30 DAY DYNAMICAL FORECASTS AT THE U.K. METEOROLOGICAL OFFICE

S.F. Milton and D.S. Richardson

Hadley Centre for Climate Prediction and research
Meteorological Office
Bracknell, U.K.

1. INTRODUCTION

In the last decade interest in producing forecasts beyond the medium-range (6-10 days) using current numerical weather prediction (NWP) and general circulation models (GCM) (referred to collectively as dynamical models) has increased as the skill of dynamical models used in the short (1-5 days) and medium-range has steadily improved. The current experience of forecasting in the extended range (days 10-30) using dynamical models has been one of modest levels of average skill but with large case-to-case variability (Tracton *et al.* 1989; Murphy 1990; Brankovic *et al.* 1990). A key issue in this area is the provision of an *a priori* prediction of forecast skill to allow the practical use of the proportion of dynamical forecasts which exhibit useable skill beyond 10 days.

One of the main methodologies for interpreting dynamical forecasts beyond the medium range has been the ensemble forecast (Leith 1974; Hoffman and Kalnay 1983), where several model integrations are run from slightly different initial states, the differences or perturbations ideally reflecting the uncertainties in an initial analysis.The ensemble represents a probability distribution of possible outcomes. Ideally, in a large, well sampled, and normally distributed ensemble, the centroid or ensemble-mean is the *best estimate* forecast. The variance or spread of this probability distribution increases with time due to the instability of the governing equations of motion to small perturbations. This growth of spread reflects the *internal* or random error growth (Dalcher and Kalnay 1987) and should ideally contain information on the uncertainties associated with the ensemble-mean forecast, enabling us to discriminate between skilful and unskilful forecasts *a priori*. Alternatively one can use the probabilistic information within an ensemble more explicitly to provide alternatives to the ensemble-mean forecast, each with their own probability of occurrence (Déqué 1988;Brankovic *et al.* 1990; Murphy 1990).

The U.K. Meteorological Office (UKMO) has been actively involved in research into ensemble forecasting in the extended range. The motivation has been to incorporate dynamical models into the UKMO operational long-range forecast (LRF), which since 1963 has provided routine predictions of temperature and rainfall for 10 U.K. districts on a twice monthly basis (Folland and Woodcock 1986).

Earlier work with dynamical models demonstrated the potential of the ensemble technique in the perfect model enviroment (Murphy 1988) and even showed that individual dynamical forecasts could outperform existing statistical techniques used in the LRF at least out to 15 days (Murphy and Dickinson 1989). A more recent study by Murphy (1990) investigated the practical impact of ensembles using the lagged average forecast (LAF) technique of Hoffman and Kalnay (1983), and showed that the potential gain in skill possible through ensemble averaging is not realised in practise due to the growth of model systematic errors, often called the *external* component of error growth. However some skill remained out to days 11-20 on average, and evidence of an ability to dicriminate between locally skilful and unskilful cases at days 6-15 using the ensemble spread was found, although the relationship was weak. Similar conclusions were reached by Brankovic *et al.* (1990) using LAF ensembles run with the T63 version of the ECMWF spectral model and by the NMC using their Dynamic Extended Range Forecast (DERF) experiments (Tracton *et al.* 1989).

Since December 1988 real-time dynamical model forecasts have been run in LAF ensemble mode at UKMO to coincide with the twice monthly LRF and have formed the basis for prediction on most occasions out to day 15. These ensemble forecasts, together with additional hindcast ensembles, provide a database of 112 LAF ensembles or 1008 individual 30-day forecasts covering 1985-91. This large database is used to address the question of the impact of the ensemble technique on skill in the extended range (section 3) and the ability of the technique to provide an *a priori* prediction of forecast skill using the ensemble spread (section 4). The size of the database allows us to test the conclusions of earlier studies performed with smaller samples of LAF ensembles and to investigate the seasonal and regional dependence of the results. In addition the question of relationships between circulation regimes and skill are also discussed in section 4 for the winter ensembles. Conclusions and plans for future work are presented in section 5.

2. MODEL AND DATABASE

At the UKMO the LAF technique was chosen to generate the ensembles as it is easy to implement in an operational enviroment. Each ensemble consists of nine model integrations of at least 32 days in length run from consecutive UKMO operational analyses separated by six hours. This is identical to the configuration used by Brankovic *et al.* (1990) but differs from that of Murphy (1990) who used a seven member ensemble run from analyses separated by twelve hours. The dates for the latest member of each ensemble, hereafter referred to as the Operational Dynamical Forecast (ODF), are shown in Table 1. Also shown are the definition of the seasons used in the remainder of the paper. Several ensembles were initialised at similar times to the ECMWF LAF ensembles, and the dataset also covers winter 1986/87 used in the NMC DERF study as well as the periods used in the LAF studies of Yamada *et al.* (1991) at the Japanese Meteorological Agency (JMA) and Déqué (1991a) at the French Meteorological Service (CNRM/Meteo France).

	1985/86	1986/87	1987/88	1988/89	1989/90	1990/91
	25.11.85	24.11.86	23.11.87	21.11.88	20.11.89	19.11.90
	09.12.85	08.12.86	07.12.87	05.12.88	04.12.89	03.12.90
WINTER	23.12.85	22.12.86	21.12.87	19.12.88	18.12.89	17.12.90
	06.01.86	05.01.87	04.01.88	02.01.89	01.01.90	31.12.90
	20.01.86	19.01.87	18.01.88	16.01.89	15.01.90	14.01.91
	03.02.86	02.02.87	01.02.88	30.01.89	29.01.90	28.01.91
	17.02.86		15.02.88	13.02.89	12.02.90	11.02.91
			29.02.88	27.02.89	26.02.90	25.02.91
			14.03.88	13.03.89	12.03.90	11.03.91
SPRING			28.03.88	27.03.89	26.03.90	25.03.91
			11.04.88	10.04.89	09.04.90	08.04.91
			25.04.88	24.04.89	23.04.90	24.04.91
			09.05.88	08.05.89	07.05.90	07.05.91
			23.05.88	22.05.89	21.05.90	20.05.91
			06.06.88	05.06.89	04.06.90	03.06.91
SUMMER			20.06.88	19.06.89	18.06.90	
			04.07.88	03.07.89	02.07.90	
			18.07.88	17.07.89	16.07.90	
			01.08.88	31.07.89	30.08.90	
			15.08.88	14.08.89	13.08.90	
		31.08.87	29.08.88	28.08.89	27.08.90	
		14.09.87	12.09.88	11.09.89	10.09.90	
AUTUMN		28.09.87	26.09.88	25.09.89	24.09.90	
		12.10.87	10.10.88	09.10.89	08.10.90	
		26.10.87	24.10.88	23.10.89	22.10.90	
		09.11.87	07.11.88	06.11.89	05.11.90	

Table 1: Initial dates of the ODF for each of the 112 UKMO LAF ensembles 1985-91

All integrations were run using a version of the UKMO global operational NWP finite difference model (Bell and Dickinson 1987) with a reduced horizontal resolution of 2 degrees latitude by 2.8 degrees longitude and 15 sigma levels in the vertical. The model contains parametrizations of the major subgridscale physical processes including gravity wave drag, penetrative convection and a climatological radiation scheme which takes account of seasonal and diurnal cycles but with the effects of cloud and humidity inferred from zonally averaged values (Corby *et al.* 1977). The sea surface temperature (SST) fields remain fixed at the initial analysed value throughout the course of each integration. These integrations differ from those in the study of Murphy (1990) where the 11-layer GCM employed in climate studies was used. Since June 1991 both the NWP model and GCM have been replaced by a new Unified Model (UM) developed at UKMO for use in both weather forecasting and climate studies (Wilson *et al.* 1990). The UM has now replaced the NWP model in the routine production of LAF ensembles at UKMO.

The UKMO long-range predictions of temperature and rainfall for the 10 U.K. districts are made for days 1-5, 6-15 and 16-30 of the month ahead. They are derived from a base prediction of mean sea level pressure (MSLP) using objective specification equations. Additional predictors for the specification of temperatures include the local SST, and for days 1-5, 1000-500mb thickness fields. The prediction of MSLP at days 1-5 is provided by the operational short-range forecasts of UKMO and ECMWF. Beyond days 1-5 the LAF ensembles and two statistical techniques called Multi-Variate Analysis (MVA) forecasting (Maryon and Storey 1985) and Surface Pressure Eigenvector Regression (SPEVR - see Harrison et al. 1991 for details) are available for use. On some occasions subjective modification of the temperature and rainfall derived from the specification equations is carried out, but this is kept to a minimum where possible. The skill in predicting the MSLP and 500mb height (H500) fields using the LAF ensembles is considered in the remainder of the paper. Forecast ranges other than those used in the operational LRF are considered as are regions other than the U.K. and Europe.

3. SKILL OF THE LAF ENSEMBLES IN THE EXTENDED-RANGE.

3.1 Potential predictability and the model climate.

As an ensemble forecast proceeds in time the dispersion or spread amongst its constituent members inevitably increases. The point at which the ensemble distribution is no longer distinguishable from the model climate distribution represents complete loss of predictability as measured in the perfect model enviroment. A quantitive measure of this potential predictability for an ensemble of M members is given by Murphy (1988,1990) as

$$< F^* > = (M/(M-1)) < s_M/w_f{}^* > \qquad (1)$$

where $< F^* >$ is the ratio of the ensemble variance ($< s_M >$) to the variance of the model about its own

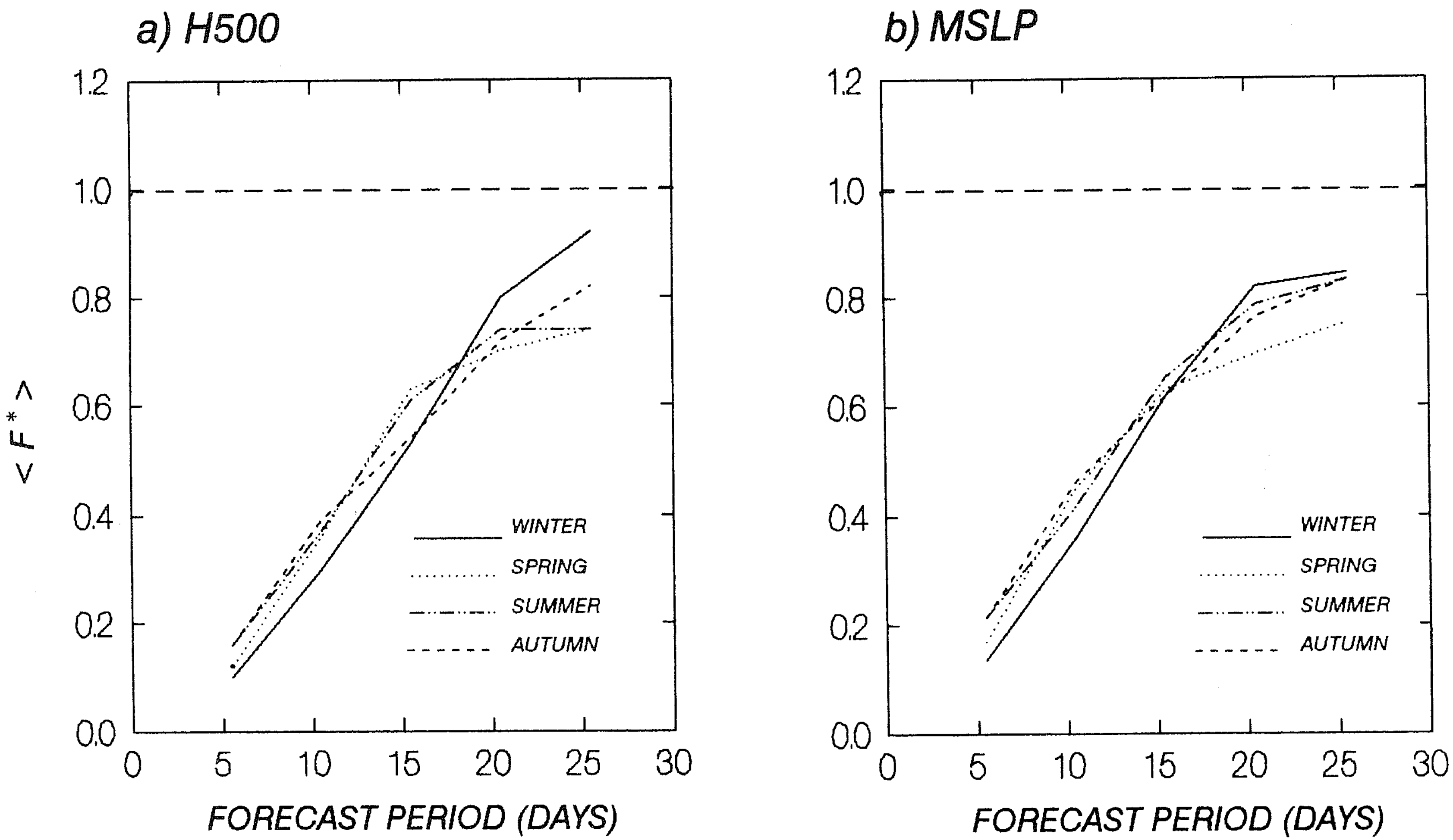

Figure 1. Growth of normalised ensemble variance <F*> for overlapping 10 day means over the Northern Hemisphere for a) 500mb heights and b) mean sea level pressure. Values plotted at the midpoint of the 10 day mean.

climate ($< w_f * >$). The operator $< >$ is an average over a large number of independent ensembles. When $< F^* >$ reaches unity all predictability is lost. The growth of $< F^* >$ for the Northern Hemisphere (NH) H500 and MSLP fields in all seasons shows a steady increase from values of 0.1 at days 1-10 to values around 0.8 at days 21-30 (Figure 1- calculated as described in Murphy (1990)). Some seasonal variation in potential predictability does occur, although differences are small. Spring is the most predictable and winter the least predictable at later forecast ranges. None of the curves have reached the predictability limit by days 21-30, suggesting that ensemble forecasts retain a potentially predictable signal out to this time range. However as pointed out by Murphy (1990) these estimates may be optimistic if the dispersion of the model solutions is unrealistically slow compared to that of the true solutions to the atmospheric equations of motion.

In deriving $< F^* >$ one can also calculate normalised values of the model's systematic error (SE) in variability and mean flow. The SE in variability ($<\beta^2>$) is given by the model's variance about its own climatology ($<w_f*>$) normalised by the observed climate variance ($<w_o>$). The measure of the SE in the mean flow ($<\alpha^2>$) is the model's mean square mean error (forecast minus analysis) again normalised by the observed climate variance. The model variability at H500 is consistently less than observed in the winter and summer seasons and generally decreases with forecast range, as seen by the decrease in $<\beta^2>$ (Figure 2). The levels of variability are higher for MSLP than for H500, and for spring actually exceed observed levels of variability at days 16-25 and 21-30. An initial decrease then recovery of model variability for MSLP is a signature of the transition seasons. The general loss of variability at H500 in all seasons and forecast ranges has serious implications for the practical application of a probabilistic technique such as ensemble forecasting, as it signifies the inability of the model to capture the full range of atmospheric flow regimes and transitions between them (Molteni and Tibaldi 1990).

