part of the error generated by numerical models is systematic. This suggests that, if the systematic error (SE) is computed from a large sample of integrations in the same period of the year, the subtraction of the SE can reduce the root mean square error of a comparable ENSEMBLE of independent forecasts (see, e.g., Arpe, 1983). The effectiveness of the application of this subtraction to a SINGLE forecast obviously depends on the statistical significance of the SE, that is on the ratio between mean and standard deviation of the error. Besides, the correction will probably not be suitable for a prediction performed with boundary conditions significantly different from those used in the computation of the SE.

On the basis of these arguments, four sets of experimental monthly forecasts were carried out with the following characteristics:

- the predicted fields were 10-day mean and 30-day mean anomalies of geopotential height and temperature at various pressure levels in the northern hemisphere;
- observed SSTs (analysed operationally by NMC) were used as boundary conditions;
- a stochastic-dynamic method, namely the LAF technique, was chosen in order to reduce the random part of the error and its performance was assessed against purely dynamical integrations;
- an estimate of the model SE, deduced from a sample of ten independent monthly integrations (appropriate for the season of year and started from random initial conditions during the previous two years), was subtracted from the LAF output to produce (statistically) corrected anomalies;
- the integrations were performed at two different horizontal resolutions.

Due to the large number of long integrations required for this study, two low-resolution versions of the ECMWF operational spectral model (Louis, 1984) were used. The expansion of the horizontal atmospheric fields in spherical harmonics was limited by a triangular truncation at wavenumber 21 (T21) and 42 (T42). Consequently, the horizontal resolution is, in both cases, isotropic and the length of the smallest wave resolved with these truncations is about 2000 and 1000 km respectively.

This report presents the results of the experiments, and is organised as follows: Section 2 is devoted to the problem of the SE of the ECMWF T21 and T42 models. Section 3 describes the results of four LAF 30-day experimental forecasts, discusses a further experiment devised to test the impact of SSTs on the forecast skill and assesses the correlation between the spread of the forecast ensembles and the forecast skill. Section 4 presents some overall results on purely deterministic extended range forecasts by analysing separately the comparatively larger database of 30-day (and some 60-day) integrations used previously to estimate the model systematic error. The conclusions are summarized in Section 5.

2. THE SYSTEMATIC ERRORS OF THE T21 AND T42 SPECTRAL MODEL

Before discussing the results of the forecasting experiments, it is worthwhile describing the main characteristics of the SE of the model employed. Emphasis will be placed on the mean 500 mb height and 850 mb temperature (referred to as Z500 and T850) over days 21-30. These parameters are chosen because they are most often used to characterize model behaviour and are widely used as forecasting tools; the period chosen is taken to be the best available to infer the model asymptotic behaviour for longer integrations. The January-February period (used later to correct the LAF forecast started from 17 January 1984) was chosen and the ensemble of ten 30 day integrations had initial dates 1, 11 and 21 January, 1 and 11 February 1982 and 1983.

2.1 500 mb height

Fig. 2.1 shows the mean observed Z500 (labelled OBS MEAN), the corresponding mean forecast (labelled FOR MEAN) for the two model resolutions and their SE (labelled SYS ERROR). Apart from a common tendency to underestimate the amplitude of planetary scale waves (see Hollingsworth et al., 1980, Wallace et al., 1983 and Tibaldi, 1986), the structure of the SE field appears to be substantially different for the two resolutions. The T21 clearly tends to reduce the latitudinal height gradient and, hence, the strength of the mean (geostrophic) zonal wind, with the maximum reduction taking place over the Atlantic and Western Europe area; here the jet diffluence is exaggerated and the anticyclonic curvature typical of this area during winter months is lost. On the contrary, the T42 model tends to increase the north-south height gradient and therefore the strength of the jet between 40° and 60°N over the Northern Pacific and over Europe. Also the jet diffluence over Western Europe is almost completely lost, together with the anticyclonic curvature. Another feature of the T42 forecasts is the damping of the Rocky Mountains and Alaskan ridge; hence the wavenumber 3 signature on the planetary scale eddy field is weakened substantially while wavenumber 2 is wrongly enhanced. South of about 40°N for the T21 and 30°N for the T42, the SE is everywhere negative for both models.

The ability of the two models to reproduce the time variability of Z500 is shown in Fig. 2.2, where the standard deviations of the observed (OBS S.DEV) and forecast fields (FOR S.DEV), and the ratio of the forecast and observed values (FOR S.DEV/OBS S.DEV) are depicted. It is evident that both models

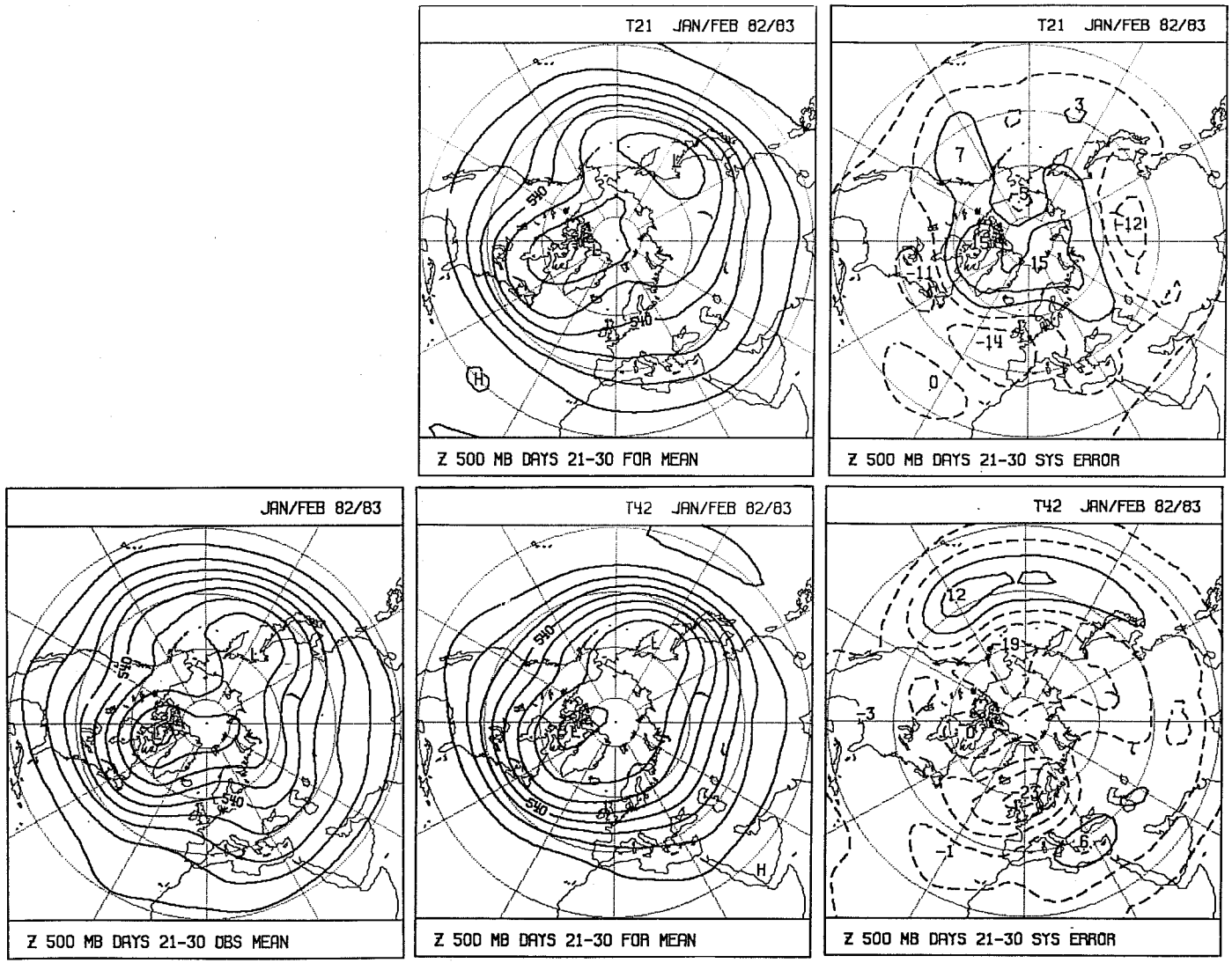


Fig. 2.1 Mean analyses (left), forecasts (centre) and systematic error (right) of northern hemispheric 500 mb height in the 21-30 day range for a sample of ten T21 (top) and T42 (bottom) monthly forecasts started in January and February 1982 and 1983.

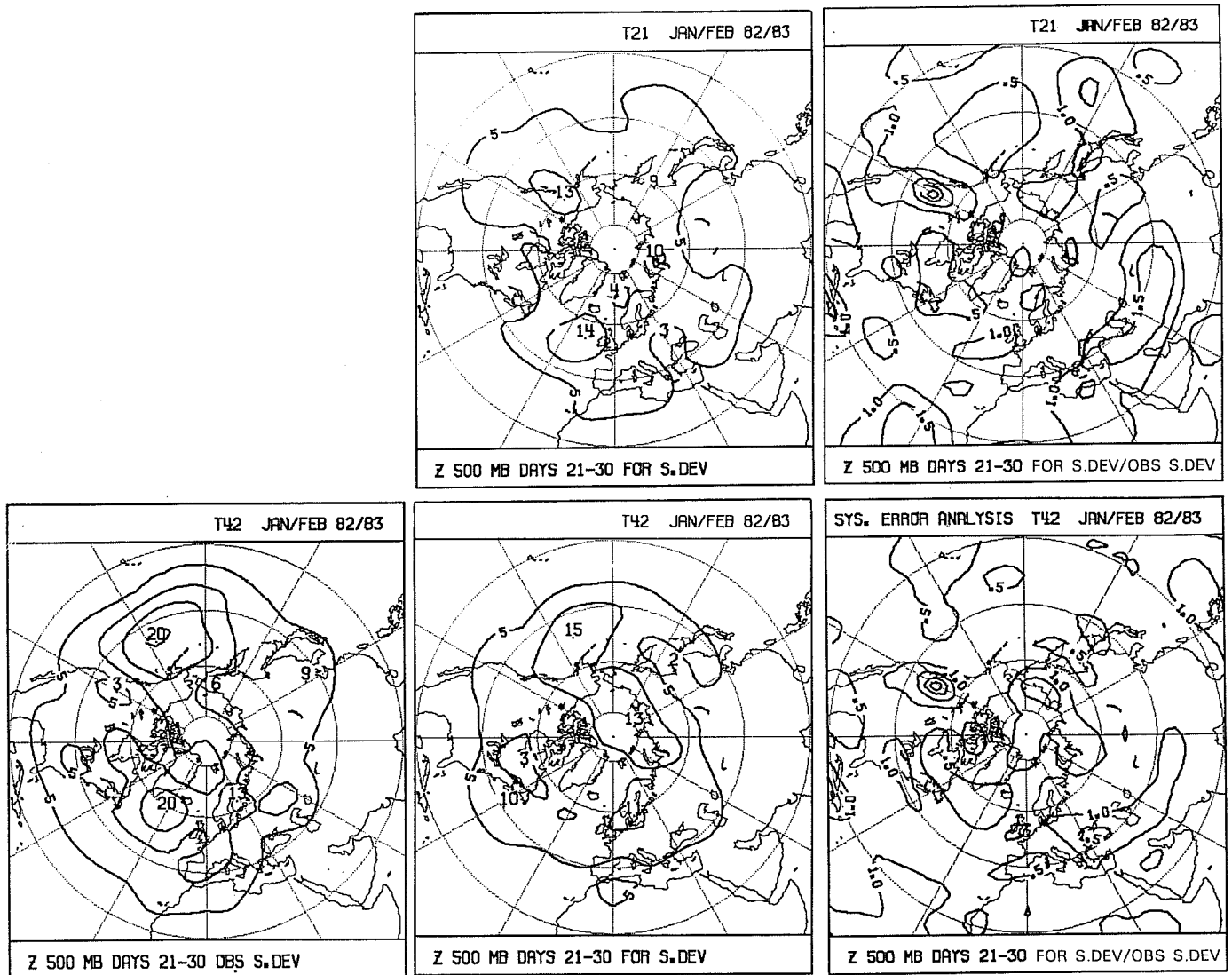


Fig. 2.2 Standard deviation of the analyses (left), of forecasts (centre) and their ratio (forecast/observed, right) of the northern hemispheric 500 mb height in the 21-30 day range for a sample of ten T21 (top) and T42 (bottom) monthly forecasts started in January and February 1982 and 1983.

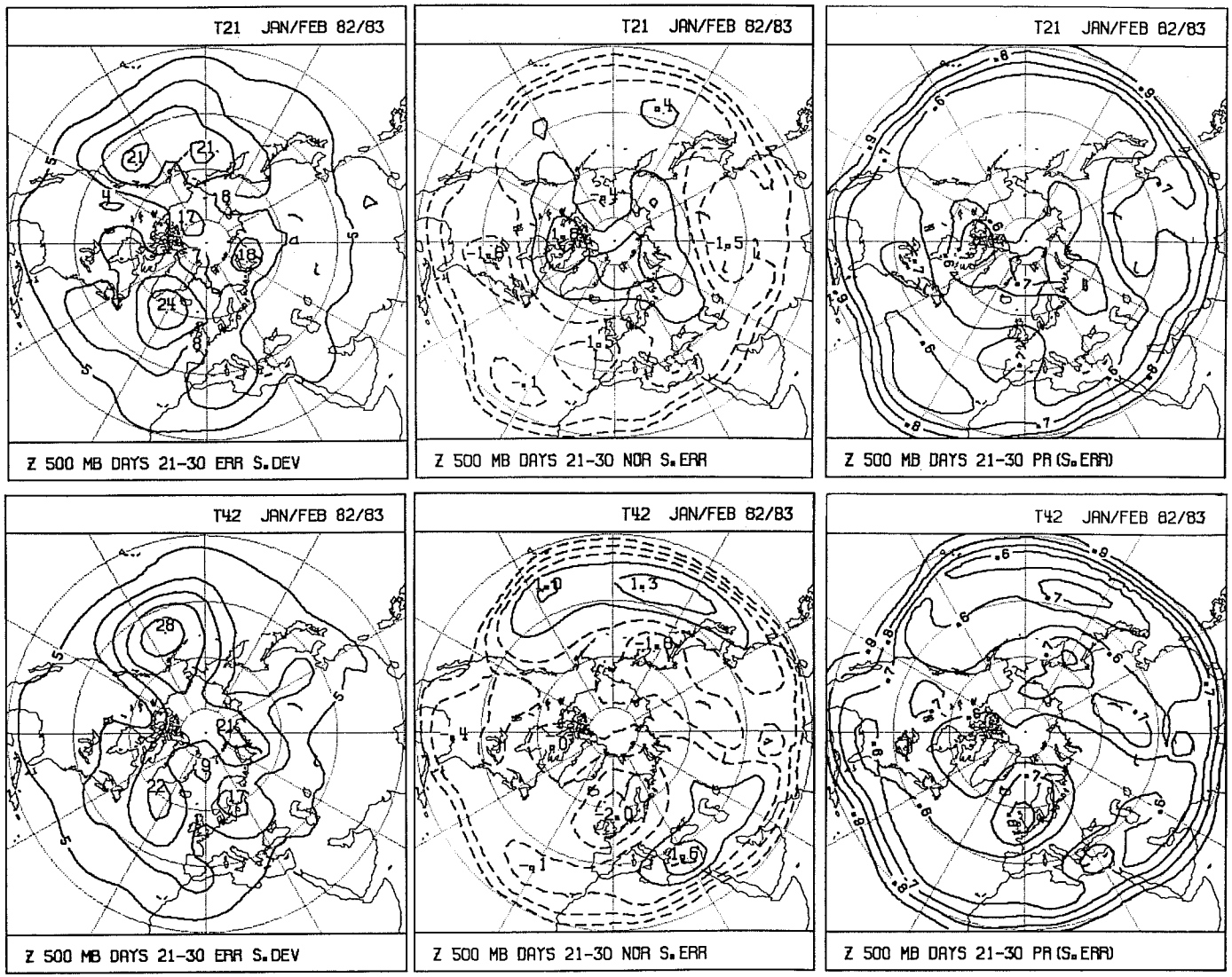


Fig. 2.3 Standard deviation of forecast error (left), ratio between mean and standard deviation (normalised SE, centre) and probability of SE (right). T21, top and T42, bottom. For more explanations see text, Section 2.

tend to underestimate such variability, particularly the T21 model over the Pacific area and the T42 model over the Atlantic. The models' standard deviation of Z500 are smoother than the observed and it appears that the influence of orography is not well reproduced (note in particular the Rocky Mountains region).

Fig. 2.3 allows the evaluation of the ratio of the random and mean error for the two model resolutions, and therefore the significance of the latter. First consider the standard deviation error (ERR S.DEV) - it is fairly high for both models and, in many areas, greater than the natural variability; this indicates that the correlation between forecast and observed fields in the period 21-30 days is likely to be very low. The higher standard deviation of the T42 model with respect to the T21 over the Pacific area is most likely due to the higher variability of the higher resolution model, rather than to a worse performance in terms of forecasting skill (as we will see later).

The ratios between mean and standard deviations of the model error, the normalised SE (NOR S.ERR) given in the centre panels of Fig. 2.3, show that for both model resolutions the most noteworthy structures of the SE are statistically significant (for a sample of 10 elements the confidence level of 95% corresponds to a ratio higher than approximately .7); over wide areas the SE is comparable or greater than the random error. This implies that, on average, we should expect improvements by using this estimate of the SE to correct our forecasts. However, as it will be shown later, on individual cases the results can be very variable.

By assuming that the probability distribution of the error around its own mean is Gaussian, it is possible to evaluate, for every point, the probability (PR(S.ERR)) that the error is nearer to its mean value than to zero - that is the probability that a SE correction reduces the total error of a forecast rather than increasing it. This probability is shown in the right panels of Fig. 2.3 and has a minimum value of 50% where the SE is zero. It can be seen how, in the northern hemisphere extratropical latitudes, for both models the term "systematic" it not always appropriate, since only in some areas is the subtraction of the SE advantageous in at least 70% of the cases. A completely different picture is true for tropical regions, where the variability of the error is much smaller than its mean value.