The seasonal evolution of SE in the mean flow $<\alpha^2>$ is more complex. In winter we see a steady growth of SE until by days 21-30 it is 77%(60%) of the observed climate variance for H500(MSLP). In spring there is a rapid growth of $<\alpha^2>$ at H500 such that by days 21-30 the mean square mean error is greater than the observed climate variance. Similar behaviour can be seen for MSLP during summer, where the mean square mean error is already greater than the observed climate variance by days 11-20. Murphy (1990) found the mean square mean error to be around 60% of the climate variance at days 21-30, although his results were averaged across all seasons, and clearly large seasonal differences occur which may ultimately affect the practical predictability of the ensemble forecasts. The impact on skill of correcting *a posteriori* for model SE in the mean flow is discussed in section 3.5. Detailed descriptions of the model climate and SE can be found in Bourke *et al.* (1991).

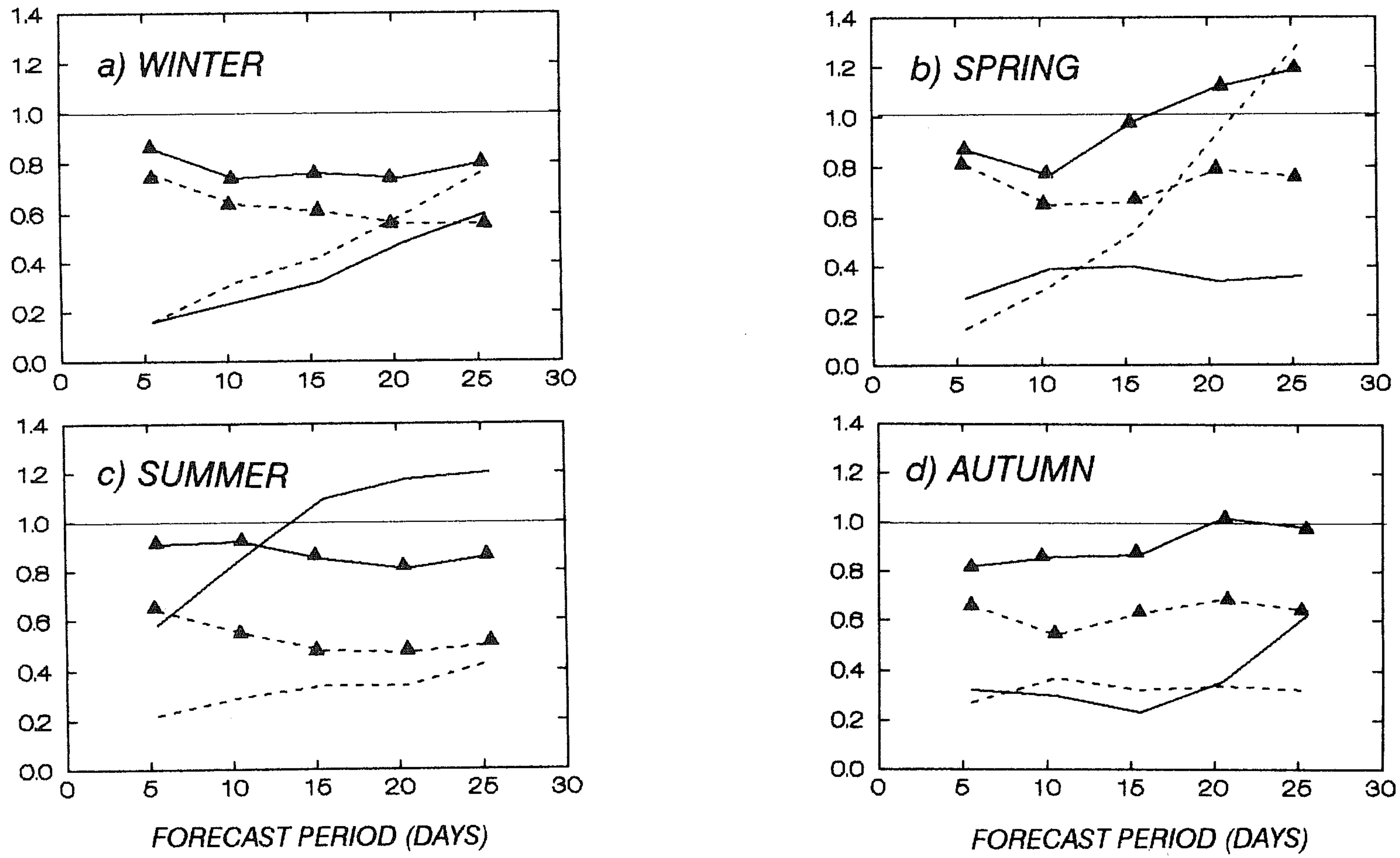

Figure 2. Seasonal estimates of the growth of normalised mean square mean error (α^2) of 500mb height (- - - -) and mean sea level pressure (——) for 10-day overlapping means over the Northern Hemisphere. Also shown are the normalised errors in variability (β^2) for 500mb height (▲- - -▲) and mean sea level pressure. (▲——▲)

3.2 Average skill on a seasonal basis.

In common with other studies we assess the practical levels of skill using the anomaly correlation coefficient (ACC) scores of the ensembles averaged over all cases in each season for H500 over the NH . All anomalies are formed with respect to a 1951-80 climatology as this minimises the climate sampling error introduced using shorter climatologies (Murphy and Dickinson 1989). The evolution of <ACC> for the ensemble-mean, persistence, and random forecasts are shown for overlapping 10 day means (1-10, 6-15, ..., 21-30) (Figure 3 - top row). Persistence is defined by persisting the anomaly averaged over the *n* days preceding the analysis time of the ODF, where *n* is the time-average period under question (i.e. 10 days in this case). The <ACC> of the random forecasts is an estimate of the noise level and is calculated by verifying each ensemble forecast against an analysis chosen at random from the same point in the annual cycle, but from an independent year of the sample.

All seasons show a rapid decay of <ACC> from values between 0.5 and 0.8 at days 1-10 to values between 0.0 and 0.2 at days 21-30. Winter is the most skilful at all time ranges, in contrast to the potential predictability estimates. In fact at days 11-20 and 16-25 the winter ensemble-mean forecasts are more skilful than those at days 6-15 in the remaining seasons. Winter and spring remain above the noise level in all time ranges suggesting some predictability, although such small <ACC> may not be practically useful. The ensemble-mean forecast performs best against persistence in winter and autumn, although in autumn the skill is negligible by days 11-20. Values of <ACC> are higher for MSLP than for H500 in all seasons (see Table 2).

An alternative skill measure to the phase based ACC is the normalised mean square error (E). Introduced by Murphy (1988) as a measure of the amplitude skill it is given by

$$E = (f - o)^2 / (< w_f > + < w_o >) \qquad (2)$$

where f and o are the forecast and observed anomalies and $< w_f >$ is the variance of the model forecast (ensemble mean or individual) about the *observed* climatology. The denominator is the mean square error (MSE) of random forecasts averaged over a large number of cases (i.e. noise level). This normalised MSE (E) has the advantage of expressing the skill in signal to noise terms such that when E=1 the forecast is no better than one made at random. In the context of ensemble forecasting E is preferable to conventional MSE as the latter score for the ensemble-mean forecast is reduced simply through smoothing introduced by ensemble averaging, which is not a property of the normalised score .

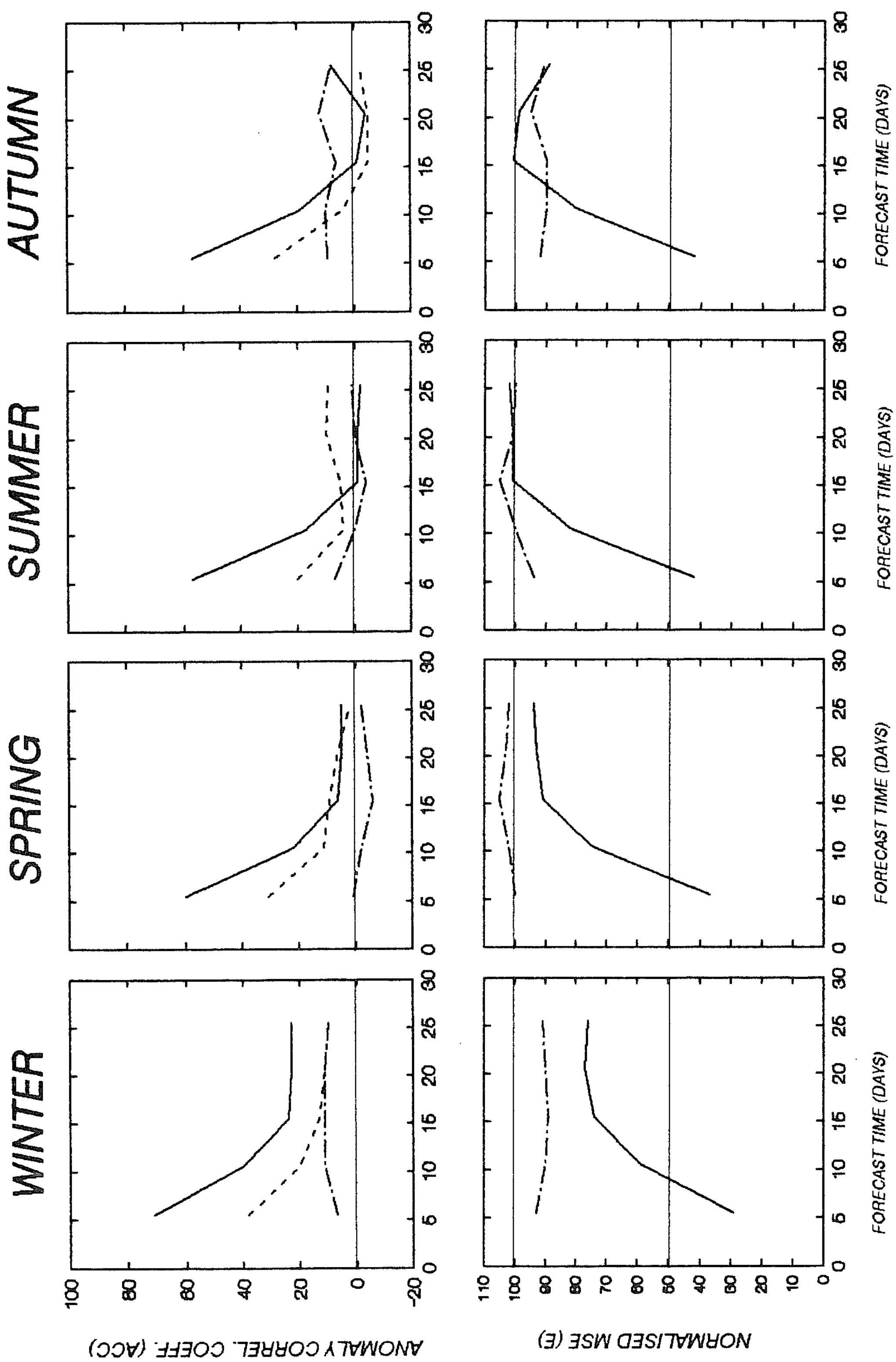

Figure 3. Average seasonal skill of 500mb heights over the northern hemisphere of the ensemble mean (——), persistence (- - - -), and random forecasts (-·-·-·) for overlapping 10-day means as measured by ACC x 100 (top row) and normalised mean square error x 100 (bottom row- persistence not included).

The growth of <E> for the ensemble-mean forecast (figure 3 - bottom row) gives very similar information to the decay of <ACC>, with a clear seasonal cycle in skill. The 0.5 (or 50%) level is the <E> value of forecasts of climatology. The ensemble-mean forecastss cross this limit within the first ten days, at around days 4-13 for 10 day means in spring, summer, and autumn at days 6-15 for the winter ensembles. Climatology provides a very stringent control forecast for MSE measures as the assymptote is 1/2 that for the MSE of a persistence forecast. A probabilistic framework has more desirable properties for assessing forecasts against climatology and this is discussed in section 3.4. The similarity between <ACC> and <E> is not surprising given that MSE based scores and ACC are related algebraically (Murphy and Epstein 1989). For an individual forecast it can be shown that

$$E = (w_f + w_o - 2(w_f.w_o)^{1/2}.ACC)/(\langle w_f\rangle + \langle w_o\rangle) \quad (3)$$

In the special case where $\langle w_f\rangle=\langle w_o\rangle$ and under certain assumptions concerning the definition of <ACC> then $\langle E\rangle=1-\langle ACC\rangle$. This relation is virtually satisfied in figure 3 although on a case-by- case basis this relation does not hold as it depends upon the variations in w_f and w_o.

These low values of average skill beyond days 1-10 over the NH are typical of those found in other extended-range forecast studies. For days 1-30 the <ACC> of the ensemble-mean forecast for H500 (MSLP) over the NH is 0.25 (0.39) averaged over all 112 ensembles. This is similar to the value of 0.27 for H500 found by Murphy (1990). During the winter season the <ACC> for H500 over the NH is 0.40 for the UKMO ensemble-mean forecasts. This compares with a value of 0.39 found by Tracton *et al.* (1989) for their DERF cases covering winter 1986/87. The <ACC> for the seven UKMO ensembles comprising the winter 1986/87 is 0.42. Yamada *et al.* (1990) found similar seasonal variations in skill with their LAF ensembles run using the JMA global prediction model. Their ensemble-mean forecast <ACC> in winter at days 1-30 for NH H500 was 0.54 for winters 1986/87 to 1989/90. The <ACC>value for the UKMO ensemble-mean forecasts over the same winters is 0.42, suggesting the JMA ensembles are slightly more skilful, although this is subject to sampling involved in running ensembles from different initial conditions.

3.3 Variability in ensemble-mean skill.

A pseudo time series of the ACC at H500 over the NH is shown in figure 4 for days 6-15, 16-30, and 1-30 of both the ensemble-mean and persistence forecasts. The first 13 ensembles are for winters 1985/86 and

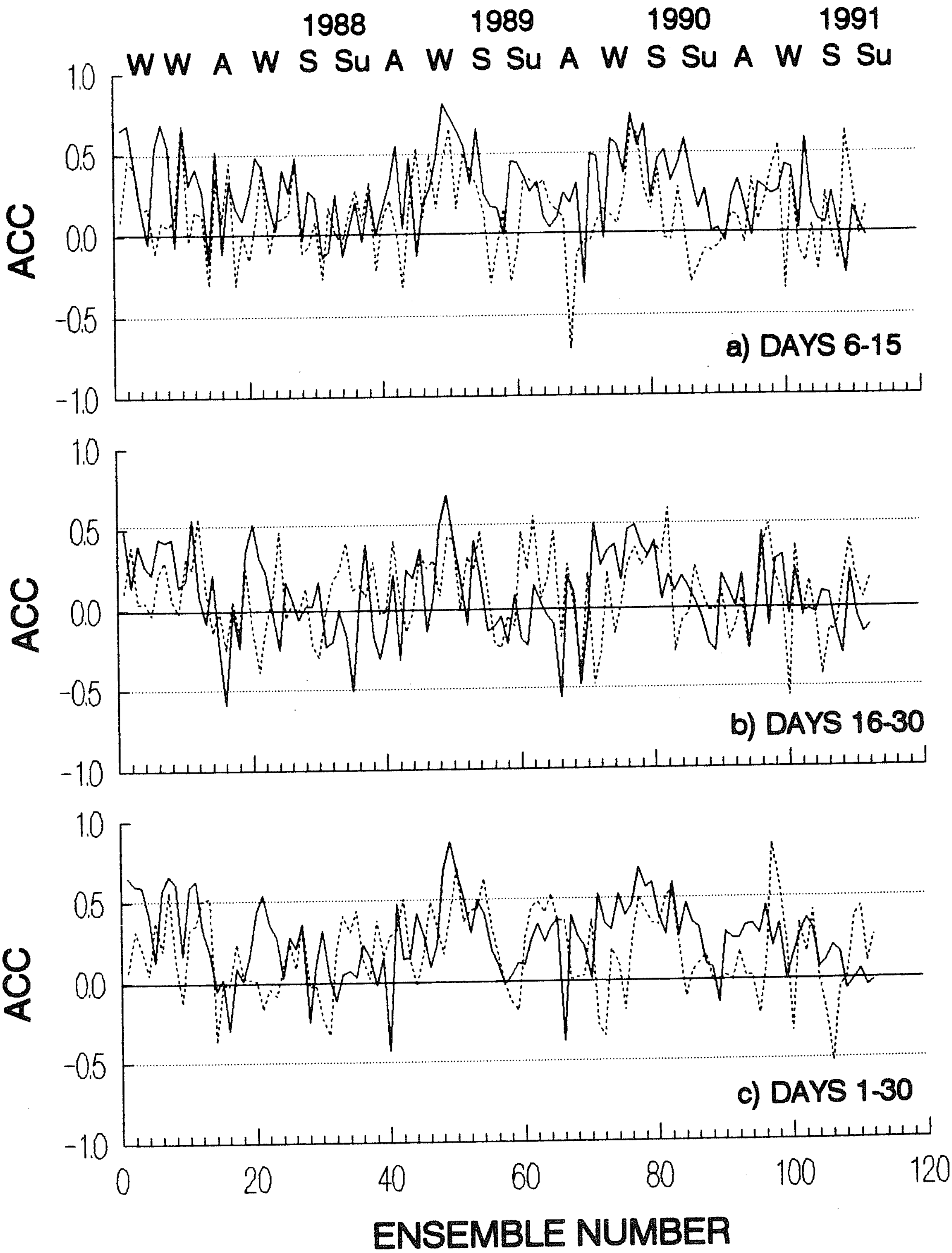

Figure 4. Time series across 112 LAF ensembles of ACC for 500mb heights over the Northern Hemisphere of the ensemble mean (——) and persistence (- - - -) at a) days 6-15, b) days 16-30, c) days 1-30. The dates of the ensembles in terms of seasons (W=winter etc.) are shown in the top panel .

1986/87 and thereafter there is a continuous time series of ensembles each separated by 14 days and beginning in autumn 1987. The ensemble-mean forecast ACC exhibits large fluctuations around the mean values averaged across all 112 ensembles of 0.27, 0.08, and 0.26 for days 6-15, 16-30, and 1-30 respectively, and is more skilful than persistence on 70%(52%) of occasions at days 6-15(16-30). There are few examples of highly skilful cases in the extended range with most ensembles having ACC between 0.0 and 0.5 at days 6-15 and 1-30. The ensembles with ACC > 0.5 all occur during winter with ensembles from 1988/89 and 1989/90 the most skilful, in agreement with the results of Yamada *et al.* (1990) and Déqué (1991a). A greater proportion of ensembles have negative ACC at days 16-30 compared to days 6-15. Of the 4 least skilful ensembles at days 16-30 (ACC ~ -0.5), three occur in autumn and one in summer. For smaller regions such as the North Atlantic sector this variability in skill is even more marked (Milton 1990), placing a large premium on being able to identify the most skilful cases in advance. The correlations between the time series of ensemble-mean and persistence ACC are 0.30, 0.20, and 0.20 for days 6-15, 16-30, and 1-30 respectively. These positive correlations suggest that the model is generally at its most skilful when persistence is high (e.g. winters 1988/89 and 1989/90). However, the correlations are not large, implying that the model does not just persist the large-scale flow on all occasions. Examples of individual cases when the observed persistence is low and model skill is high can be seen even at days 16-30 (e.g.winter 1989/90), and an example of a succesful model prediction of a regime transition at days 11-20 is discussed in section 3.8.

3.4 Improvement in skill due to ensemble averaging.

Having examined the skill of the ensemble forecasts we now consider to what extent the ensemble-mean forecast provides the optimum forecasting tool at extended range. Using ACC as a measure of skill it can be shown that under assumptions the ensemble-mean forecast ACC is an improvement over the average ACC of the individual ensemble members provided the latter score is positive (Brankovic *et al.* 1990). For the UKMO ensembles the ensemble-mean ACC is higher than that of the average ACC of the individual members in all seasons (Table 2). The largest improvement is for MSLP during winter at days 21-30 where the ensemble-mean <ACC> is 0.34 and the average <ACC> of the individual members is 0.26, representing a 30% improvement in skill. The overall improvement due to ensemble averaging is less than the potential improvements demonstrated by Murphy (1988, 1990) and reflects the growth of SE in the model noted in section 3.1.

A drawback of the LAF technique is that the ensemble members are not *a priori* equally likely in the early stages of the ensemble forecast due to the members run from earlier analysis times having effective

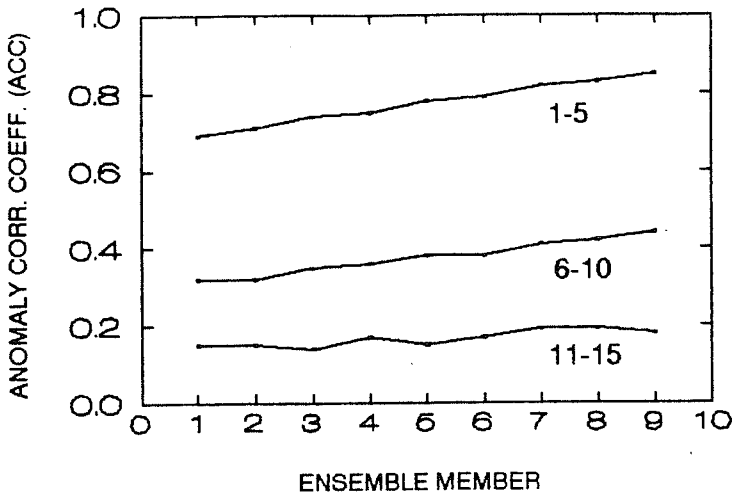

Figure 5 - Variation with ensemble member of H500 <ACC> over the northern hemisphere for days 1-5, 6-10, and 11-15. Members 1 to 9 are run from successive analyses times. (1 = -48hrs, 2 = -42.....9 = 00 hrs relative to day 0)

	Forecast range	AVG	EM	ODF
Winter	1-10	68 (66)	71 (69)	76 (73)
	6-15	37 (38)	40 (43)	43 (43)
	11-20	20 (25)	<u>24</u> (<u>31</u>)	22 (27)
	16-25	18 (23)	<u>23</u> (<u>30</u>)	22 (28)
	21-30	18 (26)	<u>23</u> (<u>34</u>)	22 (30)
	1-30	36 (45)	40 (51)	42 (51)
Spring	1-10	57 (59)	60 (62)	68 (67)
	6-15	21 (26)	22 (28)	27 (34)
	11-20	06 (13)	06 (16)	09 (20)
	16-25	04 (09)	<u>05</u> (<u>13</u>)	04 (12)
	21-30	04 (11)	05 (10)	06 (11)
	1-30	18 (29)	20 (33)	22 (37)
Summer	1-10	54 (51)	57 (54)	63 (59)
	6-15	15 (22)	17 (24)	22 (26)
	11-20	01 (14)	-01 (16)	04 (17)
	16-25	-02 (12)	-01 (15)	03 (15)
	21-30	-02 (11)	-02 (14)	04 (15)
	1-30	13 (28)	15 (31)	19 (33)
Autumn	1-10	53 (60)	57 (64)	63 (67)
	6-15	17 (26)	19 (<u>31</u>)	21 (29)
	11-20	00 (09)	-<u>01</u> (<u>12</u>)	-02 (10)
	16-25	00 (09)	-04 (<u>13</u>)	04 (12)
	21-30	05 (13)	<u>08</u> (<u>18</u>)	07 (12)
	1-30	12 (28)	16 (33)	15 (32)

Table 2 - Mean seasonal ACC x 100 of H500 (MSLP) over the northern hemisphere for the ensemble-mean (EM), average of the individual ensemble members (AVG), and the operational dynamical forecast (ODF). EM is underlined where it is more skilful than both AVG and ODF.

perturbations at day 0 which are functions of the *external* component of error growth. A trend in <ACC> across the ensemble members can be seen for days 1-5 and 6-10 for H500 over the NH averaged over all 112 ensembles (figure 5). At days 11-15 there is little trend in <ACC> suggesting that the members have become equally probable.In this forecast range the ensemble-mean forecast becomes more skilful on average than the ODF (not shown).The practical utility of the ensemble-mean forecastin the first 10 days could probably be improved by applying some form of optimal weighting procedure (Dalcher *et al.* 1988). Closer inspection of figure 5 reveals that at days 1-5 and 6-10 the difference in <ACC> is largest between ensemble members run from main synoptic hours (00 UTC and 12 UTC), whereas the <ACC> of an ensemble member run from a main synoptic hour and the <ACC> an ensemble member run from the intermediate analysis six hours later (06 UTC and 18 UTC) are very similar. This probably reflects the fact that the assimilation scheme uses the analysis from six hours earlier as its background field and hence in data sparse regions there will be no perturbation between two analyses separated by six hours. The effect is accentuated at intermediate analysis times as the quantity of synoptic observations is less than at the main synoptic hours. The failure to provide perturbations between successive analyses in data sparse areas may lead to unrealistically slow growth rates of spread for the ensemble in the early stages of the forecast.

Given the trend in skill across the ensemble, a more pertinent control forecast for the ensemble mean is the ODF. The ensemble-mean forecast performs best in winter , where it is a modest improvement over the ODF at days 11-20, 16-25 and 21-30 for both H500 and MSLP (Table 2). In spring a marginal improvement occurs at days 16-25 when skill itself is positive but small, and in summer there is no improvement over the ODF, although the ensemble-mean forecast is no less skilful on average than the ODF. By autumn the ensemble-mean forecast is again outperforming the ODF, but only at days 6-15 is the skill significantly greater than that expected from choosing random forecasts (see figure 3). The improvements due to ensemble averaging during winter are more optimistic than those noted by Murphy (1990) for all seasons. For example, at days 11-20 the winter ensemble-mean forecast and ODF <ACC> are 0.31 and 0.27 respectively compared to 0.13 for both ensemble-mean forecast and ODF in Murphy's results using eight ensembles. Contrary to our results, Brankovic *et al.* (1990) found the smallest improvement due to ensemble averaging in the winter season, which is probably a consequence of the small sample of ensembles used in their study. These seasonal variations highlight the need to assess the temporal and spatial variations in skill in order to maximise the potential benefit of the ensemble and other extended-range forecasting techniques (Livezey 1990).

On a case-by-case basis the LAF ensembles show considerable variability in the performance of the ensemble-mean forecast against the ODF. Histograms of the distribution of ACC across the ensembles at days 11-20 show the ensemble-mean forecast to be more skilful than the ODF in 24 of the 41 cases (58%), 22 of which have positive ACC scores (figure 6). In the individual cases where skill is highest the ensemble-mean forecast is often a substantial improvement over the ODF (17.02.86, 09.12.85, 15.01.90, 12.02.90, 31.12.90, and 11.02.91). This is a consequence of the earlier members of the ensemble providing positive and skilful information and would appear justification for running the nine member configuration chosen, although ideally the optimal size of the ensemble should be determined objectively. In only a few of the cases is the ODF a substantial improvement over the ensemble-mean forecast (e.g. 21.12.87, 05.12.88, and 03.12.90). These cases tend to be those with low average ACC of the ensemble.

Other relevant controls include the zero-cost forecasts of persistence and climatology. Following Brankovic *et al.* (1990) we have counted the occasions where the ensemble-mean forecast is an improvement over these zero-cost forecasts and also over the ODF at various forecast ranges for MSLP over the NH. The ensemble mean is considered optimum with respect to the controls if it is more skilful than a given control in over 50% of cases. Against the ODF the ensemble mean becomes optimum at days 11-20 for winter and autumn , although even at days 6-15 the autumn and winter ensembles reach the 50% level. Spring and summer remain below the 50% level for all forecast ranges. These results reflect those given in Table 2 for the average skill and are similar for both ACC and E measures of skill. The ensemble mean is the optimum forecast in all seasons and at all time ranges when compared to persistence, performing best in winter and autumn, but against climatology only remains above the 50% cutoff level in the first 10 days in all seasons and for days 6-15 during winter (using E as the skill measure).

Several previous studies have used the point at which the model forecast MSE crosses the climatological MSE as a measure of practical utility (Tribbia and Baumhefner 1988). However, climatology is very difficult to beat in an R.M.S. sense, and such R.M.S skill measures may not be an accurate reflection of the utility of the ensemble forecasts. A more natural framework for assessing the ensemble forecasts is a probabilistic one. Murphy (1990) verified his MSLP ensemble forecasts by categorizing them into equiprobable terces of MSLP derived from a 1951-80 climatology and assessing them using the Ranked Probability Score (RPS - Murphy, A.H., 1971). In this framework the definition of a climatology forecast is defined as one of equal probabilities across the terces. The RPS score of the ensemble-mean forecast asymptotes to the RPS score of climatology as the forecasts progress, and therefore the latter score is also a measure of the practical predictability limit of the forecasts. This is not a feature of R.M.S. based scores.