2.2 850 mb temperature

The fields shown in Figs. 2.1, 2.2 and 2.3 for the Z500 are reproduced in Figs. 2.4, 2.5 and 2.6 for the T850. The general tendency for both models to cooling the troposphere, already deducible from 500 mb height fields, is evident. Negative peaks of SE are obviously connected to underlying orographic features and the positive areas (larger for T21 than for T42) are confined to arctic regions. For the higher resolution model, noteworthy is the positive centre over the central and east Mediterranean region, including parts of East Europe and Western Asia, due to an excessive zonalisation of the model flow. The T21 error seems to display a vertical structure of the SE more equivalent barotropic than the T42.

The variability of the model's T850 (Fig.2.5) is on the whole more similar to the observed field than was found for Z500 (Fig. 2.2), with the T42 model appearing better than the T21. South of 40°N, there are large areas where the model variability appears larger than that observed, while north of this latitude the opposite occurs. It should be noted that the maximum observed variability takes place over continental masses (Alaska, Canada, Eastern Siberia) and that, with the exception of Alaska, in these areas the models display a large SE and a clear lack of variability. Even allowing for some errors due to subterranean extrapolation of the 850 mb temperature in some of those areas, these facts underline the need for a more appropriate modelling of the thermal interactions between the lower troposphere and the underlying surface.

As for Z500, Fig. 2.6 shows that for T850 the important structures of the SE are also statistically significant; however in middle and high latitudes the mean error is only really 'systematic' in limited areas, so that the correction for the SE should give, on average, positive results, but with negative effects in some cases.

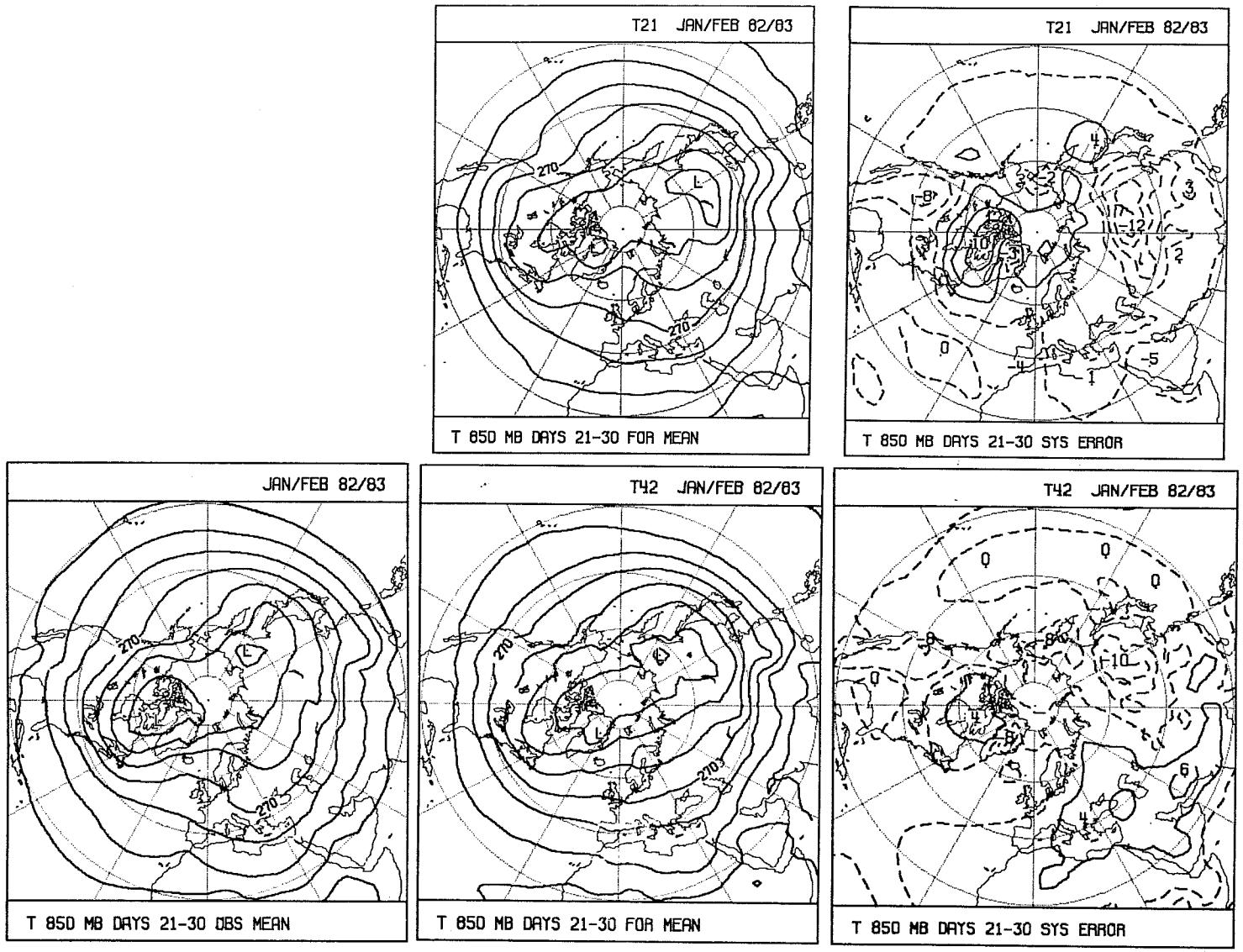


Fig. 2.4 As Fig. 2.1, but for T850

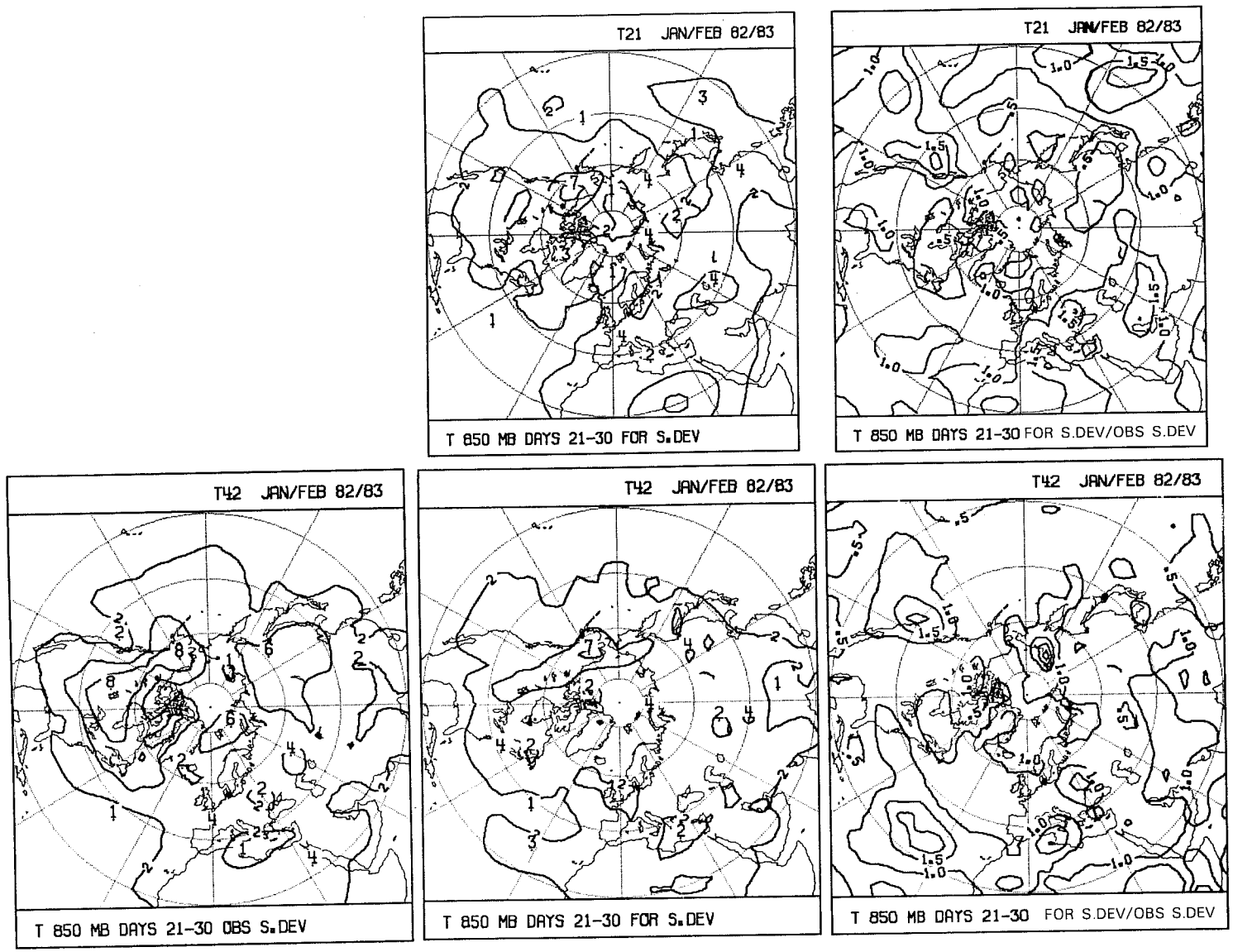


Fig. 2.5 As Fig. 2.2, but for T850

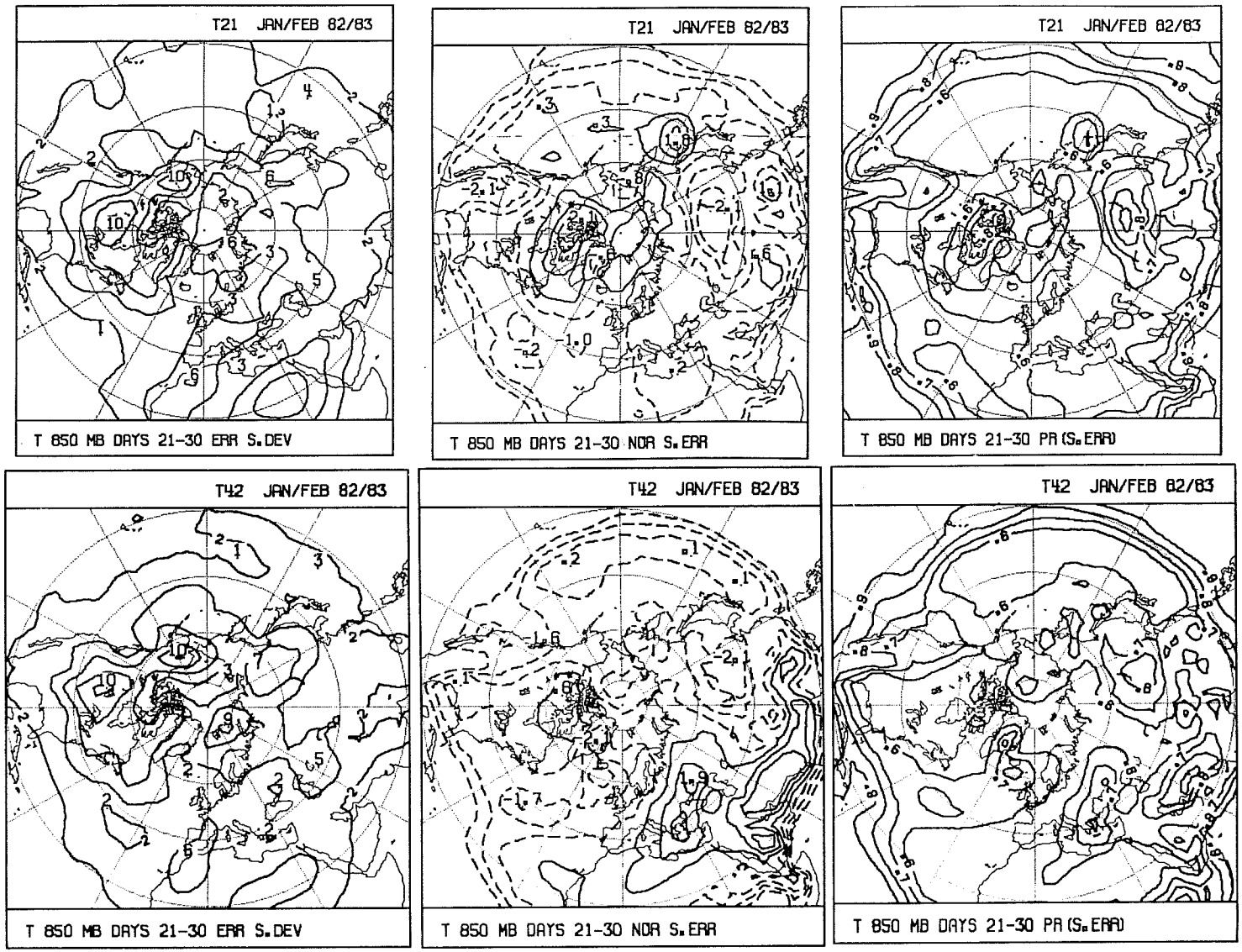


Fig. 2.6 As Fig. 2.3, but for T850

3. THE LAGGED-AVERAGE FORECASTS

3.1 Description of the experiments

The Lagged Average Forecasting (LAF) technique gives an estimate of the predicted fields by averaging the results of n integrations starting from initial times separated by a time δt. If t_o is the initial time of the most recent integration, the ensemble of forecasts start from times t_o, $t_o-\delta t$.... $t_o-(n-1)\delta t$. The construction of a LAF ensemble is illustrated in Fig. 3.1. For our experimental 30-day forecasts n=9 and δt=6h were used, and the initial times t_o were 12 GMT on 15 December 1983, 17 January 1984, 19 February 1984 and 23 March 1984. The 17 January and 19 February cases were chosen essentially because they had produced, operationally, a very good and very bad medium-range forecast respectively. The other two were chosen with the sole purpose of spanning a four-month, non-overlapping period and can therefore be considered as chosen at random.

For each of the four initial conditions and each of the two model resolutions the following operations were performed:

(a) Nine forecasts were integrated from progressively lagged initial conditions (see Fig. 3.1). The LAF was then defined as the arithmetic mean of the nine forecasts verifying at the same time. All forecasts had the same weight in such a mean (unlike, for example, in Hoffmann and Kalnay, 1983) because the main interest was on forecast times longer than ten days, for which all weights would have been practically identical.

(b) A purely dynamical deterministic forecast (DET) was obtained by spectrally truncating at T10 and T21 respectively the T21 and T42 integrations that started from the most recent initial conditions amongst the nine. The spatial truncation of the output fields was performed to remove from the DET forecasts all those scales that are filtered out by the ensemble averaging in the LAF experiments. This avoided the LAF experiments appearing objectively more skillful than the DET integrations because less variance was contained in the smaller scales.

(c) An estimate of the SE appropriate for the month of the year of each experiment was computed as the average of the forecast errors of ten independent integrations started from 1, 11 and 21 of the same month as the initial conditions and from 1 and 11 of the following month, for the two

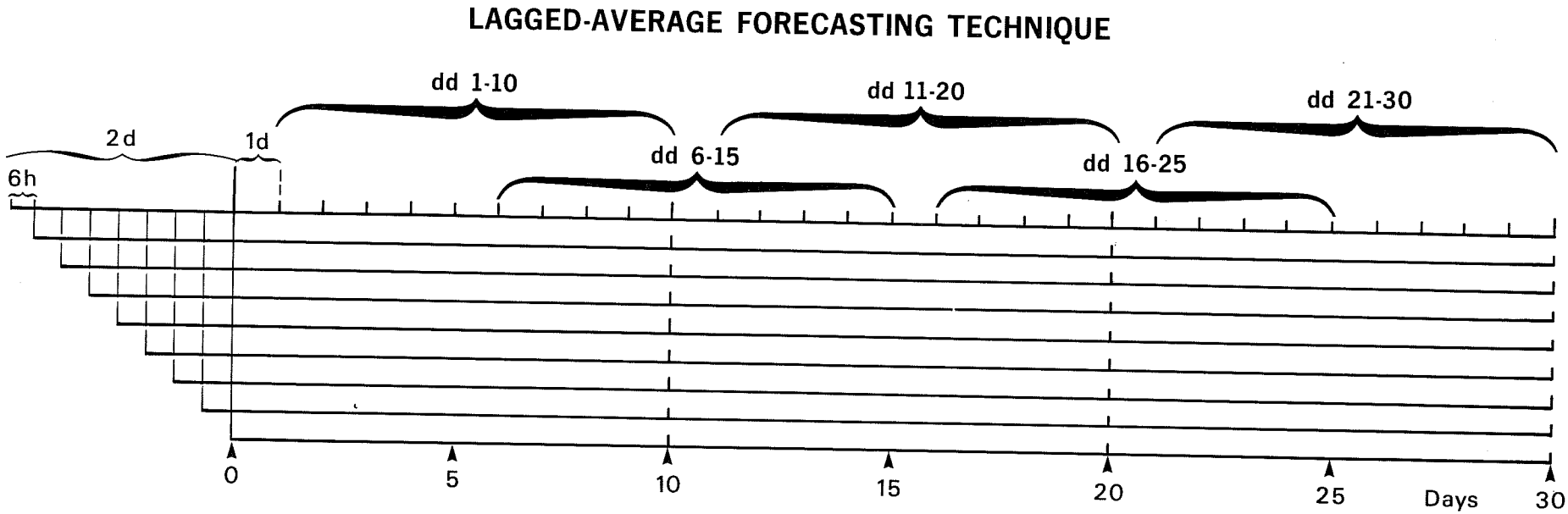


Fig. 3.1 Schematic of the construction of a LAF forecast composed of an ensemble of nine individual deterministic forecasts starting from adjacent initial analyses, progressively lagged of 6 hours. Periods of time-average are also shown.

preceding winters (1981/82 and 1982/83). This SE estimate was then used to produce statistically corrected versions of both LAF and DET forecasts, referred to as SCL and SCD, respectively.