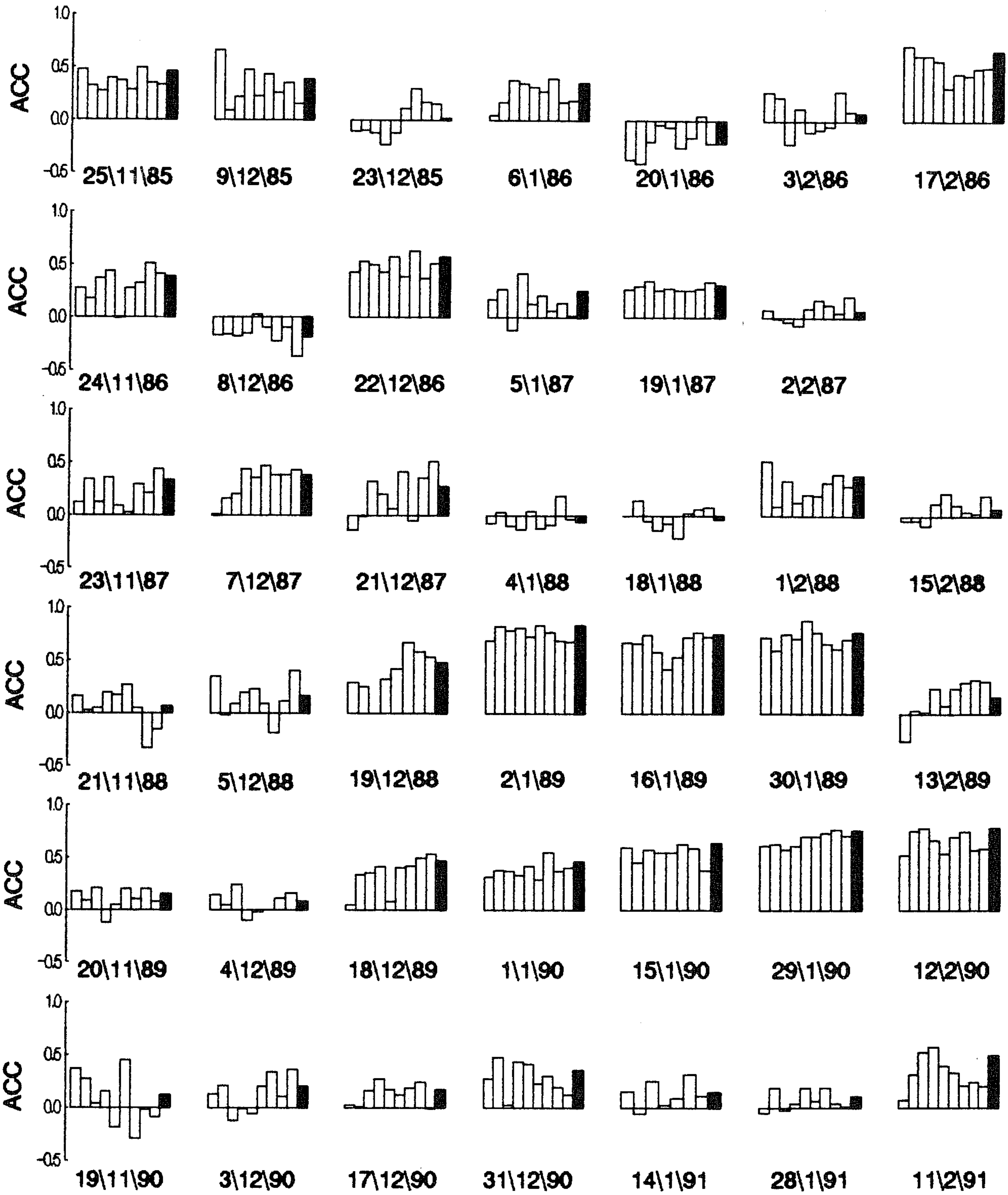

Figure 6. Histograms of ACC across the individual ensemble members for the 41 winter ensembles at days 11-20 for 500mb heights over the Northern Hemisphere. The solid bar is the the ensemble mean and individual members are plotted in order of increasing analysis time from left to right.

Murphy (1990) found that his ensemble-mean RPS score did not cross the climatology RPS score until days 16-25, which is very different to the results given by the R.M.S. measures where the ensemble-mean forecasts generally cross climatology at around days 6-15 (figure 3). Similar results have been found for this current set of UKMO LAF ensembles using the RPS score (D.S. Richardson, personal communication).

Summarising this section, we note that the ensemble-mean forecast is the optimum dynamical forecast beyond the first 10 days for winter and autumn and comparable to the ODF in spring and summer. The ensemble-mean forecast is consistently more skilful than persistence in all seasons for MSLP, but only an improvement over climatology in the first 10 days in an R.M.S. sense, although as noted above this may be a harsh measure of forecast utility.

3.5 Impact of independent systematic error correction

Given the growth of SE in the mean flow (section 3.1) another method of improving the skill of the dynamical forecasts may be to correct the ensemble-mean forecast *a posteriori* for model SE. Miyakoda *et al.* (1986) found a large improvement in <ACC> when SE calculated over a dependent set of extended-range integrations was used for correction, although Murphy (1990) showed that the use of independent integrations to form the SE led to smaller improvements, reflecting the sampling involved in estimating the SE (Déqué, 1991b). Seasonal estimates of the model SE have been calculated using the ODF integrations in the years independent of the ensemble forecast being corrected. For example 34 forecasts are used to estimate winter SE and 16 for summer. The model SE for the winter season H500 field at days 21-30 shows a tendency for lower than analysed geopotential heights over the extratropical NH (figure 7b), a feature common to other seasons and consistent with the model's cold bias in temperature (Bourke *et al.* 1991). The change in ensemble-mean forecast <ACC> due to SE correction is consistently positive in most seasons except for MSLP in autumn, with the largest improvements in spring and summer for H500 (Figure 8). At days 21-30 the correction has a negative impact in summer, autumn and winter for MSLP. The improvement is greater for days 1-30 as the SE is a larger proportion of the observed climate variance for 30 day means. Murphy (1990) found a mean increase in ACC at days 11-20 for H500 of 0.01 averaged across his eight ensembles, which is less than the value of 0.06 for the average across our sample of 112 ensembles, and 0.15 for the summer ensembles alone.

On a regional basis the impact of SE correction can be even larger. In the UKMO ensembles the biggest improvement at days 11-20 in the winter season is for the USA region (region E - figure 9) where the ensemble-mean forecast <ACC> is increased from 0.16 to a more respectable 0.40.

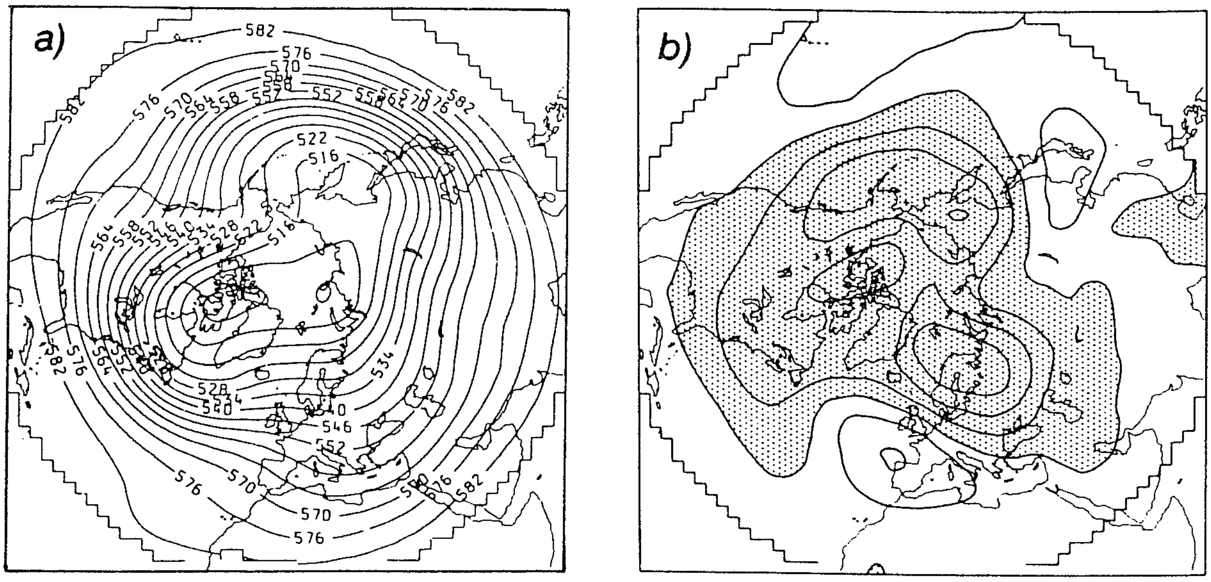

Figure 7. a) 500mb height verifying analysis corresponding to the 41 winter ODF at days 21-30. Contour interval is 6 dam b) 500mb height systematic error (forecast - analysis) 41 winter ODF at days 21-30. Contour interval is 3 dam. Values less than -3 dam are shaded.

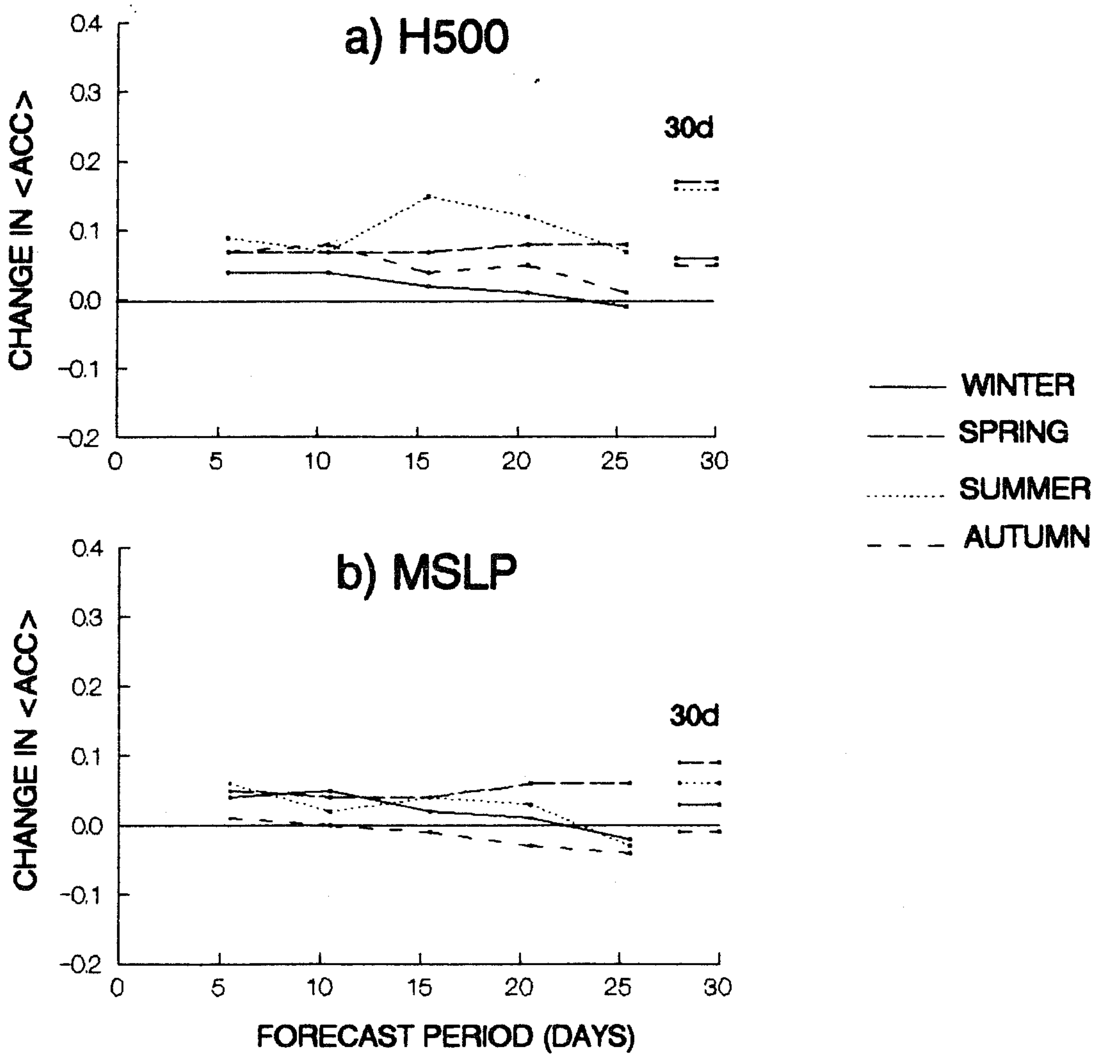

Figure 8. Average change in ensemble-mean <ACC> after removal of seasonal estimates of model systematic error for overlapping 10 day means of a) 500mb height and b) mean sea level pressure, over the Northern Hemisphere. the lines under 30d are the values for days 1-30 time averages.

3.6 Regional verification

The utility of dynamical model forecasts ultimately depends on applying them to regions of economic and societal interest. An example of the regional variability of skill defined over six regions (figure 9) is shown for the MSLP <ACC> scores of the ensemble-mean forecast, ODF, and persistence for the winter season (figure 10). Estimates of the winter SE have been removed from all the model forecasts. The <ACC> show large regional variations with highest skill over Eastern Asia (region C) and minimum over the U.K. and Europe (region A). This distribution in skill may reflect the distribution of areas of observed low frequency variability which have maxima in the blocking prone regions of the U.K. and Europe and the Eastern Pacific Ocean (region D) where regional skill is lowest . Similar regional variations are evident in the <ACC> of the persistence forecasts. The performance of the ensemble-mean forecast realtive to the ODF is greater in regions where <ACC> itself is higher. For example over Eastern Asia <ACC> for the ensemble mean is 0.57 and for the ODF is 0.40, whereas over the NH these values are 0.33 and 0.29 respectively. Therefore, on a regional basis we may expect to see greater benefit at days 11-20 from ensemble averaging than is suggested by hemispheric scores, although the current estimates of regional skill suggest that this is not the the case for the European region.

3.7 Skill of the UKMO temperature forecasts

Since the introduction of the LAF ensembles into the operational LRF they have formed the basis of prediction in the 6-15 day forecast range in the majority of cases. Milton (1990) showed that in most seasons the ensemble-mean forecast is a significant improvement over the current statistical techniques used in the LRF in providing a MSLP prediction for the UK region at days 6-15. This was not the case for days 16-30. It is worth considering whether the skill in MSLP at days 6-15 is communicated to the temperature forecasts via the objective specification equations. Given the use of the ensemble-mean forecast in the majority of cases we might expect to see the seasonal variations in temperature (and rainfall) skill to match those in ensemble-mean forecast skill for MSLP over a region surrounding the U.K. (this does not allow for additional predictability which could arise from the use of the SST field in the specification equations). The skill of the MSLP field is given by the ACC over a U.K. 30 point grid (30°W-20°E, 45-60°N) used in the LRF to determine the MSLP values for use in the specification equations. Temperature skill for the 10 U.K. districts is given by the Folland-Painting (FP) skill score (Folland *et al.* 1986), which classifies forecast and observed temperatures into equiprobale quints derived from a 1951-80 climatology and scores according to how close forecast and observed quints are to one another. A score of 100% is a perfect forecast and 0% is the score of a random forecast.

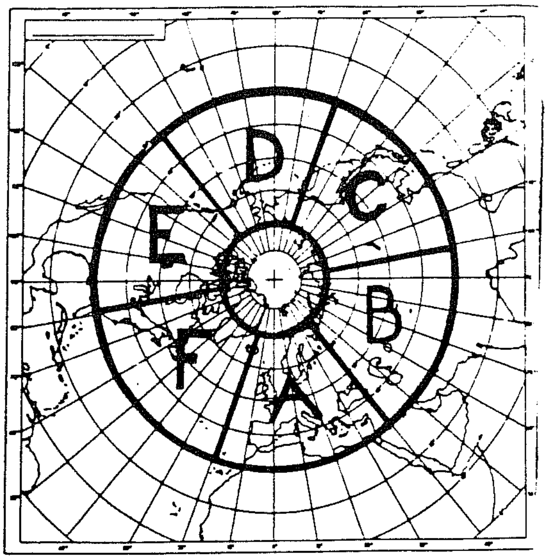

Figure 9 . Domains for regional verification.

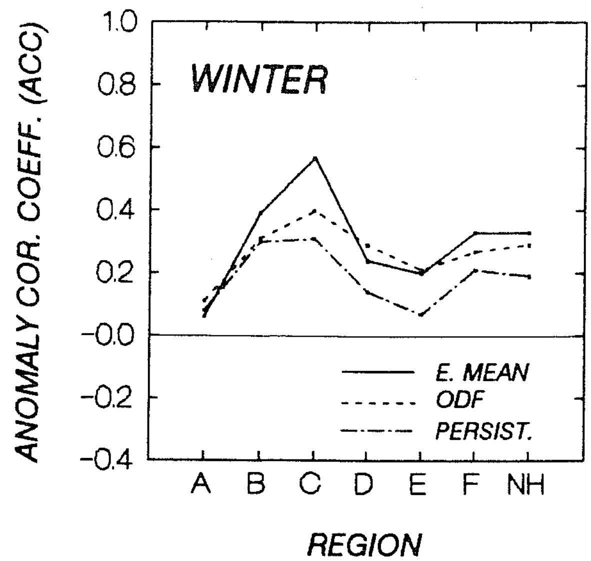

Figure 10. ACC at days 11-20 during winter for mean sea level pressure over regions A to F . Ensemble-mean (-----), ODF (- - -), and persistence (- - -).

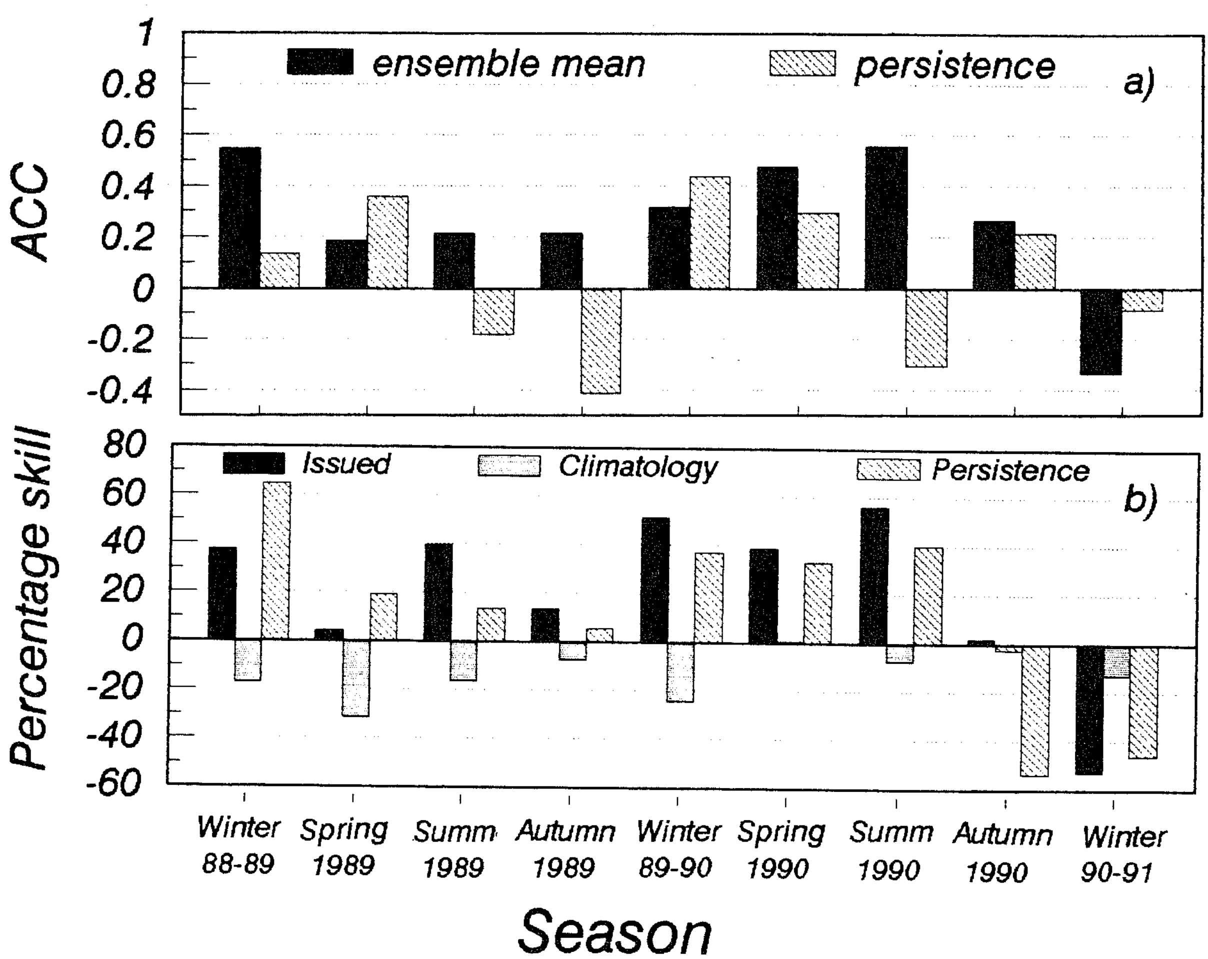

Figure 11. Variation of seasonal skill at days 6-15 for a) ACC of mean sea level pressure over a 30 point grid surrounding the U.K. for ensemble mean and persistence. b) Folland-Painting skill scores for issued, persistence and climatology temperature forecasts over 10 U.K. districts.

The seasonal variations of skill in the MSLP and temperature forecasts for 1988 to 1991 are similar, both showing skilful forecasts in winter, spring, and summer 1990, with a more recent loss of skill in winter 1990/91 (figure 11). The correlation between the two skill measures is 0.66. One exception to the above relationship is summer 1989, where the <ACC> of the ensemble-mean forecast is only 0.2 ,but the FP skill score for temperature is 40%, which is highly skilful compared to the average FP skill score of 18% for summers over the period 1964-86 (Folland *et al.* 1986). This is of interest as the ensemble-mean forecast formed the basis of the MSLP prediction in all of the LRF during this season. As noted above, additional predictability may be introduced from the SST field around the U.K., or from subjective modification of the temperatures at the time the forecasts were issued. However, it should also be stressed that it is unlikely that any one skill measure, such as <ACC>, can give us a complete description of the practical utility of the dynamical forecasts and the results from this one season may be a case in point. A more direct method of assessing the relationship between skill in the MSLP field and skill in the predictions of temperature and rainfall is to derive these quantities directly from the ensembles using the specification equations, without any subjective input and work is in progress along these lines.

In general, the correspondence between the skill of the ensemble-mean forecast MSLP fields and the issued temperature forecasts gives us some confidence that our conclusions concerning the skill of the regional MSLP fields are applicable to the temperature forecasts.

3.8 Case studies of skilful and unskilful ensemble forecasts.

To conclude the discussion of the skill of the LAF ensembles the synoptic evolution of H500 NH anomalies is described for 3 cases which showed above average skill in the ensemble-mean forecast at days 11-20 (17.02.86, 22.12.86, and 02.01.89) and one case which was unskilful at this forecast range (08.12.86).

3.8.1 Ensemble initialised on 17.02.86

This ensemble was the most skilful at days 11-20 during winter 1985/86 (see figure 6). At days 1-10 the ensemble-mean forecast and verifying analysis anomaly patterns are dominated by a strong wavenumber two component, with positive anomalies over Greenland showing the mature phase of an Atlantic blocking event from early February 1986 which both UKMO (Milton 1990) and ECMWF (Brankovic *et al* 1990) LAF ensembles, initialised around 20.01.86, failed to predict. During days 11-20 hemispheric scale circulation changes occur in the verifying analysis which are accurately predicted in the ensemble-mean forecast. The positive anomaly over Greenland is replaced by a negative anomaly , and the positive

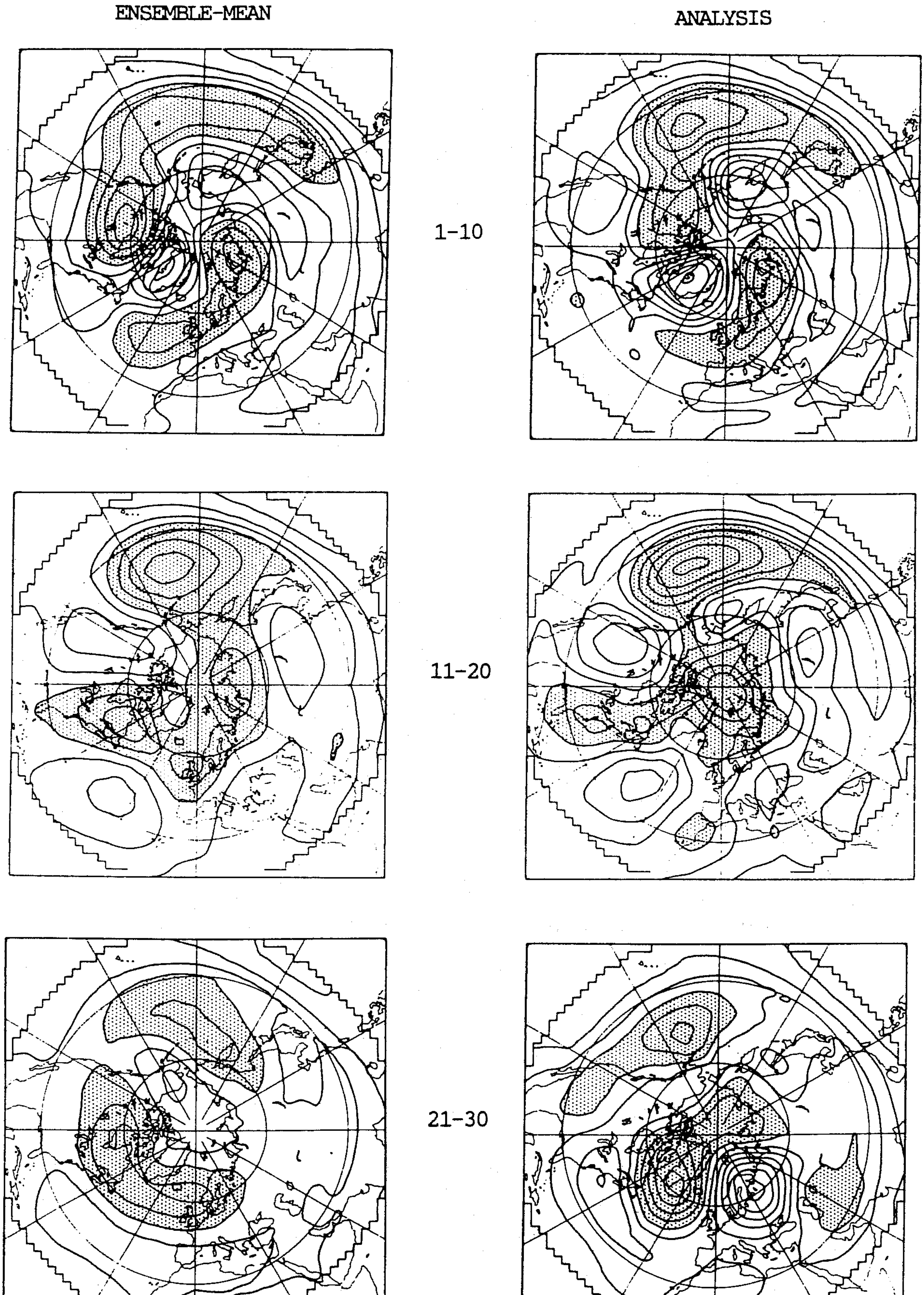

Figure 12. Ensemble mean and verifying analysis 500mb height anomalies at days 1-10, 11-20, and 21-30, for the ensemble initialised on 17.02.86. Contour interval is 6 dam and negative anomalies less than -6 dam are shaded.

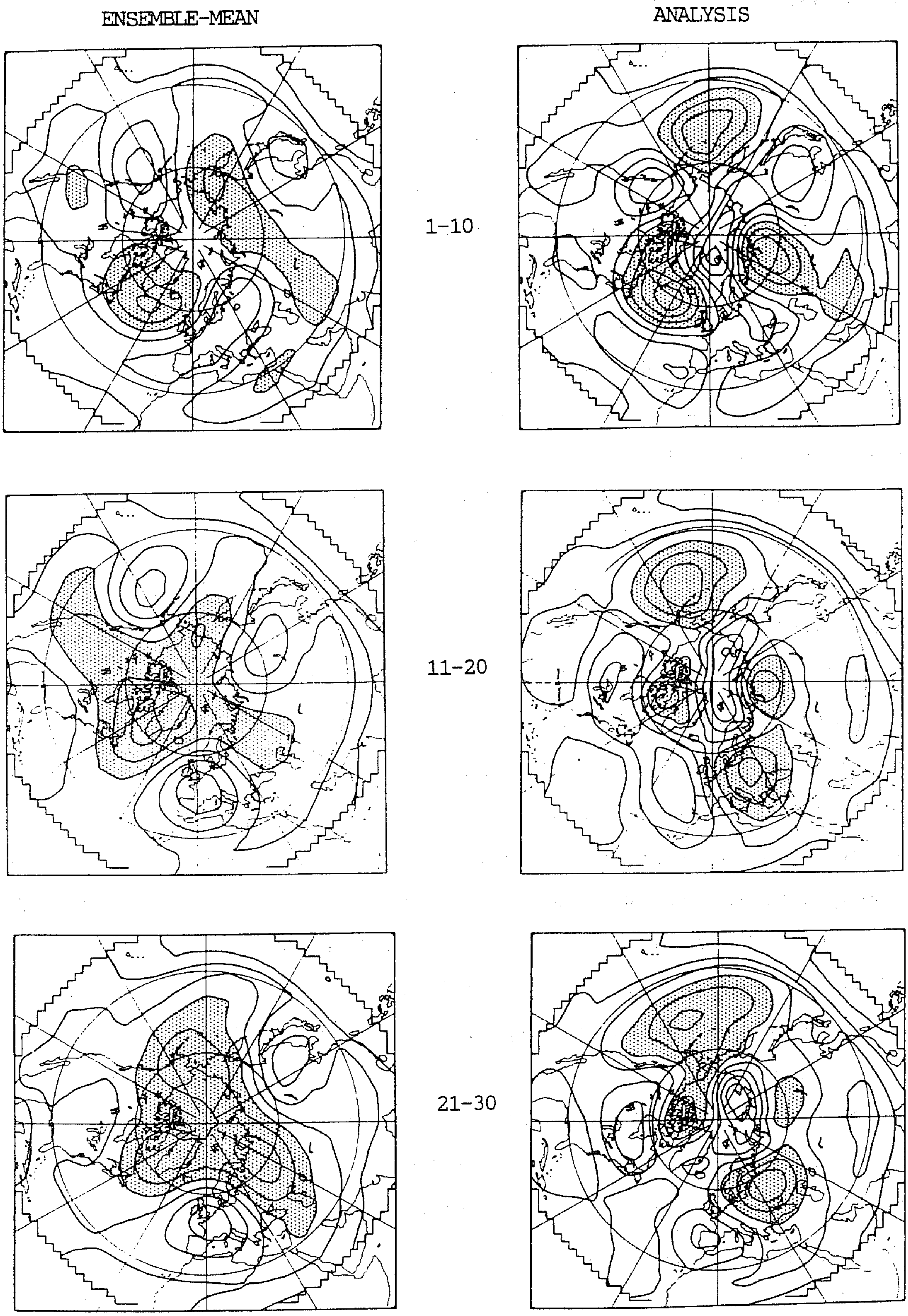

Figure 13. As figure 12 but for the ensemble initialised on 08.12.86.

phase of the Pacific-North American (PNA) anomaly pattern (Wallace and Gutzler 1981) becomes established, with low pressure over the Aleutian region and an amplification of the Rockies ridge. By days 21-30 the focus of attention has shifted to the North Atlantic and Europe where a a blocking ridge over northern Europe has formed in the verifying analysis. All of the ensemble members failed to capture this development. Brankovic *et al.* (1990) discussed this case in detail with respect to their LAF ensemble initialised on 16.02.86 which showed a remarkably similar evolution to that described above, suggesting that the predictable nature of this situation was model independent.