(d) For each forecast objective skill scores (against observed fields) of 10-day running means (days 1-10, 6-15, 11-20, 16-25, 21-30) and of monthly means (days 1-30) for 1000, 500 and 300 mb height and 850, 500 and 300 mb temperature were evaluated. This was done in order to select only the low-frequency components of the atmospheric variability which are suspected of possessing a longer predictability (Smagorinsky, 1969; Gilchrist, 1977; Shukla, 1984). The objective measures used are the anomaly correlation coefficient between forecast and analysed fields (ACC) and the root mean square (RMS) error, both computed over the extratropical northern hemisphere (20°N to 90°N).

(e) Objective skill scores were also computed for two types of persistence forecasts based on the persistence of anomalies, taking therefore the seasonal cycle into account. The first, referred to as simple persistence (PER), assumes that the forecast anomaly at the verification time (10 days or a month) is the observed anomaly averaged over the corresponding period before the initial time of each forecast. The second, referred to as long-term persistence (LTP), is similar to PER, but the observed anomaly is averaged over a period that increases with increasing forecast time. For example, LTP for the anomaly for the forecasting interval 11-20 days is the mean observed anomaly during the previous 20 days. The idea on which this less conventional persistence forecast is based is that those features that appear in a time mean over a period Δt are more likely to persist for the same period of time Δt.

In computing anomaly fields, the climatology used was derived from the ECMWF archives of the five years leading up to December 1983. This climatology is based on a comparatively short period of time and therefore probably contaminated by a degree of interannual variability. However, it was preferred to the NMC-NCAR longer-term climatology for which there are doubts about the quality of the analysis algorithms employed in constructing the objective analyses on which it is based. Although it is impossible to separate the effects of data availability, analysis quality and interannual

variability, the objective scores computed using the NMC-NCAR climate appeared to be consistently biased towards too high values due to very persistent features of the anomaly fields located over oceanic areas.

3.2 The skill of the LAF forecasts

The skill of the LAF forecasts can be compared with the DET forecasts by computing the RMS error and the anomaly correlation coefficient (ACC) between forecast and analysed fields as a function of forecast time, averaged over the four cases.

3.2.1 RMS error and anomaly correlation coefficient

Figs. 3.2 and 3.3 show the ACC and RMS error for T850 and Z500 for the Northern Hemisphere, along with the persistence scores.

These results indicate the following

- T850 is more "forecastable" than Z500 and for the T42 forecasts it outperforms persistence for the entire forecast period. The average ACC for the 30-day mean fields using the T42 SCL (LAF corrected for the systematic model error) has a value of 0.5.

- The most important element affecting objective forecast performance is the model's horizontal resolution rather than the statistical model error correction or the use of the LAF technique. However both techniques have a constant positive impact in reducing RMS errors. For T850 there is a sizeable improvement in ACC going from DET to LAF and, for T21, from LAF to SCL. Taking all the results into account, the T42 SCL gives the overall best forecasts.

- The LAF ACC does not decrease monotonically with forecast time, but shows a minimum for the 11-20 day interval for Z500 and for the 16-25 day interval for T850. For the T21 model in particular, the evidence of such 'return' of skill is emphasized by the systematic error correction technique. Similar behaviour in extended range forecasting experiments has been previously reported, e.g. Miyakoda et al. (1983), Cubasch and Wiin-Nielsen (1986), but its causes are still very much under debate.

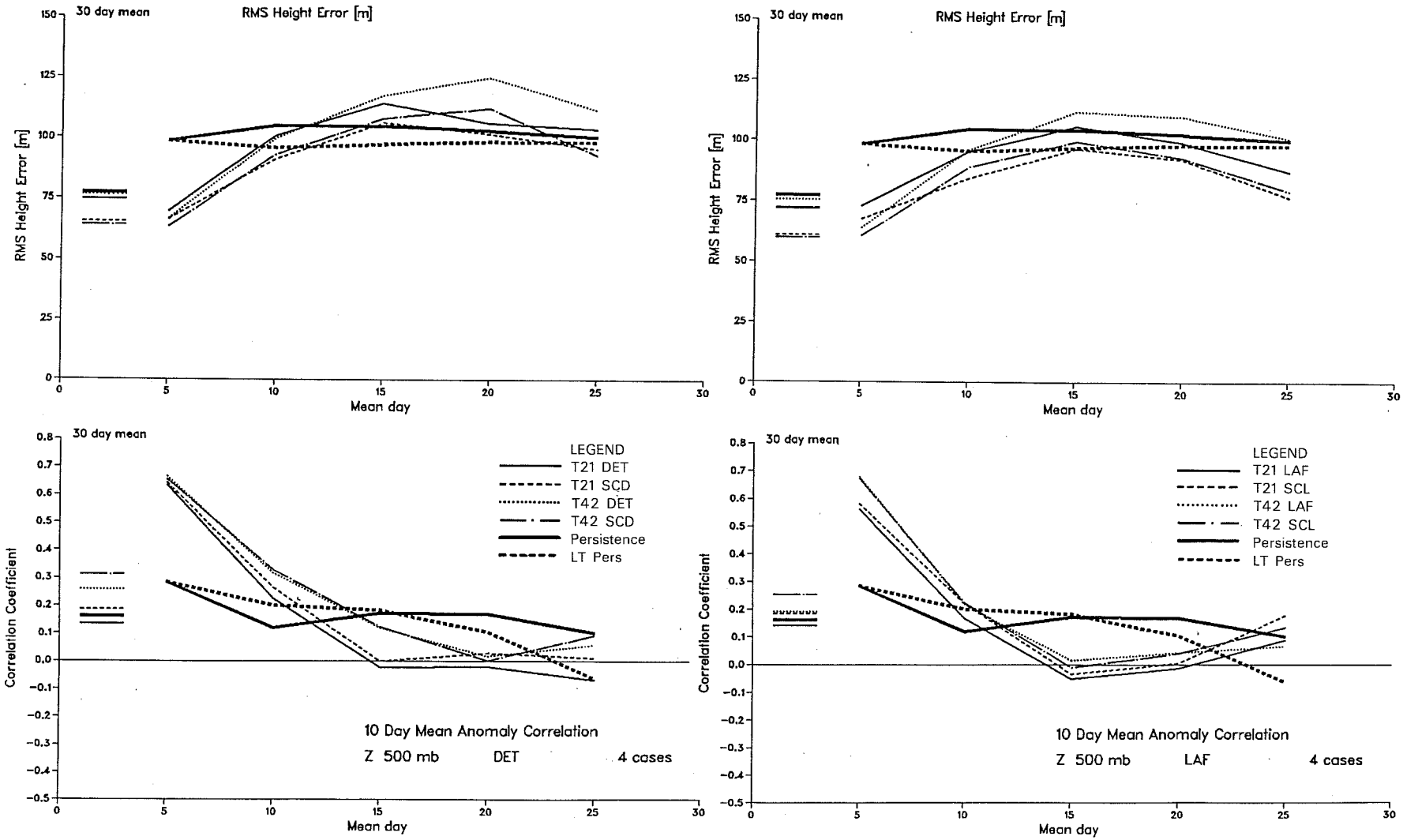
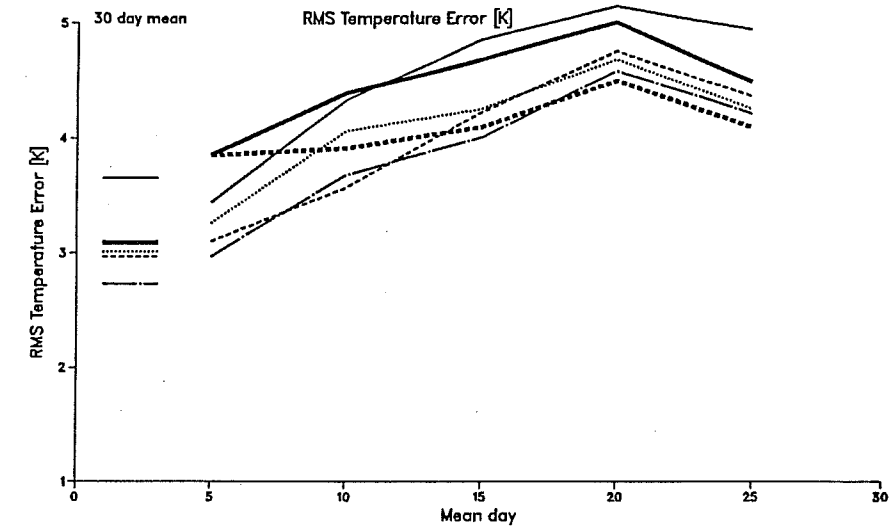
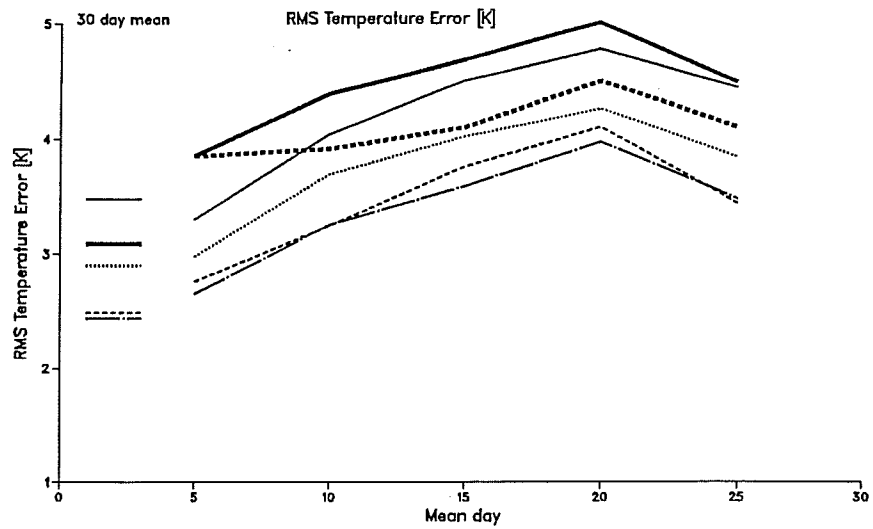
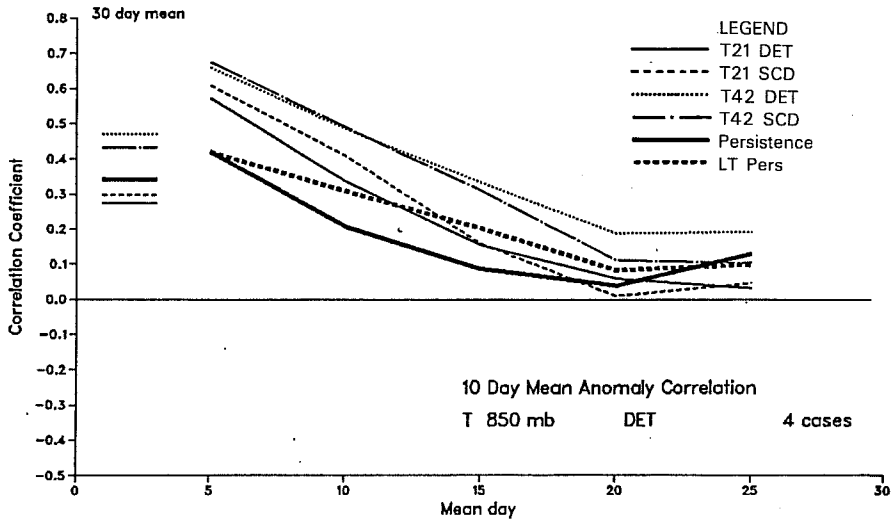


Fig. 3.2 RMS error (top) and ACC (bottom) for Z500. Average of the four cases of T21 and T42 forecasts. Left panels: DET: Deterministic forecasts. SCD: statistically corrected deterministic. Right panels: LAF: Lagged-Average Forecasts. SCL: statistically corrected LAF forecasts.







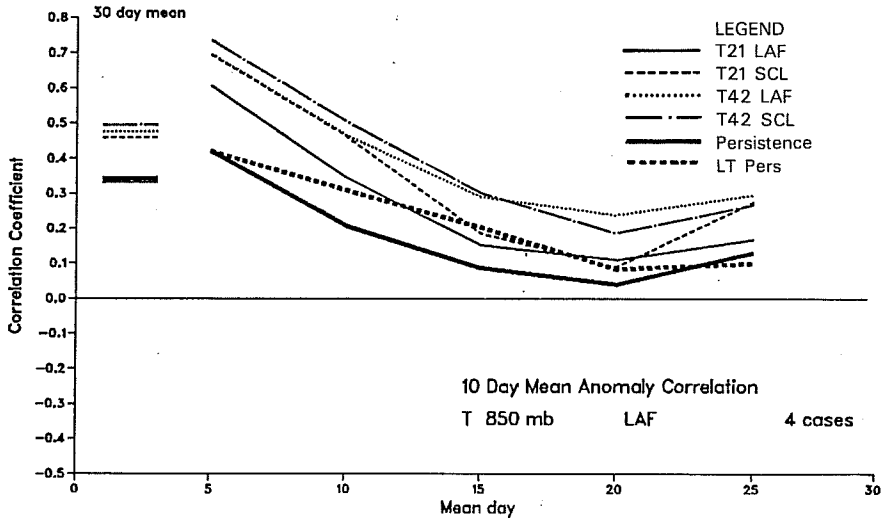


Fig. 3.3 As Fig. 3.2 but for T850.

3.2.2 Variablity of forecast skill

The variability in forecast skill within our limited sample will now briefly be discussed. It has already been seen that the average model skill in forecasting T850 and Z500 is quite different; this is also true for the case-to-case variability. Consider the 30-day mean ACC of the T42 SCL: while for T850 the mean value of .50 comes from the average of four values .43, .57, .44 and .54, the value of .25 for the Z500 is obtained from the four single values of -.05, .58, .-.04 and .52. For the (two) high skill cases, therefore, the results for the two variables are quite comparable. However, for the two low skill cases the Z500 skill drops to zero while the T850 skill only partially decreases. A possible explanation for these results is that the higher model skill in forecasting T850 could derive from the model's capability to represent the thermal interaction with the underlying surface (it should be noted however that during the 1983/84 winter there were no substantial SST anomalies). To verify such an idea, the vertical behaviour of the ACC of temperature as a function of forecast time was examined. It was found that the temperature ACC, while almost uniform with height during the first 10 days of the forecast, decreases with height during the last 10 days, going through an intermediate phase (during days 10 to 20) where the ACC has a minimum at 500 mb. This appears to be a clear indication of a positive effect of surface thermal forcing.

Now the question can be posed: is it possible to find, at least a posteriori, an indicator of predictability? Although our sample of only four cases cannot provide a proper basis for statistical reliability, it is interesting to compute the RMS variability of the 10-day mean 500 mb height fields with respect to the monthly means and examine the ratio of this to the RMS amplitude of the mean monthly anomaly. The results for observed data and for the two forecast models (after correction for the SE) are shown in Table 3.1. The observed ratio variability/mean appears to be clearly smaller in the two good forecast cases (January and March 1984) than in the two bad cases (December 1983 and February 1984). The mean monthly anomaly appears, therefore, to be more predictable the more persistent it is. This has two consequences. Firstly, it is evident that the (comparatively) low resolution models used for these experiments have difficulties in correctly reproducing the internal dynamic variability of the atmosphere on the monthly time scale.

Table 3.1a Z500 (m)

INIT. DATE	OBS		T21 SCD		T42 SCD	
	Mean	Variability	Mean	Variability	Mean	Variability
15/12/83	55	69 (1.25)	42	58 (1.39)	52	51 (.97)
17/1/84	71	58 (.82)	41	52 (1.27)	60	57 (.94)
19/2/84	52	54 (1.05)	46	45 (.99)	42	55 (1.30)
23/3/84	53	43 (.81)	38	38 (1.00)	47	51 (1.08)

Table 3.1b T850 (°K)

INIT. DATE	OBS		T21 SCD		T42 SCD	
	Mean	Variability	Mean	Variability	Mean	Variability
15/12/83	2.5	2.7 (1.07)	2.3	1.8 (.80)	2.7	2.2 (.81)
17/1/84	3.3	2.3 (.70)	2.1	1.9 (.90)	2.7	2.4 (.91)
19/2/84	2.3	2.5 (1.09)	2.5	2.0 (.82)	2.2	2.4 (1.08)
23/3/84	2.8	1.9 (.68)	2.0	1.8 (.93)	2.2	2.1 (.99)

Table 3.1 Mean and variability of observed and forecast (T21 SCD and T42 SCD) anomalies of Z500 (3.1a, meters) and T850 (3.1b, °K) for the four forecast experiments. The mean is the spatial RMS of the 30-day mean fields. The variability is computed by evaluating the RMS of the departures of the running 10-day means from the 30-day mean. The ratio variability/mean is shown in brackets.