3.8.2 Ensemble initialised on 08.12.86.

Even within the first 10 days this ensemble is different to the verifying analysis (figure 13), particularly in the Pacific and over northern Russia. By days 11-20 a marked Pacific block has developed in the ensemble-mean forecast, whereas the verifying analysis maintains the negative anomaly in the Aleutian region. This is an unusual case as the ensemble has lost skill due to the model undergoing an erroneous transition. A more common scenario is for the atmosphere to undergo a transition which the model does not capture, an example of which is the European block in the previous ensemble. Another example of erroneous Pacific blocking occurs for the ensemble from 04.01.88. By days 21-30 the ensemble-mean forecast anomaly pattern bears a strong resemblance to the models SE pattern for winter (figure 7).

3.8.3 Ensemble initialised on 22.12.86.

This ensemble run from 14 days later was the most skilful of the winter 1986/87. Again for days 1-10 the ensemble-mean forecast and verifying analysis project onto the positive phase of the PNA pattern (figure 14). Large negative anomalies are also observed over central Europe. This is a case of the model persisting the large scale hemispheric anomaly out to 20 days, with some phase differences at days 11-20 ,but generally skilful realisations over the Pacific, North America, and central Europe. Again at days 21-30 the ensemble-mean forecast resembles the model SE and has failed to capture the formation of a Scandinavian block which involved the retrogression of the positive anomaly observed over northern Siberia at days 1-10. The NMC DERF forecasts also failed to predict this block even when the blocking development was within the medium range of the forecasts (Tracton *et al.* 1989). A subsequent diagnostic study by Tracton (1990) suggested that the interaction of planetary- and synoptic-scales, which the model handled poorly, was crucial to the blocking development.

3.8.4 Ensemble initialised on 02.01.89.

This ensemble has been included as it was one of the first run operationally and marked the beginning of a period of mild and persistent conditions over the U.K. during January 1989 which were accurately

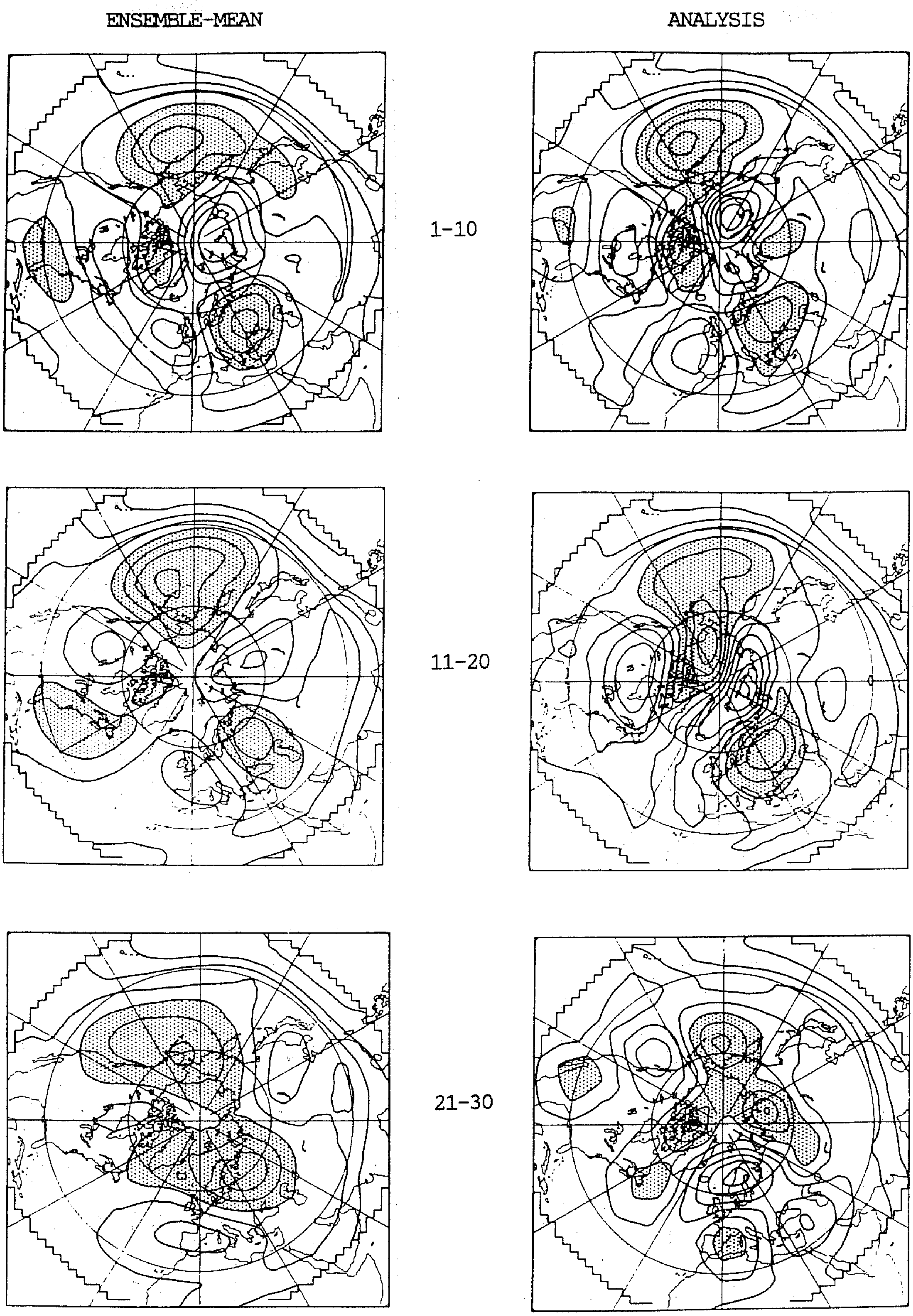

Figure 14. As figure 12 but for the ensemble initialised on 22.12.86.

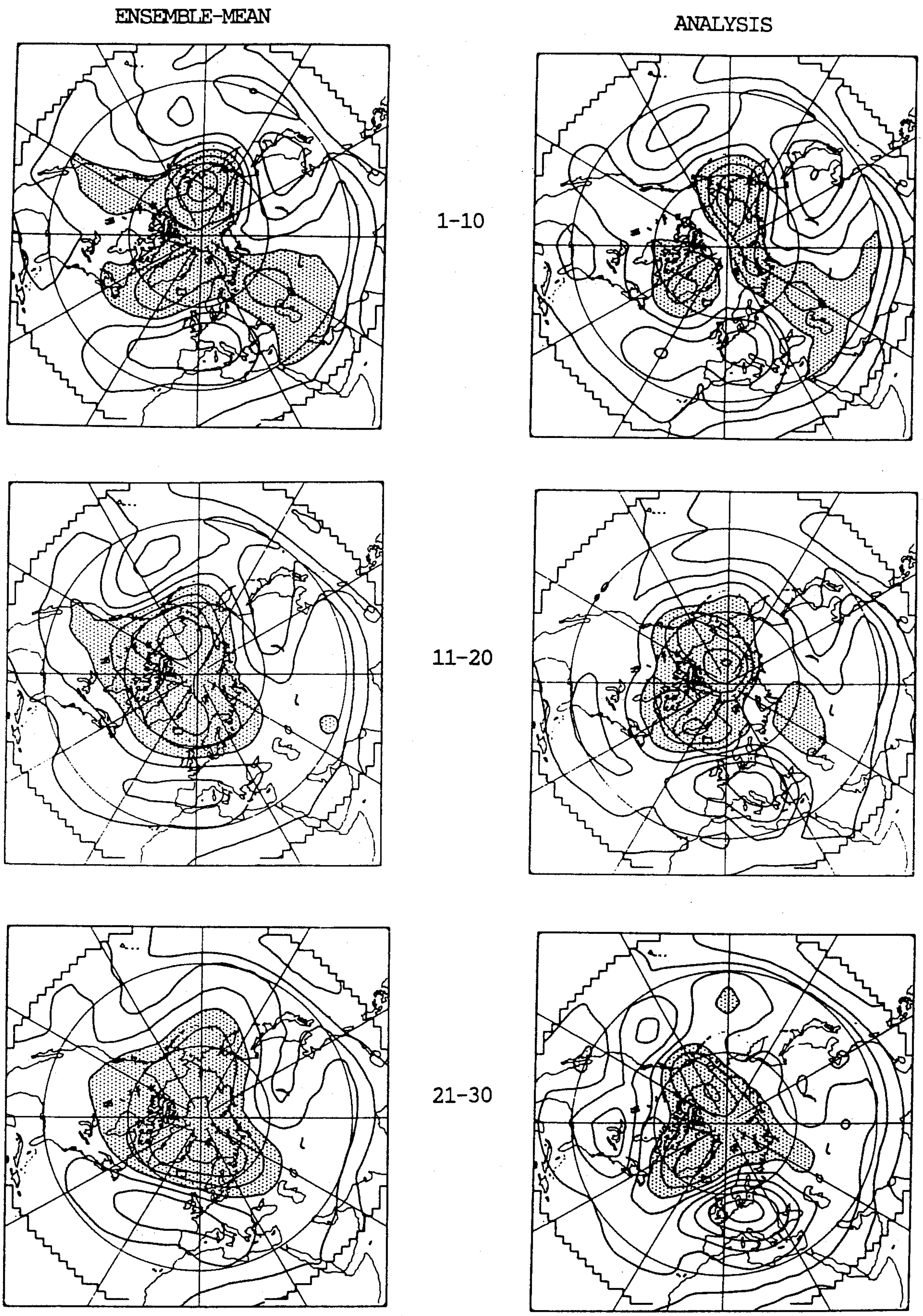

Figure 15. As figure 12 but for the ensemble initialised on 02.01.89.

predicted by this and successive ensembles. The verifying analysis has negative anomalies over the pole with large positive anomalies over western Europe and most of the Pacific, which are predicted by the model at least out to days 11-20, and even out to days 21-30 with some phase differences (figure 15). This anomaly pattern does resemble the models SE at days 21-30 and in this sense the high skill at later forecast ranges may be considered fortuitious. However removing the SE changes ACC very little suggesting that a predictable signal over and above the model SE is present. This is also suggested by the below average levels of spread for the January 1989 ensembles (see section 4).

4. PREDICTION OF FORECAST SKILL.

Having demonstrated that a number of ensemble forecasts are skilful beyond 10 days and that on certain occasions the model forecasts do predict major circulation changes, we now ask whether the ensemble provides us with a means of discriminating between the skilful and unskilful cases in advance. The following sections consider the ensemble spread as a predictor of both hemispheric and regional skill. Ideally, the spread should give us information on variability in skill arising from both variations in the quality of the initial analysis and in the stability of the flow. Such variations in stability of the prevailing flow suggest a class of alternative predictors of skill based solely upon measures of circulation regime, and some dicussion of this is given in sections 4.2 and 4.3.

4.1 Ensemble skill versus ensemble spread.

Two measures of ensemble spread are considered. The first is termed the forecast agreement and is the ACC between each ensemble member and the ensemble-mean forecast averaged over the nine ensemble members. This measures the phase correspondence of the fields within the ensemble. The second is an amplitude based measure which is the standard deviation amongst the ensemble members (i.e. $(s_M)^{1/2}$ - from section 3.1) and is referred to as the R.M.S. spread.

Table 3 gives the correlation coefficients between the ensemble-mean forecast ACC and the forecast agreement for H500 and MSLP over the NH at a variety of forecast ranges. Correlations significant at the 95% level are underlined. These significance levels are derived by assuming that the number of degrees of freedom is approximately one half the number of ensembles in a given season. This takes account of the fact that successive ensembles are not independent, being separated by only 14 days when the decorrelation time for successive 10 day means is of the order of 30 days (i.e. every other ensemble is considered independent). Thus for the 41 winter ensembles we assume 20 degrees of freedom which means a correlation of 0.36 is significant at the 95% level. For the remaining seasons correlations of 0.46 are significant at the 95% level, and for all seasons taken together this value becomes 0.23.

In the first 5 days the agreement is a good predictor of ACC in spring, summer, and autumn, with correlations generally higher for MSLP. In winter the correlations are low at 0.22 (0.23) for H500(MSLP) which by days 1-10 have risen to a significant 0.40 for the MSLP field. On a regional basis the winter correlations are higher at days 1-5, reaching 0.68 for the North Atlantic and European regions (regions F and A in figure 9). At days 6-15 and 11-20 the only significant correlations are for the MSLP field during spring.. For all seasons taken together the correlations remain significant out to days 6-15, although values are low, explaining only 11% of the variance in skill of H500 over the NH. Similar conclusions were reached by previous studies. For example Tracton *et al.* (1989) found correlations of 0.35 (0.09) at days 6-15 (11-20) for H500 over the NH during their DERF study of winter 1986/87. Murphy (1990) found higher hemispheric agreement-ACC correlations, but the results were subject to large sampling errors with only eight ensembles to consider, and for regional assessment the correlations were found to be much lower.

Forecast range	Winter	Spring	Summer	Autumn	All Seasons
1-5	22 (23)	48 (64)	17 (79)	65 (58)	43 (60)
1-10	12 (40)	59 (74)	41 (32)	63 (30)	48 (45)
6-15	14 (29)	42 (76)	34 (-12)	38 (24)	33 (29)
11-20	13 (25)	40 (53)	-04 (-21)	-02 (16)	18 (21)
16-25	20 (24)	03 (13)	-22 (-07)	25 (-32)	19 (08)
21-30	02 (31)	-17 (19)	-20 (09)	13 (-15)	03 (14)

Table 3 - Correlations of forecast agreement versus ensemble-mean ACC for H500 (MSLP) over the Northern Hemisphere for various forecast ranges.

The impact of removing the SE prior to calculating the agreement-ACC correlations is found to be positive at H500 for winter and spring, with greatest impact for the 11-20 day time range. The only other consistent positive impact on correlations is for MSLP during summer. The impact of the SE removal appears to be positive for those fields and seasons where the growth of the normalised mean square mean error (α^2) is largest (section 3.1).

Scatter plots of agreement against ACC for MSLP at day 11-20 during winter and spring show the agreement-ACC correlation in more detail (figure 16). Although the overall correlation of 0.25 for winter

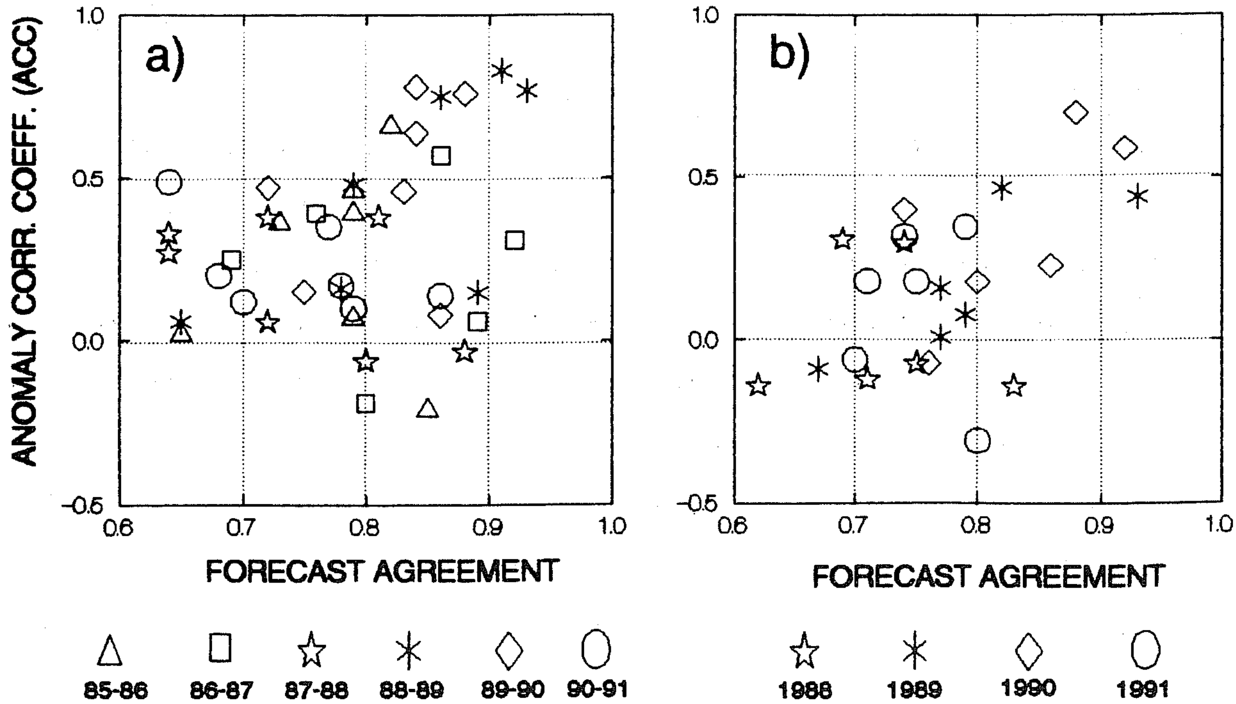

Figure 16. Scatter plots of forecast agreement against ensemble-mean ACC for mean sea level pressure over the Northern Hemisphere at days 11-20 a) Winter b) Spring.