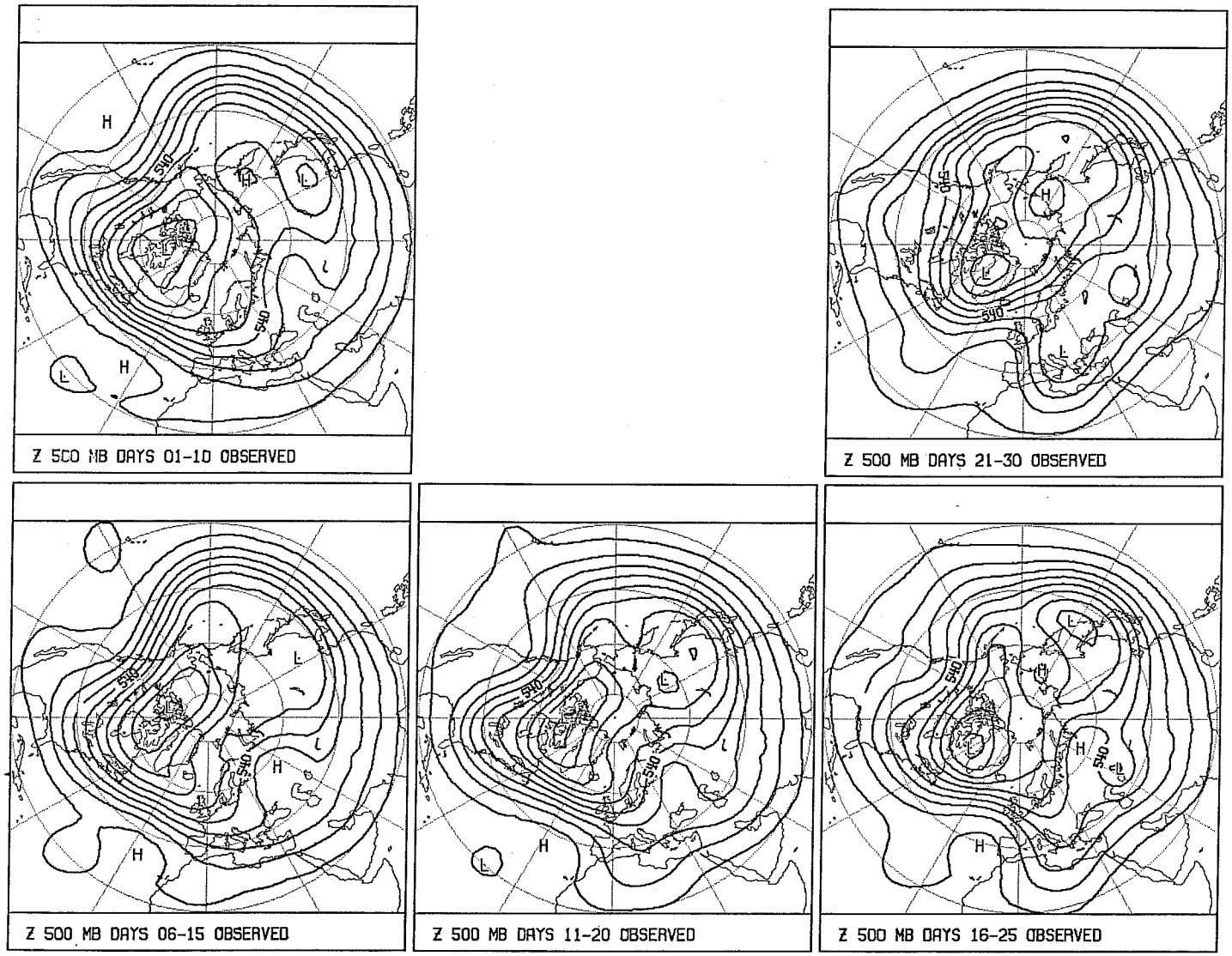


Fig. 3.4 10-day mean observed Z500 fields at 5 day intervals, corresponding to the verification periods of the 17.1.1984 experiment.

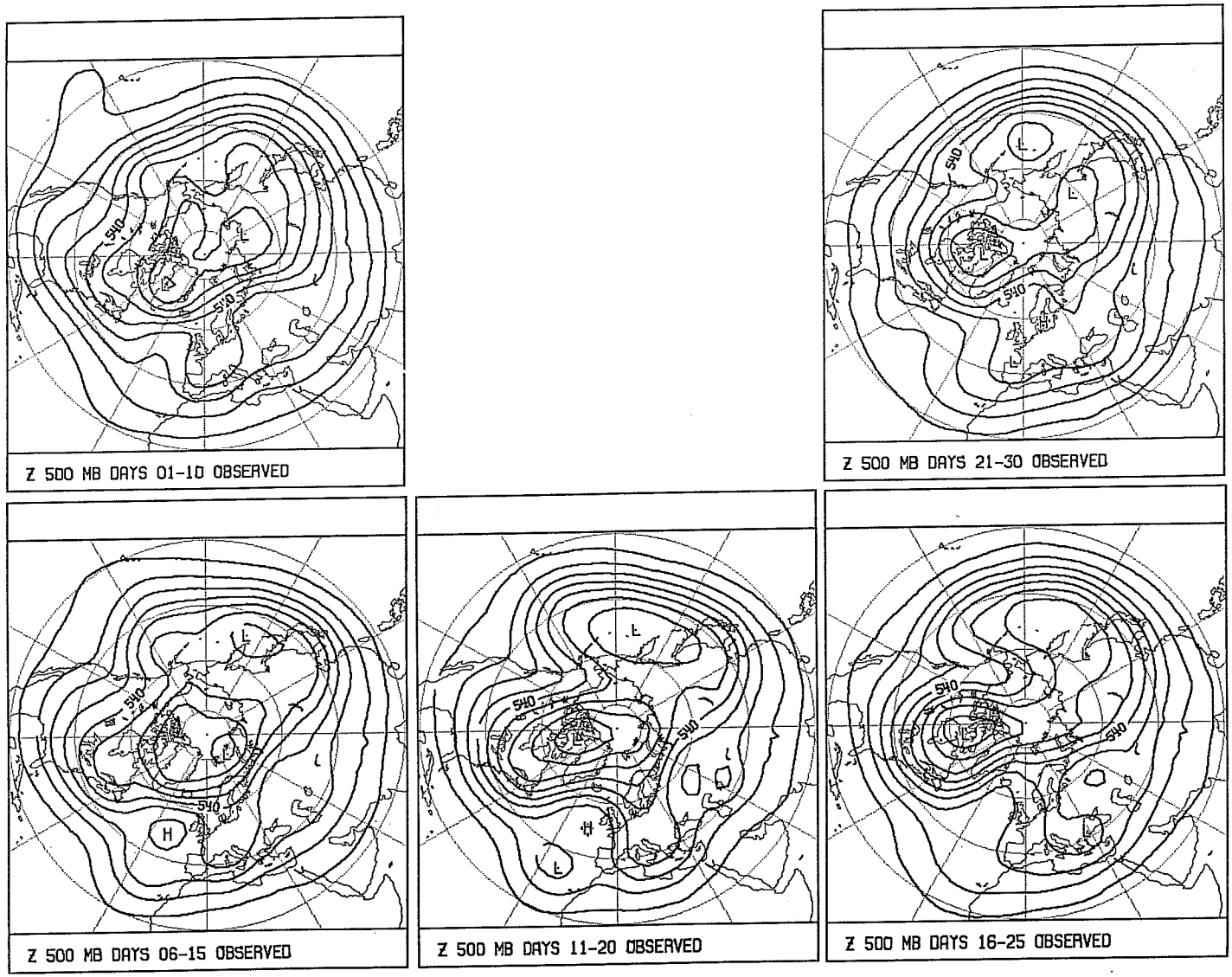


Fig. 3.5 As Fig. 3.4 but for experiment starting 19.2.1984.

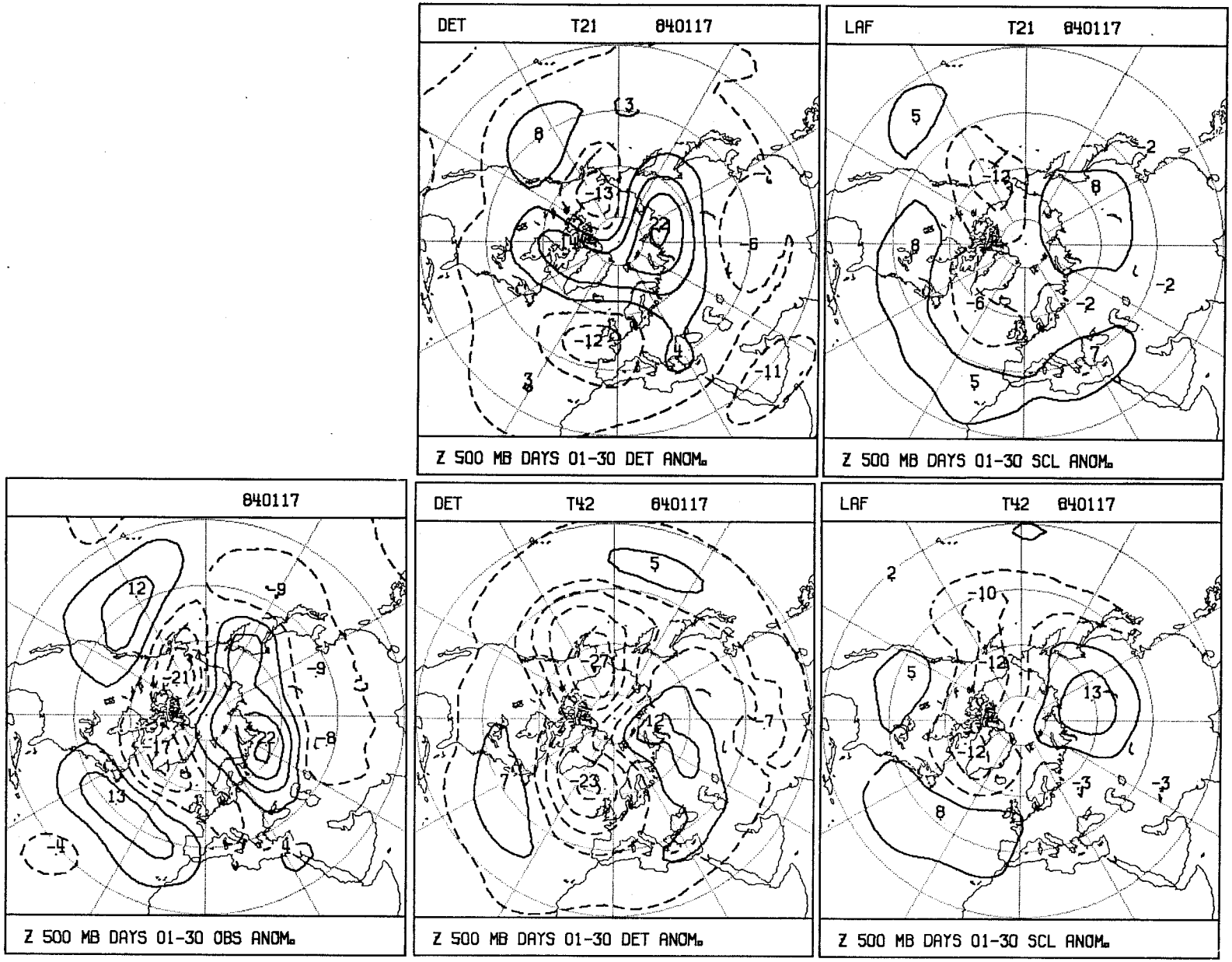


Fig. 3.6 Mean 30-day observed Z500 anomaly (bottom left), DET anomaly (Centre) and SCL anomaly (right) for T21 forecasts (top) and T42 forecasts (bottom). Experiment starting 17.1.1984.

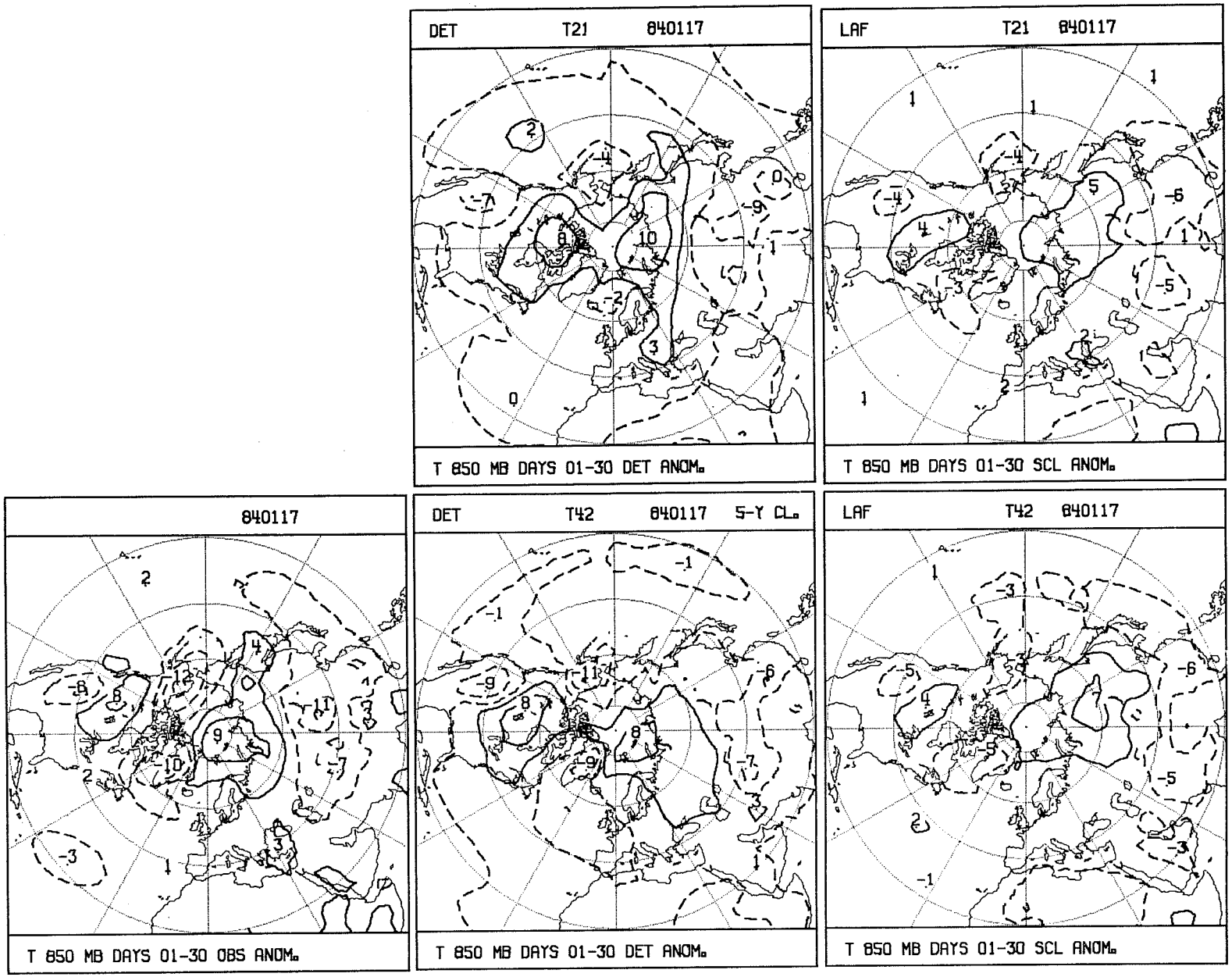


Fig. 3.7 As Fig. 3.6 but for T850.

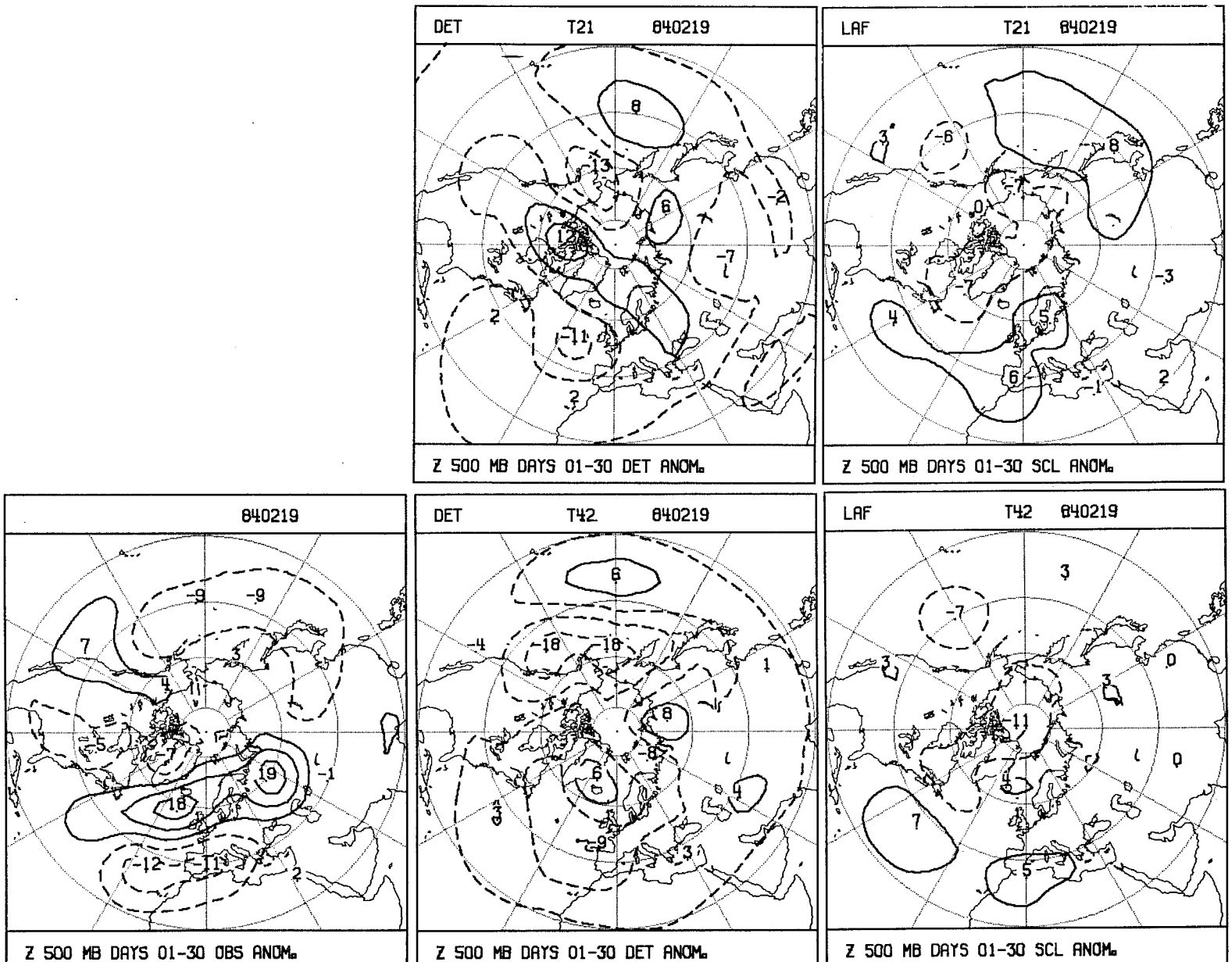


Fig. 3.8 As Fig. 3.6 but for the experiment starting 19.2.1984.