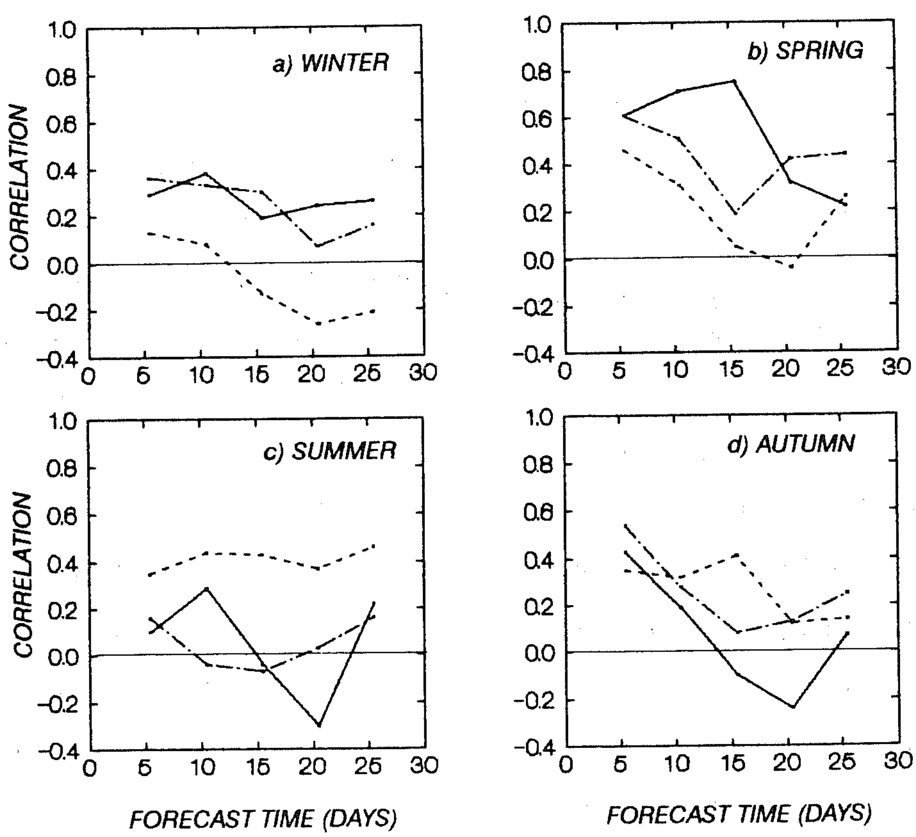

Figure 17. Seasonal correlations between variations in (i) forecast agreement and ensemble-mean ACC (——), (ii) R.M.S. spread ($(s_m)^{1/2}$) and R.M.S. error (- - - -), and (iii) anomaly magnitude of the ODF ($(w_f)^{1/2}$) and ACC of the ODF (-·-·-), for overlapping 10 day means of 500mb height over the Northern Hemisphere. All fields have an estimate of systematic error removed prior to forming the correlations.

is not significant the scatter is not totally random and some interannual variability in the agreement-ACC relationship does exist. For example in winter 1985/86 the agreement of at least 5 of the 7 ensembles gave a reasonable indication of the skill. This was also true for winters 1988/89 and 1989/90,which also had the ensemble forecasts with the largest values of ACC and highest agreements. The scatter for the other winters is much larger. A weak but positive signal is apparent in the winter season, with all ensembles with agreements less than 0.8 having ACC<0.5, whereas 47% of the ensembles with agreements greater than 0.8 have ACC>0.5 (figure 16(a)). The significant correlation of 0.53 for the spring ensembles is due mainly to the positive influence of the 5 ensembles from spring 1989 and 1990 (figure 16 (b)).

The low agreement-ACC correlation during winter is due to the class of ensembles with low ACC but high forecast agreement. At least one ensemble in each winter falls into this class, which represents low frequency transitions occurring in the atmosphere which are missed by the model. It is possible that the inability of the model to predict major regime transitions is due to a sampling problem, and that 9 members are not sufficient to capture the full range of possible outcomes. However it seems more likely that some fundamental dynamical mechanism may be misrepresented in the model in the extended range. Candidates include the interaction between synoptic and planetary scales and tropical extratropical interactions.

Correlations for the R.M.S. spread against R.M.S. error are generally lower than agreement-ACC correlations in winter and spring and higher in summer and autumn (figure 17), although significant correlations only occur at days 1-5 (not shown). The higher correlations in autumn can be trivially explained by the impact of the seasonal cycle on both R.M.S. spread and R.M.S. error. Also shown in figure 17 are the correlations between the forecast anomaly magnitude ($(w_f)^{1/2}$) of the ODF as predictor and the ACC of the ODF as predictand. Palmer and Tibaldi (1988) discussed the fact that both ACC and agreement measures are sensitive to the anomaly amplitude and found that the latter was a better predictor of the ACC of individual forecasts in the medium range than the ACC based spread between two forecasts initialised twelve hours apart. For our results the forecast anomaly magnitude is comparable to the forecast agreement as a predictor of skill in winter summer and autumn, but has much lower correlations than agreement in spring beyond days 1-10. The higher correlations observed for the agreement-ACC relationship in comparison to that for R.M.S. spread-R.M.S. error relationship are partially a reflection of the correlation of forecast agreement and ACC with the anomaly magnitude.

Regional spread-skill correlations (not shown) are similar to those on the hemispheric scale. As an example of the information contained in regional agreement measures we have considered the agreement-ACC relationship in a probabilistic way by constructing a 5 x 5 agreement-ACC contingency table for

MSLP over the U.K. and Europe region (region A in figure 9) and categorizing the 41 winter ensembles according to their agreement and ACC values at days 6-15 (Table 4) and at days 11-20 (Table 5). This is similar to the approach adopted by Murphy (1990) except that it is relevant to a specific region whereas his results were integrated over a number of regions in the NH. In addition we have 41 ensembles all for one season, compared to his eight ensembles covering all seasons. All fields have been corrected for seasonal estimates of the model SE.

	ACC categories					
Agreement categories	1 (0.8)	2 (0.4)	3 (0.0)	4 (-0.4)	5 (-0.8)	Number of ensembles.
1 (0.95)	.50	.17	.17	.00	.16	6
2 (0.85)	.28	.39	.22	.06	.06	18
3 (0.75)	.00	.20	.50	.30	.00	10
4 (0.65)	.00	.66	.12	.12	.00	6
5 (0.55)	.00	.00	.00	1.00	.00	1
All	.19	.34	.27	.15	.05	1.00/ 41

Table 4 - Contingency table of forecast agreement and ensemble mean ACC for MSLP over the U.K. and Europe (region A) at days 6-15. ACC categories are of width 0.4 and forecast agreement categories of width 0.1. The midpoints of both ACC and agreement categories are shown in brackets.

	ACC categories					
Agreement categories	1 (0.8)	2 (0.4)	3 (0.0)	4 (-0.4)	5 (-0.8)	Number of ensembles.
1 (0.95)	.00	1.00	.00	.00	.00	1
2 (0.85)	.45	.27	.18	.09	.00	11
3 (0.75)	.15	.15	.23	.31	.15	13
4 (0.65)	.00	.17	.33	.33	.17	6
5 (0.55)	.00	.20	.20	.30	.30	10
All	.17	.22	.24	.24	.13	1.00/41

Table 5 - As table 4 but for days 11-20.

The agreement ACC correlations are 0.38 (0.43) for days 6-15 (11-20) and are the largest regional values and larger than the corresponding hemispheric values. For both periods the contingency tables show that for the high agreement categories (1,2), containing 58% (30%) of the 41 ensembles at days 6-15 (11-20), a significant peaked distribution towards the skilful end of the ACC categories is observed. For example at days 6-15 only 19% of all cases fall within the most skilful ACC category. However if we confine ourselves to those cases with high agreement (category 1) this figure rises to 50%. Similar conclusions can be drawn at days 11-20 for the second highest agreement category (2).

Summarising, we note that measures of the ensemble spread are unreliable predictors of the hemispheric skill of H500 and MSLP beyond 10 days, although a small but significant correlation between forecast agreement and ACC is evident at days 6-15 when measured across all seasons, and at days 11-20 for MSLP in spring. There is evidence of a practically useful prediction of regional ACC using regional agreement for the winter season at days 6-15 and 11-20 if one considers the agreement-ACC relationship in a probabilistic manner as in Murphy (1990).

4.2 Dependence of forecast skill on circulation regime.

Several recent studies have identified possible relationships between circulation regime and forecast quality. Palmer (1988) found that both skill of medium-range forecasts at ECMWF and the extended-range forecasts run by Miyakoda *et al.* (1986) were dependent on the sign of the PNA pattern. He hypothesised that the negative phase of this pattern was associated with increased barotropic instability and hence larger error growth rates than the positive phase. Similar conclusions were reached by O'Lenic and Livezey (1989) using the NMC medium-range forecasts. Molteni and Palmer (1991) used this relationship as one predictor in a practical skill prediction scheme applied to ECMWF medium-range forecasts. In the remainder of the paper we consider whether the skill of the UKMO winter LAF ensembles is related in any way to some measure of regime either in the forecast or verifying flow.

4.2.1 Dependence of forecast skill on the PNA mode of variability.

The PNA anomaly pattern arises as one of the principal modes of variability in a number of objective observational studies to identify regimes (Wallace and Gutzler 1981;Barnston and Livezey 1987; Molteni *et al.* 1990). Following the study of Barnston and Livezey, rotated principal components have been used to derive a set of modes using H500 for half months during November and December from 1973-88. The dominant modes resemble the North Atlantic Oscillation (NAO - figure 18(a)) in its negative phase (as described in Wallace and Gutzler 1981) and the PNA mode in its negative phase (figure 18(b)). To be consistent with the definition of NAO and PNA in Wallace and Gutzler all indices derived by projections

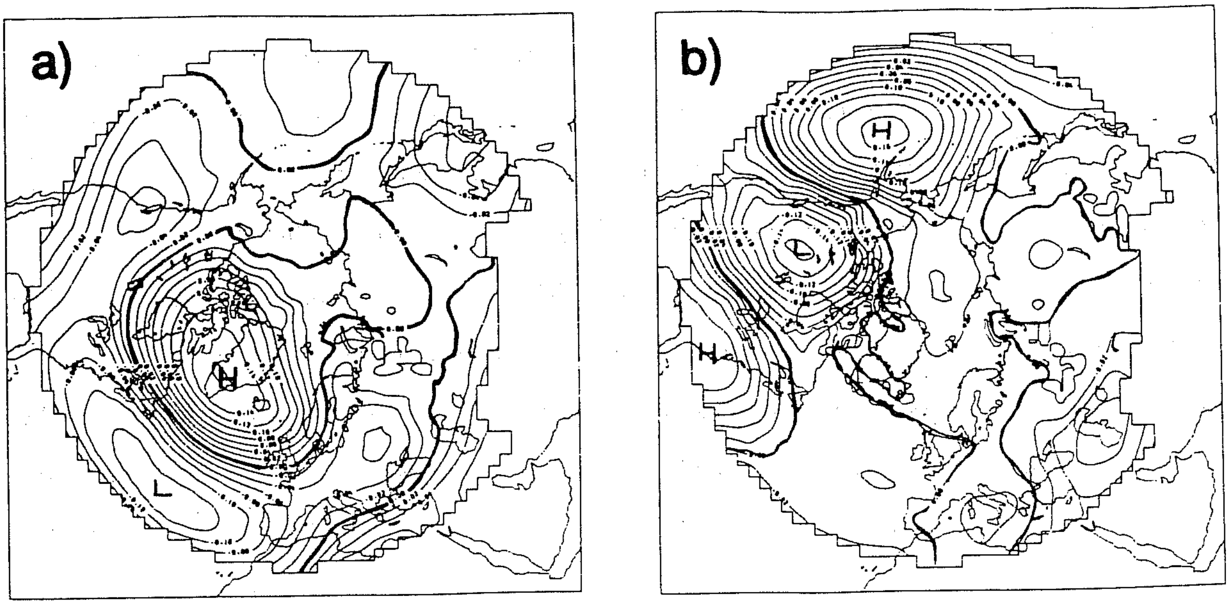

Figure 18. Loading patterns derived from a principal component analysis of the 500mb height field for half-months from November to December 1973-88. a) NAO mode b) PNA mode.