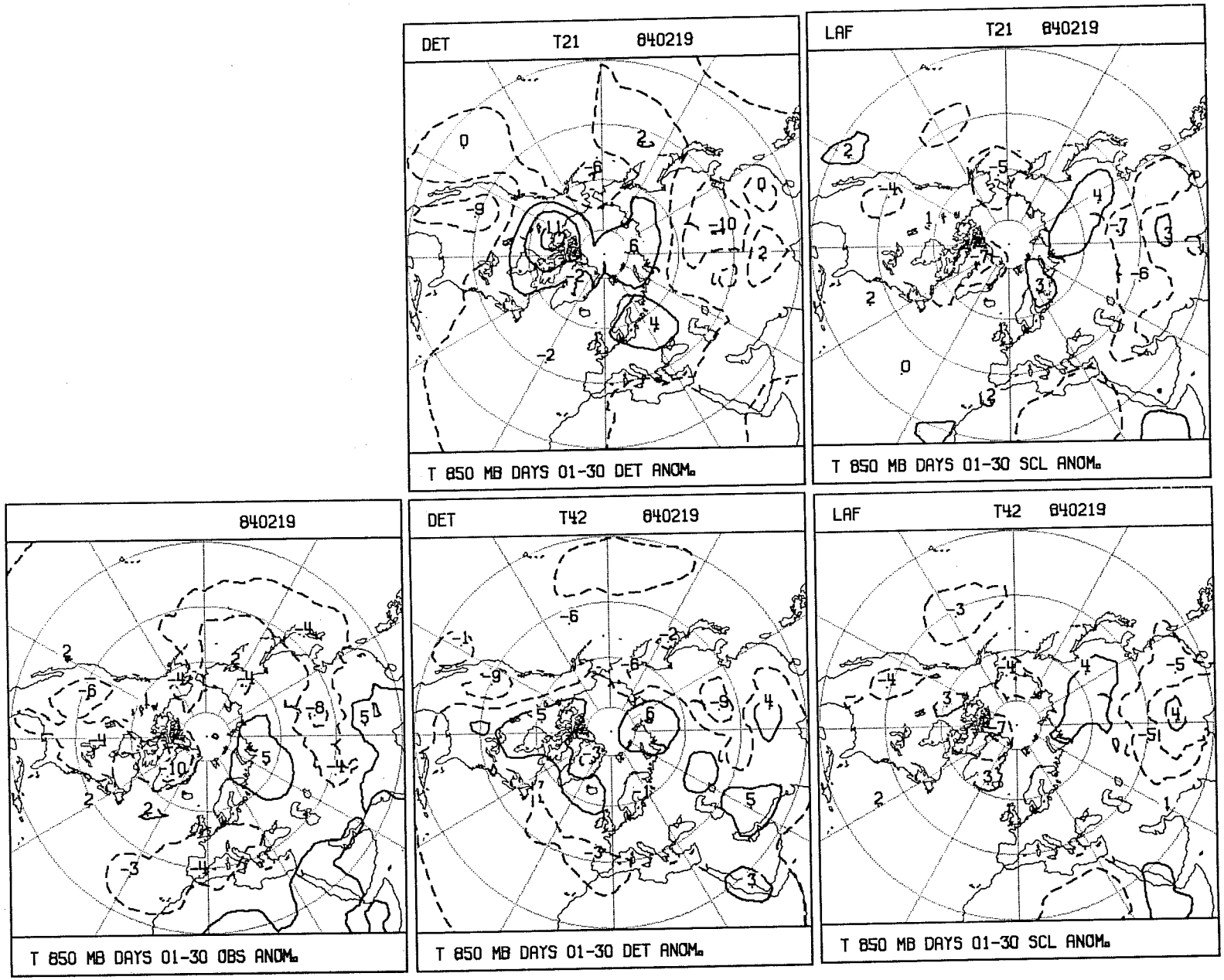


Fig. 3.9 As Fig. 3.7 but for the experiment starting 19.7.1984.

Secondly, since anomalous boundary forcing are usually conducive to increased persistence of atmospheric anomalies, the idea of a positive influence of boundary forcing on atmospheric predictability is strengthened.

Amongst the physical processes that are invoked to explain the low-frequency variability of the atmosphere is (linear and) non-linear resonance. This has been recently found to be (Benzi et al., 1986) a possible source of multiple equilibria states of the Northern Hemisphere extratropical atmosphere as diagnosed for example, by Sutera (1986) and Hansen and Sutera (1986). Blocked and zonal situations would then be explained as metastable equiblibrium states of the extratropics involving a relatively high and low amplitude of the planetary scale waves respectively. Sutera (private communication) has suggested that a substantial proportion of the global models' errors in the medium range might be associated with the inability to represent transitions between zonal states and high planetary wave amplitude states.

Supporting evidence to such ideas comes from comparing the Z500 10-day mean observed fields from the two January and February 1984 cases shown in Figs. 3.4 and 3.5. In the first case (the one with the highest score) the planetary scale waves remained at a high amplitude for almost the entire 30-day period, though they decreased towards the end. However, in the second case, a very poor forecast, the amplitude of the large scale waves was small during the early part of the month but amplified strongly during the central period which coincided with a blocked situation over the Eastern-Atlantic sector and a large amplitude Rockies ridge. Figures 3.6 to 3.9 show, for comparison, the observed and forecast (for DET and SCL) anomalies for the 30-day mean fields for the same two experiments. Their comparatively good and bad (respectively) synoptic value is evident and confirms the results of the objective scores.

All this pinpoints the need for more systematic studies of the predictability and 'forecastability' dependence upon atmospheric large scale regimes, both in the medium and extended range.

3.3 A test on the impact of SSTs

The LAF results discussed in the preceding sections did suggest that surface boundary (thermal) forcing might be responsible for increasing the predictability of the atmosphere and, consequently, the model's forecasting skill (at least for those large-scale patterns that are expected to be more influenced by the atmosphere's thermal interaction with the underlying surface). Since, however, all the LAF experiments were performed during the 1983/84 winter period, during which the anomalous boundary forcing due to SST anomalies was not particularly strong, it seemed interesting to perform a further experiment for a period in which such forcing was likely to be of importance. The winter 1982/83, characterized by a strong El-Nino event, provided an ideal example on which to test the impact of surface forcing in the ECMWF spectral global model. Two sets of model integrations (both LAF and DET) will be described here; the initial date for all integrations was 19 January 1983, 1200 GMT. One set of model runs (CLI) used climatological SSTs, while the other (OBS) was integrated using an SST field constructed by adding to the climatological field the mean anomaly for January 1983, as analysed by the NOAA Climate Analysis Center (Reynolds, 1983). The model used was the ECMWF T42 spectral model and all integrations were carried out to 60 days. In evaluating results, LAF ensembles are compared to the most recent of the nine ensemble elements, referred to as the deterministic forecast (DET). Statistically corrected (SC) versions of all integrations were also produced by subtracting from the forecast fields an estimate of the model's systematic error produced by averaging the forecast error of an independent set of 10 60-day integrations carried out from randomly chosen initial conditions during the 1983/84 and 1984/85 winters with observed SSTs.

Fig. 3.10 shows the correlation coefficients between 500 mb height forecast and observed anomaly fields averaged over running 10-day periods, and the corresponding RMS errors for DET and LAF forecasts, using both OBS and CLI SSTs and for corrected (SC) and uncorrected integrations. These time series can also be compared with the skill of the persistence forecast (PER).

Several interesting points emerge from the analysis of such skill scores. Firstly, there is an anomalously high performance for persistence during this

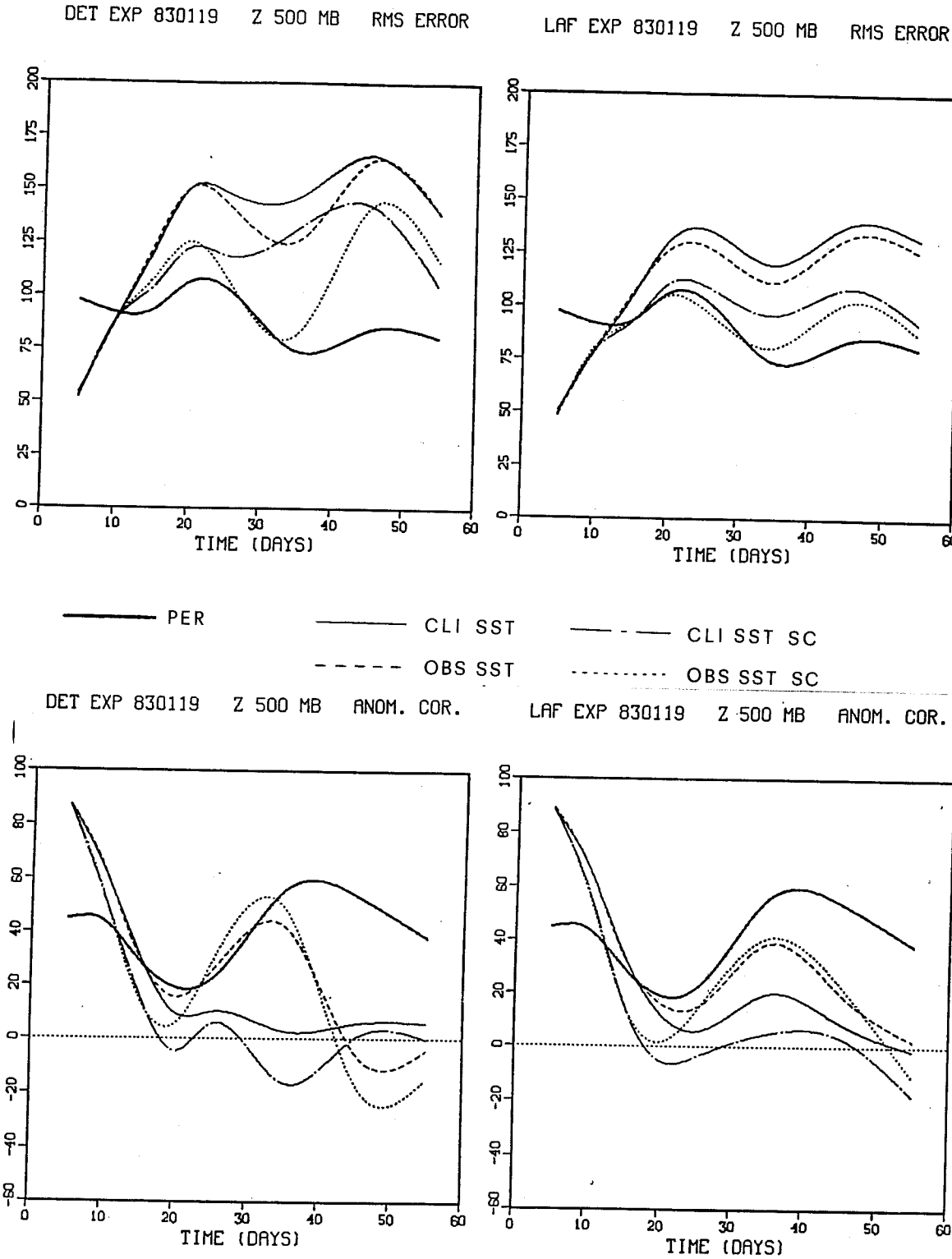


Fig. 3.10 RMS errors (top panels, meters) and anomaly correlation coefficients (bottom panels, %) of 10-day running means of 500 mb height. Deterministic (DET) experiments, left panels. LAF experiments, right panels. The skill scores are shown as a function of forecast time and for persistence forecasts (thick continuous lines), climatological SSTs forecasts (thin continuous lines), observed SSTs forecasts (dashed lines), climatological SSTs statistically corrected forecasts (dashed-dotted lines) and observed SSTs statistically corrected forecasts (dotted lines). All experiments starting on 19.1.1983 1200 GMT.

period. The anomaly correlation of persistence reaches the 60% level around day 40. This seems to be fairly characteristic of this El-Nino period, during which the characteristic PNA (Pacific/North American) signature remains a feature of the Northern Hemisphere circulation for a long period of time. Secondly, there is a clear positive impact of the SSTs in the OBS integrations during the day 20 to 50 forecast period. Around the day 30 mark, this improvement is more evident in the DET forecasts than in the LAF, but in the LAF the positive impact of observed SSTs lasts until the end of the integration period.

A further interesting point is the very sizeable "return" of skill again evident in the OBS integrations, suggesting that such a return of skill might in this case be linked to a higher atmospheric predictability induced by surface thermal forcing. It should be noted that the CLI LAF forecast (but not the DET) also shows a return of skill, albeit much weaker. This could be due to the fact that the atmospheric initial conditions "know" about the surface anomalous forcing even if, during the model integration, this knowledge is eliminated. It is plausible, however, that the residual predictability associated with the initial conditions (and their "knowing" about El-Nino) can only be detected after the effects of the smaller scale random forecast errors have been partially eliminated by applying the LAF technique.

The impact of LAF and of the systematic error correction are both clearly positive on the RMS scores, but not so on the ACC. It should be noted that, during the first 20 days of model integration, the growth of the error of the uncorrected LAFs is almost identical to the one shown by the DET forecasts and the impact of the SC is negative on both. This essentially means that, during this interval, the T42 forecast error depends predominantly upon the initial conditions. Thereafter, the beneficial impact of the LAF technique (and of the SE correction) increases with forecast time, so much so that the SC LAF integrations show the maximum forecast error around day 20.

The amplitude of the "return" in forecast skill taking place in this particular set of experiments, and the fact that it takes place also in the DET forecasts with OBS SSTs, suggest it is due to the failure of the model

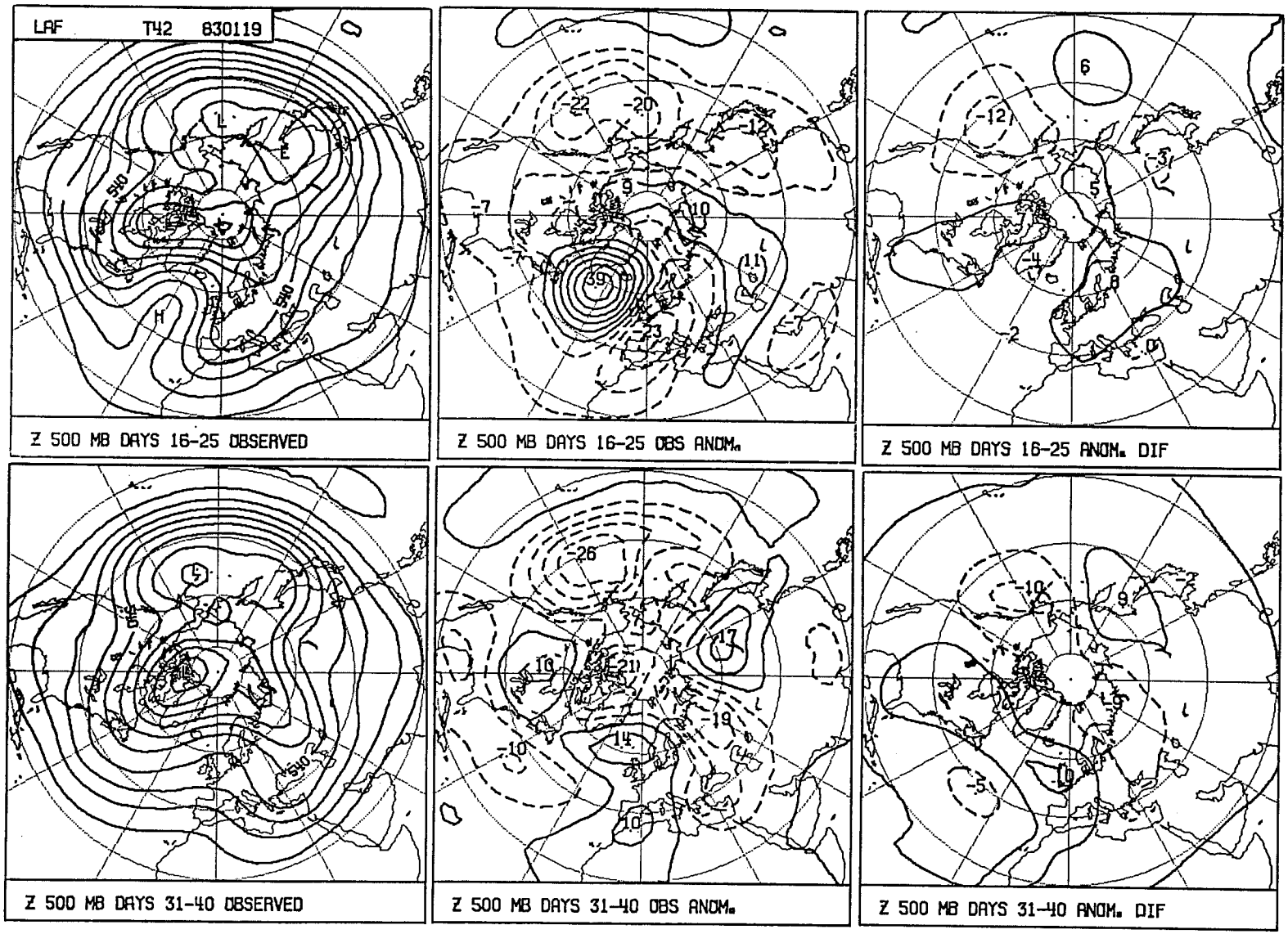


Fig. 3.11 Mean observed 500 mb height fields (left panels), observed anomaly (centre panels) and difference between forecast anomalies (OBS-CLI, right panels) for the two forecast intervals days 16-25 (top panels) and days 31-40 (bottom panels). All experiments are LAF starting on 19.1.1983 1200 GMT.

to represent a synoptic development of limited lifetime that, in turn, affects little the evolution of the global circulation on a longer (>20 days) time scale.

It then becomes interesting to evaluate what proportion of the observed atmospheric anomaly can be thought of as a response to the surface boundary forcing and how much can be explained by internal dynamics. We would then expect internal dynamics to dominate the early, low skill, part of the forecast period and the boundary forcing response to be more important during the latter part of the 60 days. We can estimate this by comparing the total observed atmospheric anomaly with the difference fields between the two model integrations CLI and OBS, assuming that the latter fields might be representative, albeit approximately, of the atmospheric response to surface forcing. To have more stable and statistically significant fields we will use the difference between the two LAF (sets of 9) integrations. Fig. 3.11 shows this comparison for the Z500 averaged over the two time intervals day 16-25 and day 31-40, for which the OBS LAF shows, respectively, a relative minimum and a relative maximum of forecast skill. It is evident that, during the first interval, the observed anomaly is dominated by the onset and life-span of an Eastern Atlantic blocking episode. This blocking does not seem to be too influenced by the tropical Pacific thermal forcing, since both LAF integrations fail to model its development: in such conditions, the similarity between the observed anomaly map and the forecast difference (OBS-CLI) map is virtually non-existent over most of the Northern Hemisphere, with the exception of the North Pacific. During the second half of the integration period, on the contrary, the blocking pattern disappears and the observed anomaly and the forecast difference map become highly correlated, suggesting that the atmospheric anomaly is mainly the product of the SST forcing.

In summary, it is arguable that, during the central part of the integration period, the Northern Hemisphere extratropical atmosphere has switched from a regime dominated by internal dynamics variability to a regime driven mainly by anomalous boundary forcing. The model integrations shown seem, furthermore, to be unable to reproduce the former while they show considerable skill in representing the latter.

3.4 Ensemble spread and forecast skill

In the previous sections, the results of the lagged-average predictions were evaluated in terms of the skill of the ensemble mean forecast; it was shown that the improvement obtained with the LAF method in comparison with a purely deterministic forecast was much more evident in the RMS errors than in the anomaly correlation coefficients. However, the anomaly correlation improves with increasing horizontal resolution, so that the actual advantage of the LAF technique might be questioned since for a given computational time an extended-range LAF is as expensive, in computer resources, as a deterministic forecast with a model of considerably higher resolution.

The perspective would change if it could be proven that the LAF produces not only a marginally better forecast than a deterministic prediction, but can also provide an estimate of the reliability of the forecast that can be deduced from the spread of the ensemble. Theoretically, for a perfect model, the spread of the ensemble could be considered as a measure of the instability of the circulation regime that is present in the initial conditions. A correlation between the spread and the error of the mean forecast should then be expected.

3.4.1 Effects of systematic errors

In practice, the effects of the model deficiencies can strongly affect the spread-skill relationship. Rinne and Karhila (1974) and Pitcher (1977) found that the inclusion of a random forcing term (used to parameterize the "external" sources of error) in the equations of the barotropic models used in their stochastic-dynamic experiments was necessary to obtain a reasonable correspondence between the standard deviations of the estimated quantities and their errors. Also Hoffman and Kalnay (1983) had to take into account the existence of model-generated errors in developing a statistical relationship between forecast error and ensemble spread.

The quantitative effects of a model systematic error on the correlation between spread and skill are difficult to evaluate. Once again, very much depends on the actual "systematic" behaviour of this error. If it was simply a superimposed error pattern that depends on forecast time but very little on the initial circulation regime, then a general reduction of the spread

(proportionally greater for large forecast times) could be expected; apart from a time-dependent scaling factor, the correlation between skill and spread should be maintained. If, on the other hand, the "systematic" error strongly depends on the initial conditions, then an ensemble of forecasts dominated by the systematic error (whose skill is expected to be very poor) might show a considerably lower spread than another set started from initial conditions that do not give rise to strong biases in the forecasts.

3.4.2 Evaluation of the local correspondence between forecast spread and forecast skill

The small number of LAF experiments described here is obviously not sufficient to investigate the correlation between global or hemispheric measures of spread and errors as shown in Hoffman and Kalnay (1983). However, it is possible to try to test whether there exists a local correspondence between these quantities, taking into account that both the errors and the spread are likely to be large over areas where the variability of the predicted fields is large.

For each grid point of a 3.75° x 3.75° regular lat-lon grid covering the Northern Hemisphere (from 22.5°N to 90°N), the RMS values (among the 4 experiments) both of the ensemble error and standard deviation of a given meteorological field were computed. (The ensemble standard deviation is the RMS difference between each of the 9 individual forecasts and their ensemble mean). Then a standardised error ($\hat{E}$) was computed for each grid point and LAF experiment as the ratio between the local error and its RMS value; in this way, $\hat{E}$ reflects only the local variability of the error among the 4 experiments and tends to be independent of the local error variance. In an analogous way, a standardised spread ($\hat{S}$) was defined for each grid point and experiment by dividing each local value of the standard deviation by its RMS value among the 4 experiments.

The range of $\hat{S}$ was divided in 4 classes:

class 1	$0.7 > \hat{S} \geq 0$
class 2	$1.0 > \hat{S} \geq 0.7$
class 3	$1.3 > \hat{S} \geq 1.0$
class 4	$\hat{S} \geq 1.3$

and all the values of $\hat{E}$ corresponding to a value of $\hat{S}$ falling in a given class were averaged in an RMS sense (an area weight proportional to the cosine of the latitude of the point was also used for this average).

Instead of comparing 4 spatially averaged values of error and spread, it is possible to compare 4 values of standardised error averaged over all the points and cases in which the value of the standardised spread fell in a given range: such a procedure may be able to reveal a correlation between spread and skill even if it exists only over certain areas.

The results of this computation, for both the LAFs and the SCLs, are shown in Fig. 3.12 for the monthly means of geopotential height at 1000, 500 and 300 mb, and temperature at 850, 500 and 300 mb (obviously the subtraction of the systematic error affected the values of the errors but not those of the standard deviations). In comparing the results of the LAFs with those of the SCLs, it is necessary to take into account that the central classes of $\hat{S}$ included more points than the extreme ones.

3.4.3 Discussion of the results

The results can be summarised as follows:

- Apart from the Z1000, a correlation between spread and error is not present for any variable or level using the T21 model, even when the systematic error is subtracted.

- A clear correlation can be found for the height fields predicted by the T42 model, especially at 500 and 300 mb; at these level, a slight correlation also exists for the temperature fields. The removal of the systematic error is in general beneficial: the greatest improvement is found for Z500, while for the T850 only the uncorrected LAF shows a marginal correlation.

It is interesting to see how the correlation evolves with the forecast time: Fig. 3.13 shows the results for the 10-day mean fields (as well as for the monthly mean) of Z1000 and Z500 deduced from the T42 SCL. At both levels, the

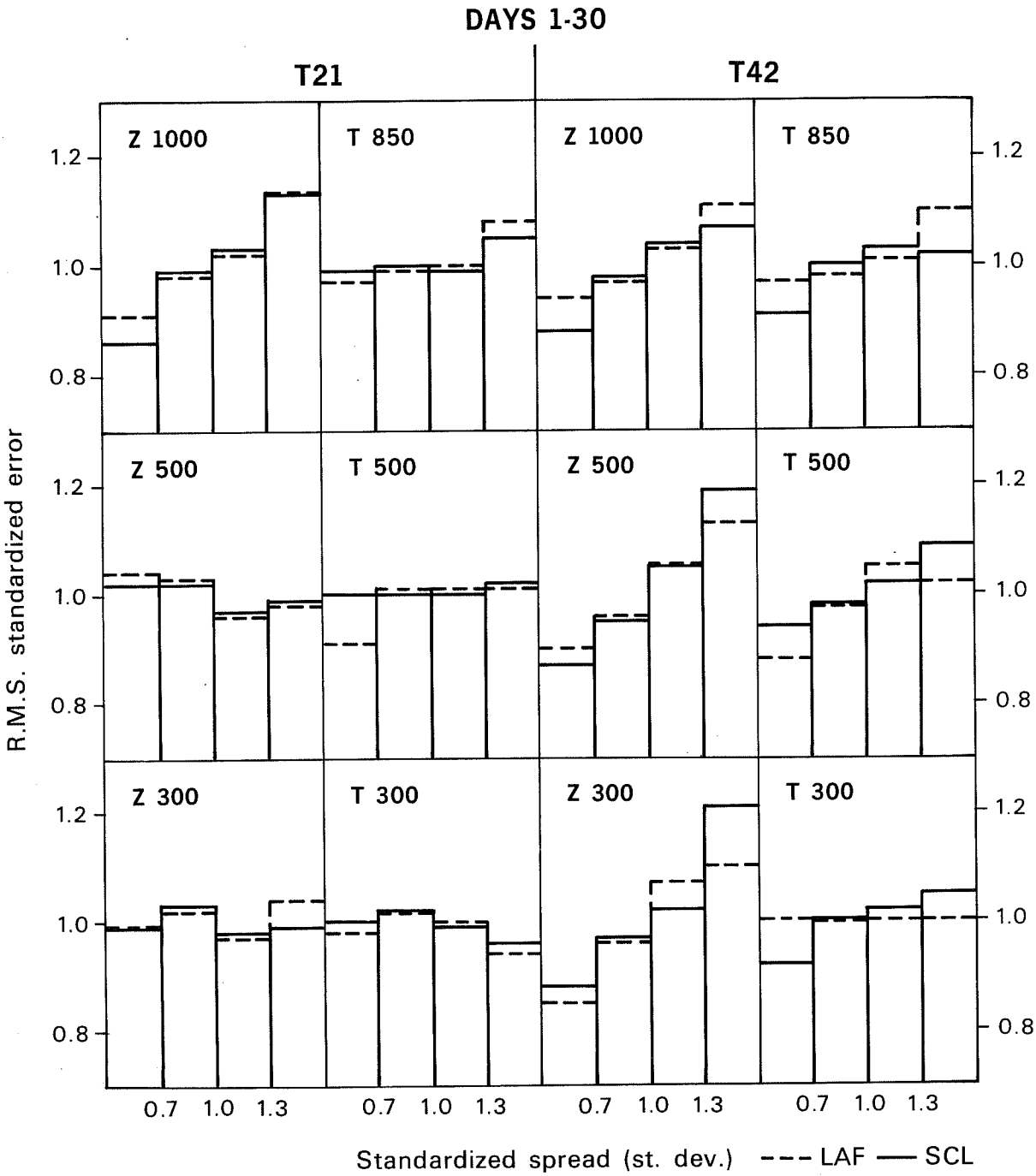


Fig. 3.12 Mean day 1 to 30 spread-error relationship for four classes of standardized spread (standard deviation). Top 1000 (850) mb, centre 500 mb, bottom 300 mb, for both height and temperature. Both LAF experiments (dashed) and SCL experiments (full) are shown. For more explanations see section 3.4.

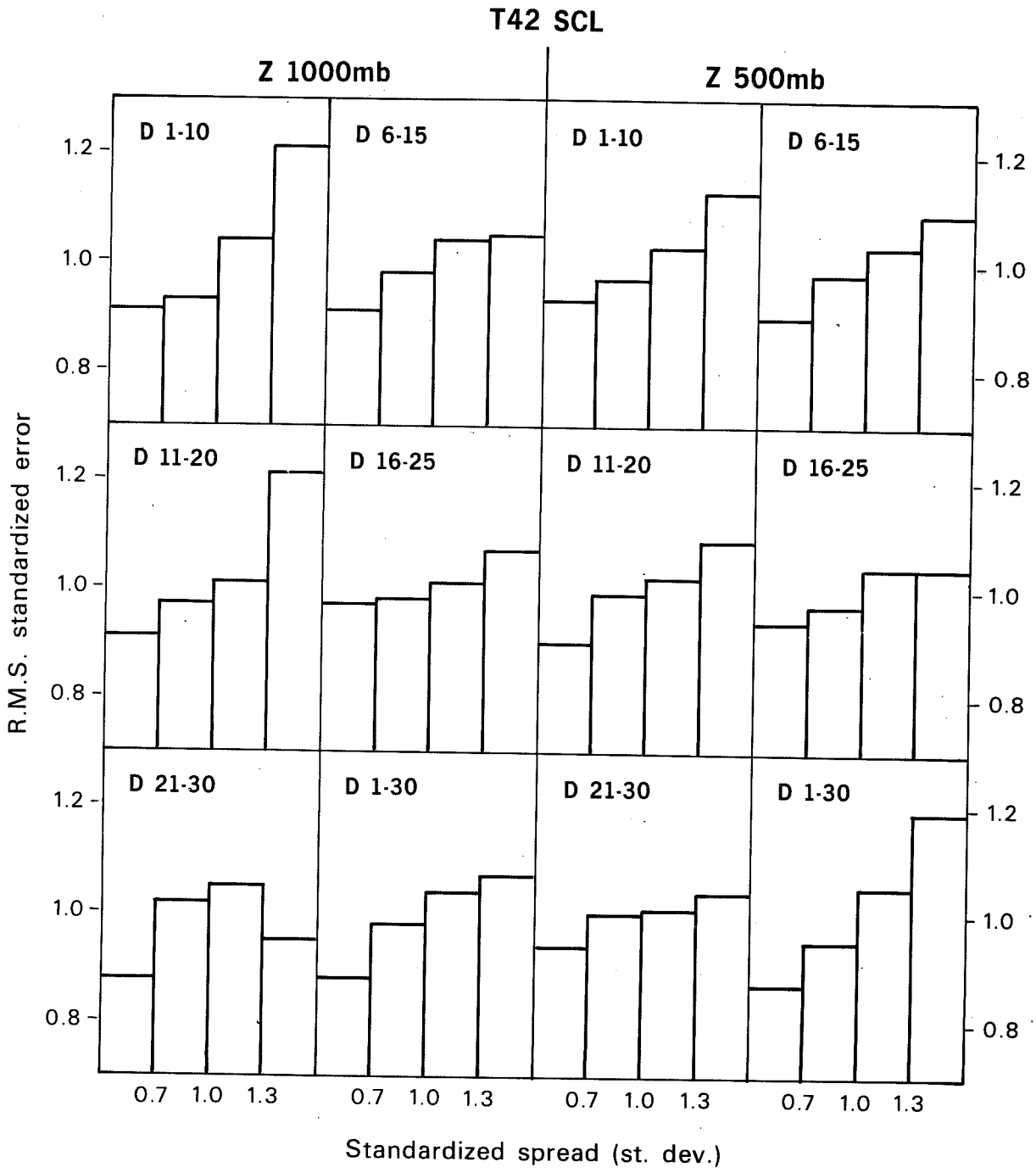


Fig. 3.13 10-day mean histograms of spread-error relationship for the same four classes of standardized spread for T42 SCL forecasts. Top dd 1-10, centre dd 11-20, bottom dd 21-30. For more explanations see section 3.4.

correlation decreases with forecast time, but it is more stable at 500 mb, even if slightly lower in the beginning of the forecast. At this level, the correlation is stronger for the monthly mean than for any of the 10-day means, while at 1000 mb the results for the 30-day period are very similar to those obtained in the 6-15 day period.

The fact that the correlation is much larger in the T42 results is surprising in some ways, though not in others. It seems obvious that the spread of an ensemble of forecasts can be a good indicator of the probable range of the error only if the variability of the model is comparable with that of the real atmosphere. Therefore, the fact that the T42 shows a larger variability than the T21 can partially explain these results. On the other hand, the mean error of the T42 is less systematic than that of the T21, and on this basis a better correlation for the T21 model should be expected. In these results, the former effect appears to be more important than the latter and the superiority of the T42 model over the T21 is confirmed.

In conclusion, the higher computational cost required by the LAF technique (and to a lesser degree by the correction of the systematic error) appears to be justified not only by a moderate improvement in the forecast skill but also by the existence of a significant correlation between the spread of the ensemble and the error of the mean forecast, at least for some variables and time intervals. The necessary condition for the existence of such a correlation is an internal variability of the forecast model comparable with that of the real atmosphere. The importance that the users of the NWP products attribute to a prediction of the forecast skill is probably the ultimate basis on which to decide if stochastic-dynamic predictions are worth becoming an operational tool, but their superiority over deterministic forecasts is undoubtedly much more than a theoretical speculation.

3.5 Summary

An attempt has been made to assess the relative improvements that can be achieved on purely deterministic numerical forecasting by either using the LAF technique and/or by attempting to correct the model's systematic bias by subtracting an estimate of it derived from an independent set of model integrations during a similar seasonal period. The results of this assessment can be summarized as follows.

- The skill of the forecasts, in particular those performed with the LAF technique, does not decrease monotonically with time, but has a minimum in the 11-20 day period for Z500 and in the 16-25 day period for T850. This effect could be partially ascribed to low-frequency periodicity of the atmosphere, as the similar recovery of the persistence forecasts for T850 would suggest.

- Both in an absolute sense and in comparison with persistence, the results for T850 are clearly better than those for Z500. For the former, the T42 SCL is better than persistence in all the verification intervals, both in terms of RMS error and of ACC (in a RMS sense it is also better than climate, apart from the 16-25 day period); for the latter, a consistent improvement over persistence, apart from the first ten days, can be seen only in the monthly mean.

- The LAF technique and the correction of the systematic error lead to a clear and constant improvement of the RMS error; this is most evident for T850 where a positive impact of the LAF method on the ACC can also be detected. The impact of the correction of the systematic error on the ACC is generally positive, but notable exceptions can be found among the deterministic forecasts performed with the T42 model. More evident, and without exception, is the improvement in the ACC due to the increase of horizontal resolution: the T42 model, even with no correction for its bias, has a consistently higher ACC than the corrected T21.

- Even though for some particular variable or score it is possible to find a forecast method that gives a comparable or slightly better performance than the T42 SCL, but with a lower computational cost, a comprehensive evaluation of the results shows that the T42 SCL has the highest skill among all the combinations of resolution/error filtering methods that we have explored; as pointed out earlier, for T850 the T42 SCL is more skillful than either climate or persistence for the whole monthly range. The SCL also has the advantage of providing an indicator of the reliability of the forecasts, since a considerable correlation between the spread of the forecast ensemble and the skill of the SCL was found for individual geographical locations.

- The results of a further experiment to test the impact of SSTs on the forecast skill during the El-Nino period is consistent with the idea that there is considerable predictability (and 'forecastability') coming from the boundaries, but that even in cases of strong forcing this can be easily obscured and overwhelmed if the model fails to represent an important synoptic event whose development is dominated by internal dynamics, e.g. the onset of blocking. When (and if) the effects of such low-frequency transient development fades away, the model seeems to be able to recover its capability to represent the effects of boundary forcing.

4. THE DETERMINISTIC FORECASTS

4.1 Description of the experiments

This section reports the results of a further study to investigate purely deterministic extended range forecasts which have been selected at regular 10 day intervals in order to simulate operational conditions. The database is the one generated to estimate the T21 and T42 models' systematic error used to correct the LAF experiments described earlier. All initial conditions come, therefore, from the four winters 1981-82 to 1984-85 and give a sample large enough to explore the interannual variability of the forecasting skill. The results of this study should therefore be representative of the typical performance of numerical extended range forecasts to be expected under operational conditions for the winter-spring period. However, only an estimate of the lower end of the forecasting skill can be obtained, due to the comparatively low resolution of the models employed.

The experiments consisted mainly of 30-day forecasts, though some 60-day forecasts were also carried out. The starting dates of the 38 cases are given in Table 4.1. The initial data have been derived from the operational ECMWF assimilation system by direct spectral truncation to the model resolution. Throughout the integrations the values at the lower boundary of the atmosphere, i.e. SST, deep soil moisture and deep soil temperature, have been kept constant. For 1981 and the first half of 1982, the SSTs have been set to their climatological values, while from winter 1982/83 onwards SSTs analysed by NMC have been used.

Corresponding persistence forecasts have also been evaluated based on either the 10- or 30-day average of the days preceding the initial date. These persistence forecasts provide a baseline against which the dynamical forecast can be compared.

1981	1982	1983	1984*	1985*
	1. 1.	1. 1.	1. 1.	1. 1.
	11. 1.	11. 1.	11. 1.	11. 1.
	21. 1.	21. 1.	21. 1.	21. 1.
	1. 2.	1. 2.	1. 2.	1. 2.
	11. 2.	11. 2.	11. 2.	11. 2.
	21. 2.	21. 2.		
	1. 3.	1. 3.		
	11. 3.	11. 3.		
	21. 3.	21. 3.		
	1. 4.	1. 4.		
	11. 4.	11. 4.		
2.12.	1.12.			
11.12.	11.12.			
21.12.	21.12.			

*60-day integrations

Table 4.1 The initial dates of the forecast experiments

4.2 Analysis of results

All experiments have again been evaluated in terms of 10 and 30 day averages. The objective verification has been carried out by calculating the ACC and the RMS errors of 10- and 30-day means for the northern hemisphere between 20°N and 90°N and the ensemble anomaly correlation over the 38 cases using the z-statistics (Seidman, 1981).

4.2.1 Intra-seasonal variability

For only two of the winters (1981/82, 1982/83) were enough cases integrated to evaluate the mean monthly forecast scores as a function of the initial date of the forecast. Fig.4.1 shows the correlation coefficients of 30-day means for the two model resolutions, as well as for persistence, for Z500 in the northern hemisphere. In both winters the persistence forecast scores between 10% and 60% with a smooth transition between periods with good and bad forecasts. The scores of the model forecasts appear to be uncorrelated with those based on persistence. The T42 model simulation nearly always stays above the 25% level and is therefore a more reliable predictor than persistence (at least for the monthly means) whilst the skill of the T21 model undergoes fluctuations which are as large as those found for the persistence forecast. The time evolution of the 10-day mean forecast skill (not shown) exhibits much more variation than for the 30-day mean forecasts, but a correlation between the quality of the persistence and model forecasts could not be found on this shorter timescale either.

4.2.2 Interannual variability of skill

The period between 1 January and 11 February is covered for all 4 years. This gives an opportunity to assess the interannual variability of the forecast quality based on a mean of 5 forecasts per year. The T42 model produces the best scores in terms of anomaly correlation and RMS error for the 500 mb height field (Fig.4.2) during 1985, which is also the worst year for the persistence forecast. The 1982/83 El-Nino winter produces persistence forecasts with considerable skill (particularly in the 30-day mean), but the T42 model forecast quality falls below the average. Conversely, for the T21 model, the 30-day mean scores are quite high. This contradictory result might be explained by the different structure of the systematic error of the

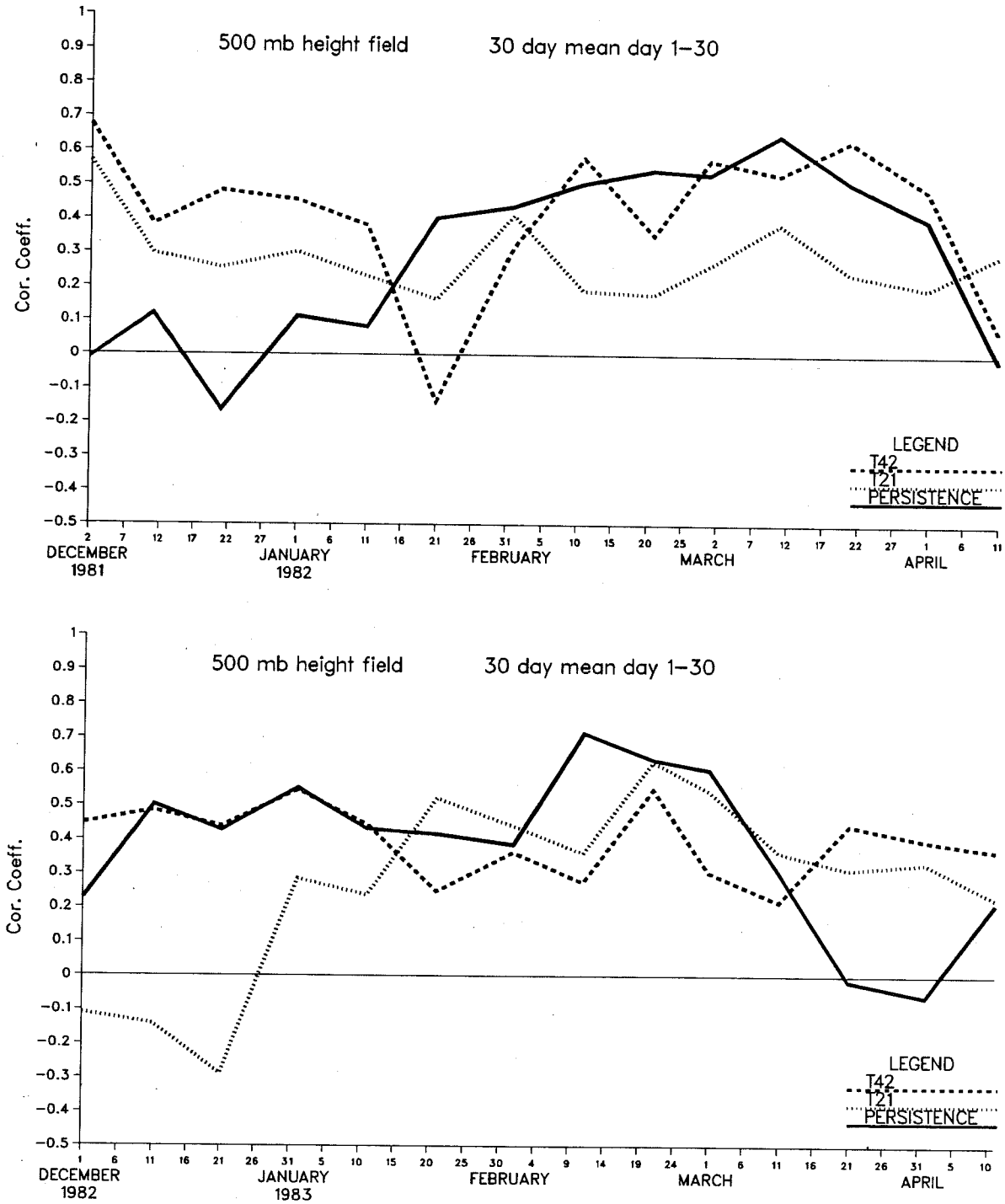


Fig. 4.1 The 30-day mean correlation coefficient for the Z500 as function of the initial day; a) experiments during winter 1981/1982; b) experiments during winter 1982/1983.

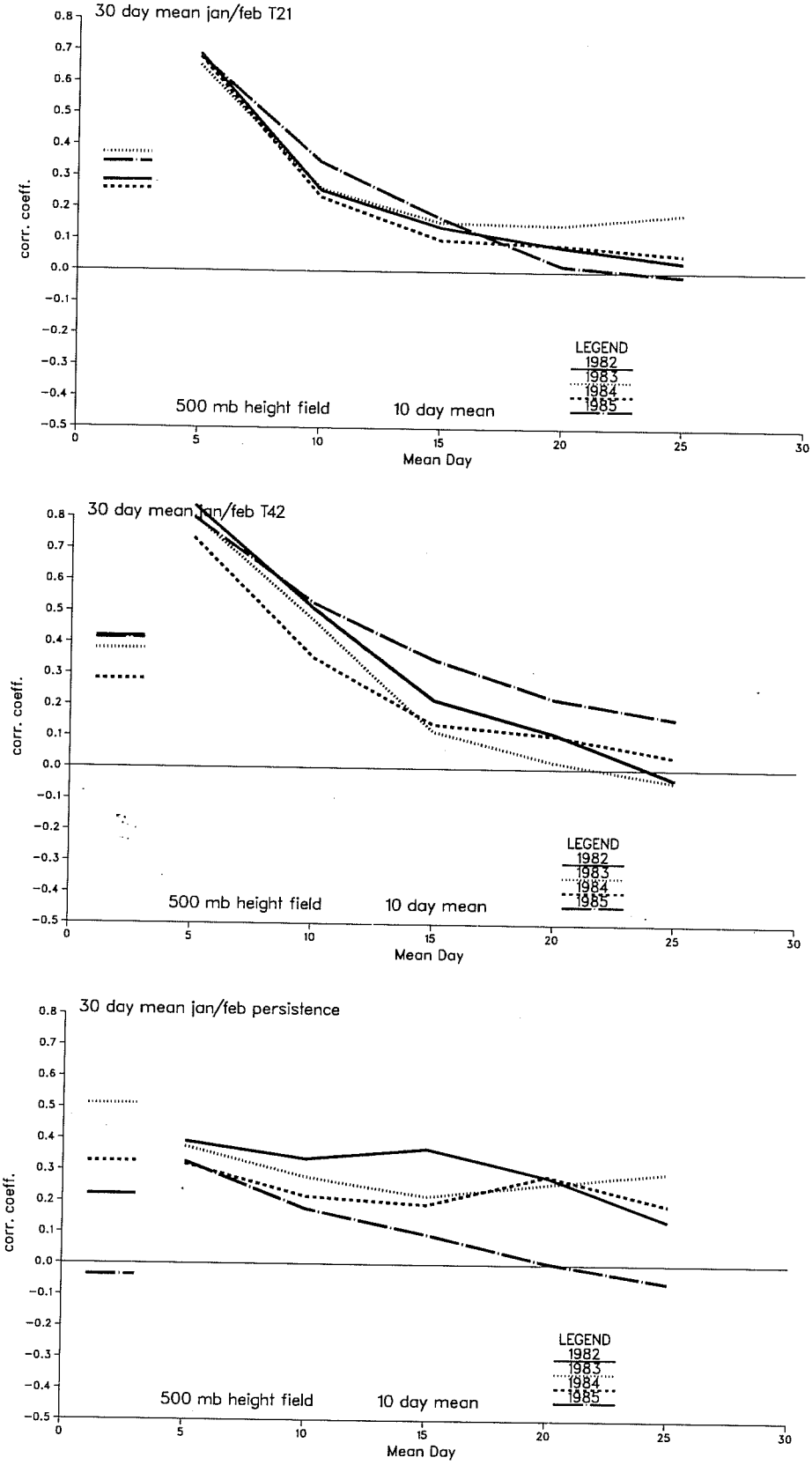


Fig. 4.2 The time evolution of the correlation coefficient for the Z500 averaged over the experiments started during the months of January and February for different years a) for the T21 model forecasts; b) for the T42 model forecasts; c) for the persistence forecast.

two models (see section 2 and Cubasch and Wiin-Nielsen, 1986). Regarding the better skill attained by the T42 model in the experiments of subsection 3.3, it should be remembered that a different, and more accurate, SST anomaly was used for those integrations. In the experiments described in this section, the El Nino signature was largely suppressed by the operational NMC SST analysis scheme.

The scores for the 850 mb temperature field confirm the findings for the height field.

4.2.3 The mean performance of the Z500 forecasts

To evaluate the mean forecast quality in the extended range, the skill scores have been averaged over all 38 cases (Fig.4.3). Since this ensemble is dominated by cases from the January/February period, the result is biased towards these months. Additionally the best and the worst scores for every forecast period have been extracted and displayed together with the mean skill in order to give an indication about the spread of the forecast quality. From the standard deviation of the forecasts scores a threshold tube has been calculated which encloses the skill which should be achieved by 95% of the forecasts (assuming a normal distribution). It coincides fairly well with the area of the minimum-maximum spread.

For a persistence forecast in the Northern Hemisphere, the mean Z500 anomaly correlation coefficient shows values between 20 and 30% (Fig. 4.3) and an RMS error of about 100 m (Fig. 4.4), both of which are quasi-independent of forecast time. On the contrary, the dependency of the model skill on forecast time is very strong at both resolutions: after a sharp drop during the first 10 (T21) and 15 (T42) days, the ACC slowly approaches the zero level. In terms of ACC, the T21 model is less skillful than persistence after days 6 to 15, and the T42 after days 11 to 20. With regard to the RMS error, during the first half of the forecast period the T42 model is more skillful than the T21 model and both models are better than persistence; during the second half, the situation is reversed. The larger RMS error of the T42 model compared to the T21, even in the presence of better ACC scores, can be explained by the higher variability of the T42 forecasts during the second part of the integration period.

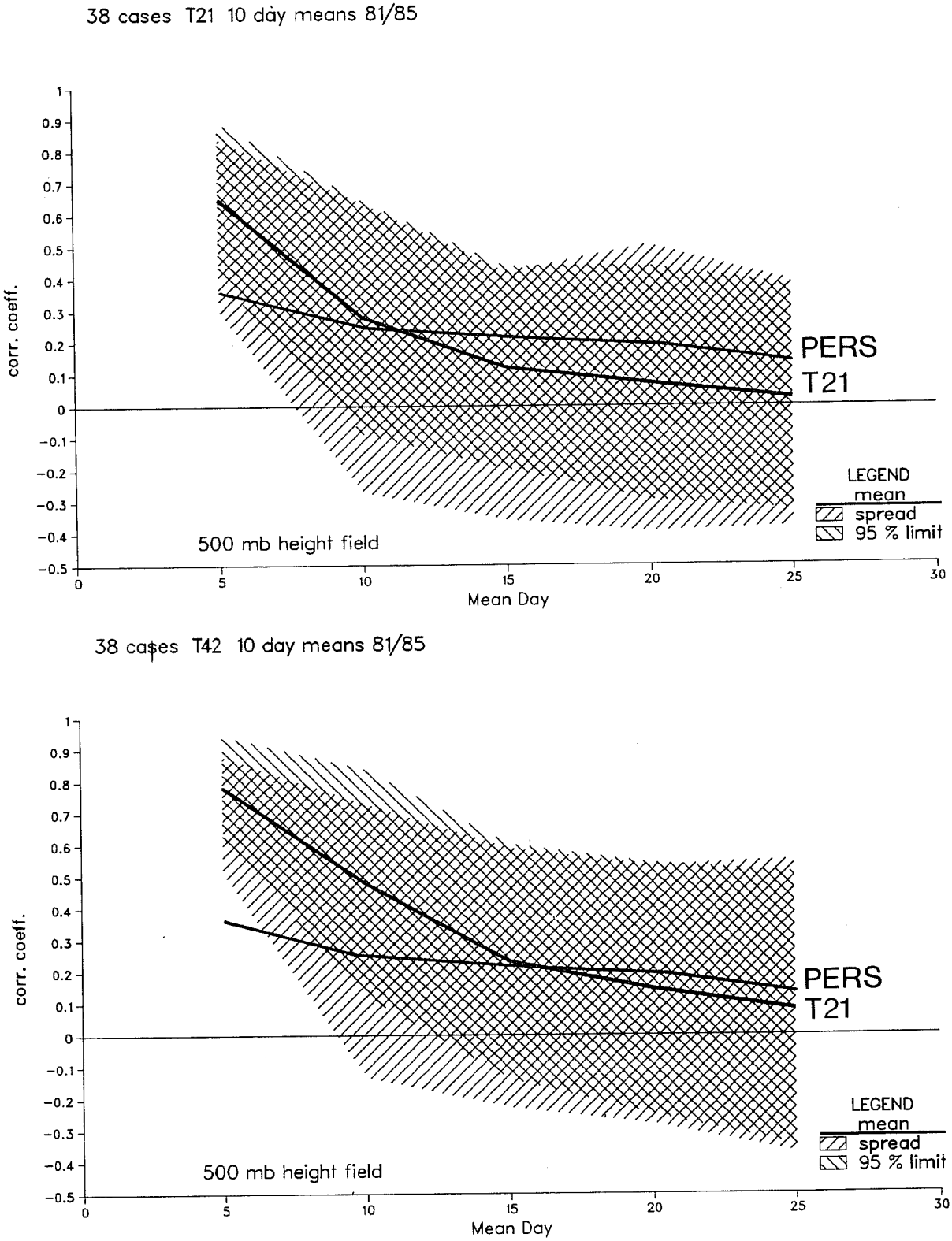


Fig. 4.3 The time evolution of the correlation coefficient and the RMS error for the Z500 averaged over all of the 38 cases; a) for the T21 model forecast; b) for the T42 model forecast. Persistence forecast is also shown in both panels. ///hatching indicates the max-min spread amongst the 38 cases; \\\ hatching indicates the skill variability corresponding to ± 2 standard deviations.

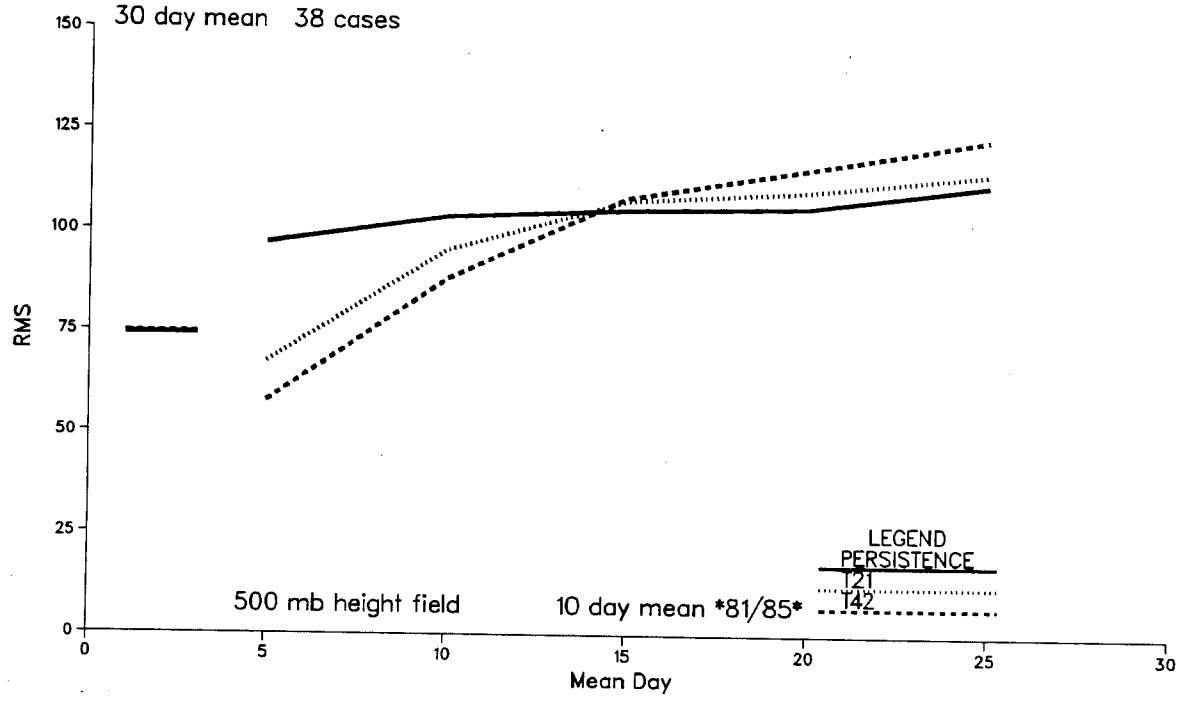


Fig. 4.4 The time evolution of the 500 mb height RMS error averaged over 38 cases: T21, T42 and persistence.

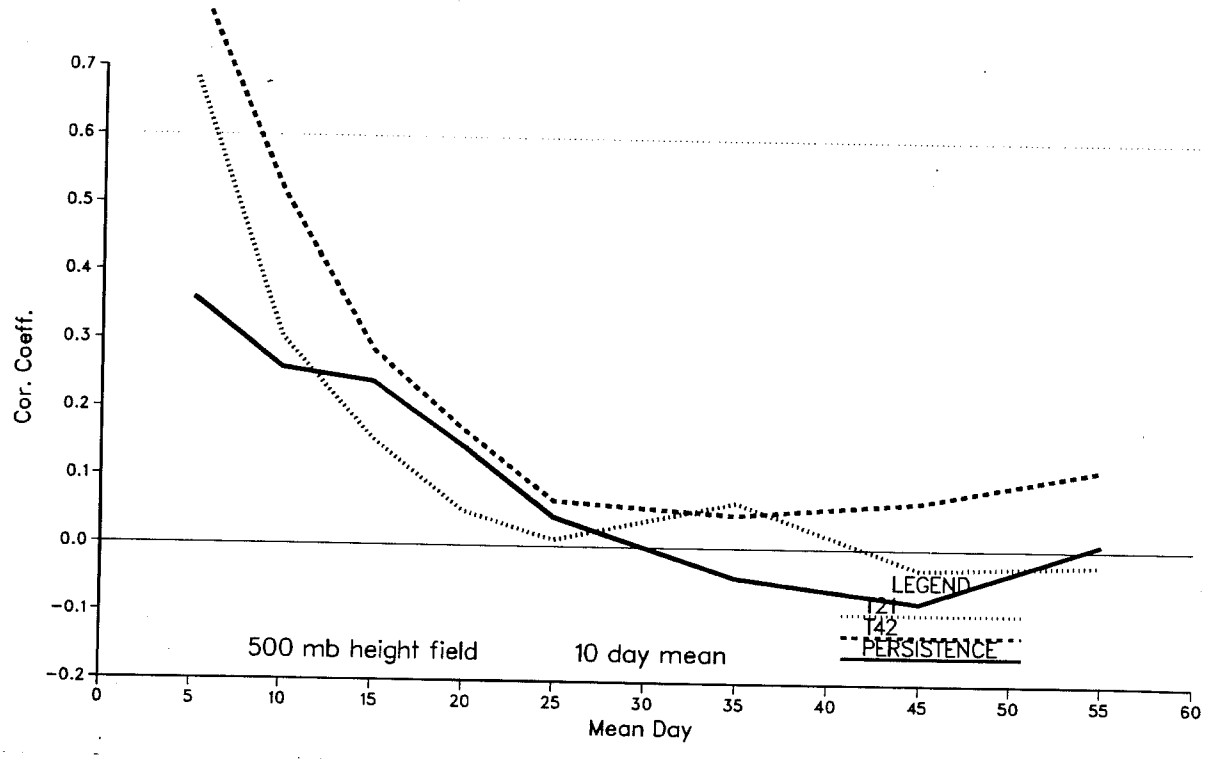


Fig. 4.5 The time evolution of the 10 day mean correlation coefficient for Z500 in a sample of ten 60-day forecasts.

The decay of model skill with forecast time is accompanied by an increase of the spread of the forecast skill (see again Fig. 4.3). The slower drop of the ACC for the T42 model during the first part of the forecast period has therefore its counterpart in a more contained spread of model skill, particularly at days 1 to 10. This is a further proof of the superior performance of the T42 model in the medium-range.

4.2.4 60-day forecasts

Ten experiments have been integrated up to 60 days. These experiments were used to investigate whether there is a recovery of skill during the later part of the forecast, a phenomenon discussed earlier.

The mean behaviour of the 10 cases considered (Fig.4.5) is similar to the larger ensemble discussed before with a sharp drop in correlation during the first time intervals and a levelling afterwards. The 10-day mean scores for the T42 model stay around the 10% level up to the end of the forecast period. It is interesting to note that, in this limited sample, the performance of the T42 model is always better than persistence, while the skill of the T21 model is about the same as persistence. The recovery in forecasting skill could not be found in the average over the 10 cases; this may indicate that it only occurs in selected cases, for example in cases where the tropical sea surface temperature plays a dominant role or during transitions between flow regimes.

The dynamical forecasts of 30-day means performed with the numerical models are always better than the persistence forecasts, and the skill of the model forecasts improves with resolution.

4.3 Summary

Based on the analysis of the objective skill of a set of 38 extended range forecasts starting from independent winter conditions separated by 10 days, the following conclusions can be drawn.

- The T42 model provides a more reliable forecast of 30 day means than persistence. However, the quality of these model forecast seems to be uncorrelated with the the quality of the persistence forecast. This is at variance with the results, obtained on a much smaller sample, described in Section 3.

- There exist a substantial variability in the predictability during different years.

- Up to day 10 the T21 model forecast is better than persistence, whilst the T42 model forecast stays superior to persistence up to day 15. The higher resolution model outperforms the T21 model in most respects and especially during the earlier part of the forecast period.

- A return of skill in the later part of the forecast period (studied with a subset of the experiments integrated up to 60 days), which had been observed in other forecasting experiments, could not be confirmed as a consistent feature. In these experiments the skill remains low after the drop in the earlier part of the forecast and appears not to recover.

The results indicate that, even using low resolution models, some skill is present on average well beyond the current operational forecast period (10 days at ECMWF).

5. CONCLUSIONS

Since summaries of the results of both LAF and deterministic experiments have been given in Sections 3.5 and 4.3, this section will be confined to recalling those points that are either most relevant to future research work or that have possible operational consequences.

Our results have shown that both purely deterministic and Lagged-Average numerical forecasts of 10-day and 30-day means of 500 mb height and 850 mb temperature outperform both persistence and climate forecasts well beyond the currently accepted practical limit for deterministic instantaneous forecasts, estimated around one week to ten days. This happens despite the comparatively low resolution of the models employed in our tests.

It has also been shown how the single most important model characteristic that influences forecast skill is, by far, horizontal resolution, at least in the resolution band examined. Stochastic-dynamic ensemble averaging techniques, although generally beneficial, do not show nearly as much impact as model resolution. However, a word of caution is necessary: it would not be legitimate to export our conclusions on the impact of model resolution to the upper limits of the resolution scale. Further numerical experimentation is needed, with higher resolution models, to estimate if and at what stage resolution ceases to be the most sensitive model characteristic and whether ensemble averaging might assume a much more important role in increasing model skill.

It has been shown that for the LAF technique the use of the spread of forecasts within an ensemble to estimate a priori forecast skill is viable and gives very promising results. Again this conclusion should be tested with higher resolution models and on a much larger statistical basis, encompassing seasons other than winter alone. Another area where more effort is needed.

Systematic error correction techniques have given somewhat inconsistent results. Although they clearly improve the RMS errors of the model, they have little impact on ACC at the highest model resolution tested, for which model internal variability seems to be large enough to make the truly systematic part of the error difficult to estimate from a small sample of independent

integrations. Such error correction techniques appear to be more applicable, as theory dictates, to ensemble of forecasts than to single deterministic integrations. When applied to the LAFs, they show the additional advantage of improving the spread-skill correlations.

The 'return' of forecast skill found during the latter part of extended range forecasts appears to be a recurrent, but not altogether consistent, feature of such experiments. More work is needed to understand its origin.

The ability of the models used in this set of experiments to represent and forecast the low-frequency variability of the Northern Hemisphere extratropics is itself very variable both on the inter-annual timescale and from case to case. Surface boundary forcings seem to be reasonably well modelled and are very likely to participate in increasing predictability and forecastability, while examples of remarkable model failures in reproducing internal dynamical variability on the 5 to 15 days time scale, e.g. blocking, are common. The understanding of the reasons for such failures is one of the most important topics for furure research in dynamical extended-range forecasting.

Acknowledgements

Large numerical experimentation programmes such as the one presented here are always a collective effort. A large number of individuals, both in the Research and Operations Departments of ECMWF, have in various ways contributed to this project by providing software and assistance, by participating in discussions and exchanging ideas or simply by bearing with us while we were making a nuisance of ourselves in mopping up computer resources. We thank them all. We would like to mention, in particular, R. Riddaway and K. Arpe for carefully reading an earlier version of the manuscript and making many useful suggestions to improve the readability of the text. One of us (FM) participated to this project partly supported by research funds of ENEL, the Italian National Electricity Board, and partly on leave from Istituto di Cosmogeofisica, Italian National Research Council, Turin, Italy.

References

Arpe, K., 1983: Diagnostic evaluation of analysis and forecasts: Climate of the model. ECMWF Seminar/Workshop on Interpretation of Numerical Weather Prediction Products, 13-24 September 1982, ECMWF, Reading, U.K., 99-140.

Arpe, K. and E. Klinker, 1986: Systematic errors of the ECMWF operational forecasting model in mid-latitudes. Quart.J.R.Met.Soc., 112, 181-202.

Arpe, K., A. Hollingsworth, M.S. Tracton, A.C. Lorenc, S. Uppala and P. Kållberg, 1985: The response of numerical weather prediction systems to FGGE level IIB data. Part II. Forecast verification in implications for predictability. Quart.J.Roy.Met.Soc., 111, 67-102.

Benzi, R., P. Malguzzi, A. Speranza and A. Sutera, 1986: The statistical properties of general atmospheric circulation: observational evidence and a minimal theory of bimodality. Quart.J. Roy.Met.Soc., 112, 661-674.

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.

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

Gilchrist, A., 1977: An experiment in extended range prediction using a general circulation model and including the influence of sea surface temperature anomalies. Beitr.Phys.Atmos., 50, 25-40.

Gleeson, T.A., 1968: A modern physical basis for meteorological prediction. Proceedings First National Conference Statistical Meteorology, Boston, Mass., USA., Amer. Meteor. Soc., 1-10.

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

Hansen, A.R. and A. Sutera, 1986: On the probability density distribution of large-scale atmospheric wave amplitude. Submitted to JAS.

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

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

Horel, J.D., and J.M. Wallace, 1981: Planetary-scale atmospheric phenomena associated with the Southern Oscillation. Mon.Wea.Rev., 109, 813-829.

Lorenz, E.N., 1963: Deterministic nonperiodic flow. J.Atmos.Sci., 20, 130-141.

Lorenz, E.N., 1969a: The predictability of a flow which possesses many scales of motion. Tellus, 21, 289-307.

Lorenz, E.N., 1969b: Atmospheric predictability as revealed by naturally occurring analogues. J.Atmos.Sci., 26, 636-646.

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

Louis, J.F. (editor), 1984: The ECMWF forecasting model. ECMWF Research Manual Vol.2 and Vol.3, ECMWF, Shinfield Park, Reading, UK.

Madden, R.A., 1976: Estimations of the natural variability of time-averaged sea-level pressure, Mon.Wea.Rev., 104, 942-952.

Miyakoda, K. and O. Talagrand, 1971: The assimilation of past data in dynamical analysis. I. Tellus, 23, 310-317.

Miyakoda, K., and J.-P. Chao, 1982: Essay on dynamical long-range forecasts of atmospheric circulation. J.Met.Soc. Japan, 60, 292-308.

Miyakoda, K., T. Gordon, R. Caverly, W. Stern, J. Sirutis, W. Bourke, 1983: Simulation of a blocking even in January 1977. Mon.Wea.Rev., 111, 846-869.

Miyakoda, K., J. Sirutis and J. Ploshay, 1986: One-month forecast experiments - without anomaly boundary forcings. Submitted to Mon.Wea.Rev.

Namias, J., 1969: Seasonal interactions between the North Pacific Ocean and the atmosphere during the 1960's. Mon.Wea.Rev., 97, 173-192.

Nicholls, N., 1980: Long-range weather forecasting: Value, status and prospects. Rev.Geophys.Space Phys., 18, 771-788.

Nicholls, N., G. Gruza, Y. Kikuchi and R. Sommerville, 1984: Long-Range Weather Forecasting: Recent Research. LRFR Publication Series No.3, WMO, Geneva.

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.Roy.Met.Soc., 112, 1001-1039.

Pitcher, E.J., 1977: Application of stochastic dynamic prediction to real data. J.Atmos.Sci., 34, 3-21.

Rasmusson, E.M., 1983: Ocean effects. Proceedings WMO-CAS/JSC Expert Study Meeting on Long-Range Forecasting, Princeton, 1-4 December 1982, Long-range Forec. Res. Publ. Ser. No. 1, WMO, Geneva, 97-122.

Ratcliffe, R.A.S., and R. Murray, 1970: New lag associations between North Atlantic sea temperature and European pressure applied to long-range forecasting. Quart.J.Roy.Met.Soc., 96, 226-246.

Reynolds, R.W., 1983: The sea surface temperatures during the 1982-83 El Nino Event. Proceedings of the Eighth Annual Climate Diagnostics Workshop, NOAA, US., 86-91.

Rinne, J., and V. Karhila, 1974: Stochastic forecasts computed with an EOF model. The GARP Programme on Numerical Experimentation, Report of the International Symposium on Spectral Methods in NWP, Copenhagen, 12-16 August 1974. (Report No.7 of the GARP-WGNE). WMO, Geneva, 333-340.

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

Shukla, J., 1981: Dynamical predictability of monthly means. J.Atmos.Sci., 38, 2547-2572.

Shukla, J., 1984: Predictability of time averages. Problems and Prospects in Long and Medium Range Weather Forecasting, D.M. Burridge and E. Källén, eds. Springer-Verlag, Berlin and New York., 109-206.

Shukla, J. and J.M. Wallace, 1983: Numerical simulation of the atmospheric response to equatorial Pacific sea surface temperature anomalies. J.Atmos.Sci., 40, 1613-1630.

Smagorinsky, J., 1969: Problems and promises of deterministic extended range forecasting. Bull.Am.Met.Soc., 50, 286-311.

Sutera, A., 1986: Probability density distribution of large scale atmospheric flow. Advances in Geophysics, Vol. 29, ed. by B. Saltzman. Ac. Press, Orlando, 227-250.

Tibaldi, S., 1986: Envelope orography and maintenance of the quasi-stationary circulation in the ECMWF global models. Advances in Geophysics, Vol. 29, ed. by B. Saltzman. Ac. Press, Orlando, 339-374.

Von Storch, H., E. Roeckner and U. Cubasch, 1985: Intercomparison of extended-range January simulations with general circulation models: statistical assessment of ensemble properties. Beitr.Phys.Atmos., 58, 477-497.

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

Wallace, J.M., S. Tibaldi and A.J. Simmons, 1983: Reduction of systematic forecast errors in the ECMWF model through the introduction of an envelope orography. Quart.J.Roy.Met.Soc., 109, 683-717.

Walsh, J.E., 1983: Sea ice, snow cover and soil moisture. Proceedings of the WMO-CAS/JSC Expert Study Meeting on Long-Range Forecasting, Princeton, 1-4 December 1982, Long-Range Forec. Res. PUbl. Series No. 1, WMO, Geneva, 84-96.

Walsh, J.E., and M.B. Richman, 1981: Seasonality in the associations between surface temperatures over the United States and the North Pacific Ocean. Mon.Wea.Rev., 109, 767-783.