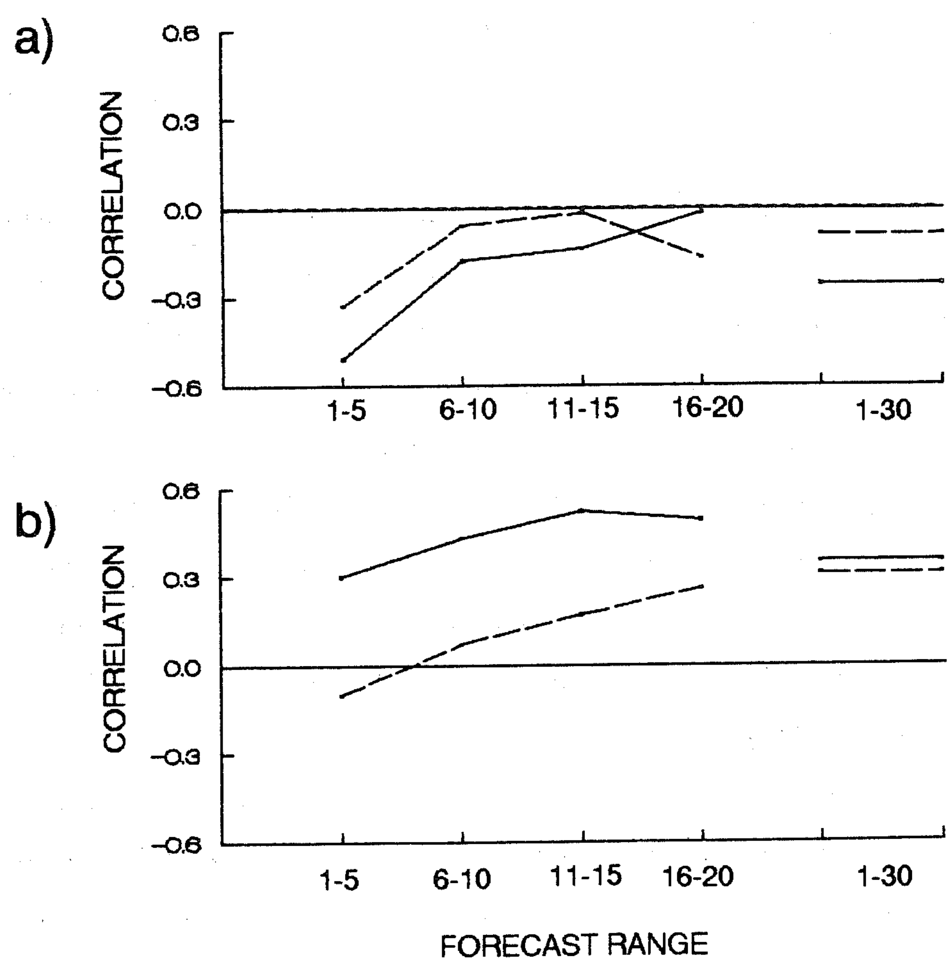

Figure 19. Correlations at various forecast ranges between variations in a) the PNA mode in the forecast flow and R.M.S error over the U.S.A, of 364 dynamical forecasts comprising the 41 winter ensembles (- - - - -), and the PNA mode in the flow of 41 verifying analyses and R.M.S. error over the U.S.A of the 41 winter ensemble means (——). b) as a) but for NAO mode and ACC over the North Atlantic and Europe. All indices and scores are for 500mb height.

onto figure 18a and 18b are multiplied by -1 so that a posititive index represents the positive phase of each pattern. This simple two dimensional phase space defined by the NAO and PNA patterns is used to assess the dependence of model skill on regime, although we acknowledge that these regimes are in no way a complete description of the possible modes of variability in the atmosphere or the model.

The H500 anomalies in the ensemble forecasts and verifying analyses from our winter sample covering 1985-91 have been projected onto these patterns for various forecast ranges to give us a measure of the amplitude of these modes in both the model and real atmosphere. These amplitude indices are then correlated with the forecast skill of the ensembles. Figure 19 shows the correlations between the skill of each ensemble member in the sample (364 individual forecasts) and its projection onto the two modes of variability in the forecast flow for various forecast ranges. Also shown is the correlation between the ensemble-mean forecast skill and the projection of the verifying analyses onto the two modes. The correlations shown are for regions and skill measures which gave the highest correlations with the PNA and NAO modes. For the PNA mode this was the U.S.A region (160 -60°W, 30-70°N) and R.M.S. error as the measure of skill (figure 19(a)), and for the NAO mode the North Atlantic and Europe region (60°W- 40°E, 30-70° N) and ACC measure of skill (figure 19(b)).

At days 1-5 the highest correlation of -0.51 is between the PNA mode in the verifying analyses and R.M.S. error of the ensemble-mean over the U.S.A. This correlation implies that when the PNA mode is in its negative phase (i.e. that shown in figure 18(b)) in the verifying analysis, then R.M.S. error is large and the forecast is unskilful, agreeing with the results of Palmer (1988). For the remaining pentads (6-10, 11-15 and 16-20) the correlations are small and not significant, and for days, and for days 1-30 the correlation is slightly higher at 0.26. The above correlations involving the verifying analyses give us possible diagnostic relationships between skill and regime, but in order to have an *a priori* prediction of forecast skill we must consider whether the forecast flow regime itself is a potential predictor. The correlation between the PNA mode in the forecast flow and R.M.S. error over the U.S.A shows a similar trend to that for the verifying analyses but the correlations are smaller, being -0.33 at days 1-5 and -0.09 at days 1-30. This implies that the PNA mode is not a reliable predictor of the skill in our sample of 30-day forecasts. However, in common with the agreement-ACC relationship we note that some interrannual variability does exist in the PNA-skill relationship. Scatter plots of the PNA index in the forecast flow against R.M.S error for days 1-30 over the U.S.A. show that for winters 1985/86 and 1986/87 there is a good relationship between skill and the sign of the PNA mode in the forecast flow, with correlations of -0.48 and -0.77 respectively for the two winters (figure 20). However the opposite correlation occurs for winters 1988/89 and 1990/91 implying that when the forecast flow contains a PNA-

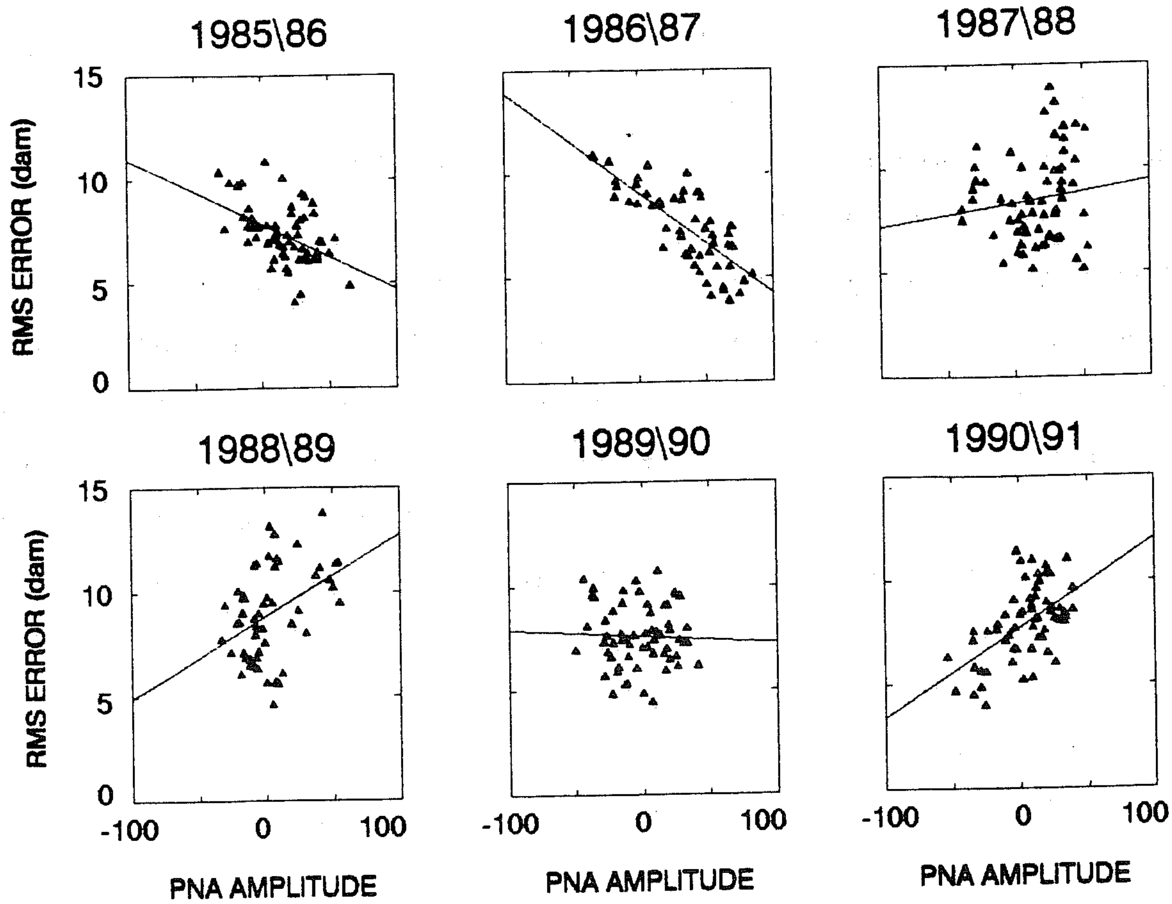

Figure 20 Scatter plots of PNA index for the forecast flow and forecast R.M.S. error at days 1-30 for 500mb heights over the U.S.A . The 364 dynamical forecasts from the 41ensembles for winters 1985/86 to 1990/91 are plotted.

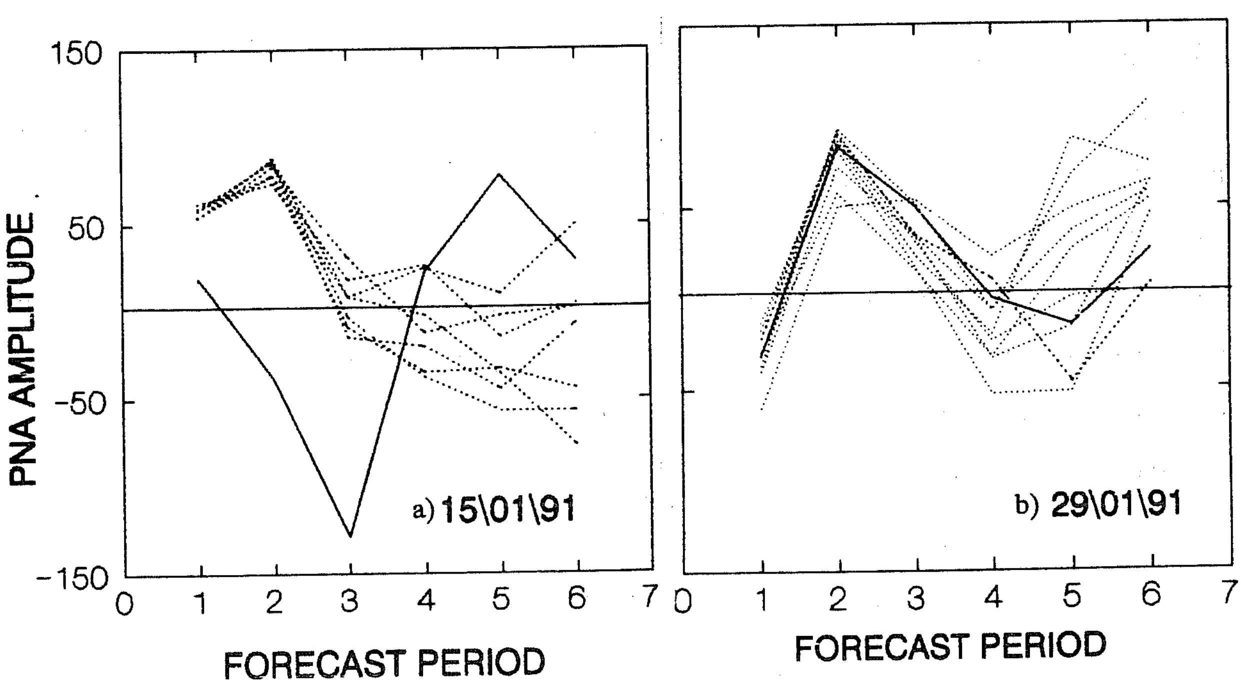

Figure 21. Regime transitions involving the PNA index for 500mb height anomalies of ensemble forecasts (- - - -) and verifying analyses (——) for ensembles initialised on the a) 15.01.91 and b) 29.01.91. Periods are forecast pentads (1=1-5, 2=6.10....6=26-30).

anomaly pattern the forecast is skilful. A possible explanation for this conflicting behaviour is provided by the fact that during these two winters the model accurately predicted low amplitude PNA- patterns, but during observed transitions to high amplitude PNA- the model tended to move to a PNA+ pattern. This is discussed in more detail in the next section.

For the NAO index the largest correlation is again for the verifying analyses and skill, but this time it is an increasing function of forecast range such that by days 11-15 the correlation is 0.52 (figure 19b). Therefore, when the NAO is in its negative phase (i.e. that of figure 18(a)) then ACC is lower than average and forecasts are unskilful. The negative phase of NAO represents a reduction of the westerlies over the Atlantic and is related to Atlantic blocking. The relationship between the NAO pattern and skill reflects a systematic deficiency in this model and others, which is the inability to enter a blocked type of regime with sufficient frequency in the medium- and extended-range (Tibaldi and Molteni 1990). However,we note that the correlation between the forecast flow and ACC is also an increasing function of forecast range and at days 1-30 the correlation of 0.31 is comparable to the diagnostic correlation involving the verifying analyses. This is difficult to explain physically and may be due to the positive influence of the ensemble forecasts from winters 1988/89 and 1989/90, where anomalous persistent zonal flow (figure 15) was predicted with unusually high ACC values in the extended range, where ACC is generally low on average.

4.2.2 Regime predictability.

Although a reliable predictor of forecast skill based upon prevailing regime has not been found, it is still useful to consider the performance of the model in handling persistence in regimes and transitions between them. This is particularly true if the relationship between skill and regime is non-linear, a fact ignored when using linear correlations. In order to investigate this a transition probability matrix for the PNA mode has been calculated using the 41 winter verifying analyses and the 364 individual 30 day forecasts from the LAF ensembles. This is done by splitting the PNA mode into the following four categories of high and low amplitude PNA+ and PNA- according to their index :

PNA+ (H)	index > 40
PNA+ (L)	0 < index < 40
PNA- (L)	-40 < index > 0
PNA- (L)	index < -40

then classifying both forecasts and verifying analyses according to which category they fall in at days 11-15 and also which category the initial conditions fall into. This gives a transition probability matrix for transitions between (or persistence in) PNA states between day 0 and days 11-15 in the verifying analyses (Table 6) and dynamical forecasts (Table 7). (Note: the values corresponding to the distribution of cases (%) at day 0 amongst the four PNA states are slightly different for the analyses and forecasts as 4 of the 41 LAF ensembles had less than 9 members, which introduces a non-uniform weighting into the distributions. However this does not significantly affect the conclusions).

For the verifying analyses the values along the diagonal of the transition probability matrix are all 10%, and therefore 40% of the 41 verifying analyses are associated with persistence in one of the PNA states. The off diagonal elements representing transitions between the various states are generally smaller, with the largest values occurring for transitions to the high amplitude PNA+(H) state.

If we now consider the corresponding matrix for the forecasts we immediately note that at days 11-15 there is an underpopulation of both positive and negative high amplitude PNA states compared to the verifying analyses, although this error is largest for the PNA-(H) states. This failure to maintain high amplitude planetary wave states is a common deficiency of dynamical models in the extended range. Interesting assymetries are apparent in the model performance in either sign of the PNA mode. For example in persistence of PNA+(H) states there are 7% of forecasts compared to 10% of verifying analyses, whereas for persistence of PNA-(H) there are only 2% of forecasts compared to 10% of verifying analyses. The model appears to perform better in large amplitude PNA+ states than in large amplitude PNA- states, but its performance in persistence of PNA+(L) is similar to that in persistence of PNA-(L). The figures in brackets alongside the forecast transition probailities give the % of those forecasts observed in that category which are correctly predicted. These figures show that on average a larger proportion of correct predictions occurr on the PNA+ side of the distribution than the PNA- side. In particular , the model predicts transitions from PNA+ to PNA- with the correct frequency but few of the predictions are correct (both high and low amplitude states are considered together), whereas a greater proportion of transitions from PNA- to PNA+ are correctly predicted. Examples of this regime behaviour were given in the case studies from 08.12.86 and 17.02.86 discussed in section 3.8. Obviously with such small samples of independent forecasts and analyses these figures are subject to considerable sampling error. However we note that the general conclusions are in agreement with the results of Molteni and Tibaldi (1990) for the ECMWF models.

A graphic example of transitions involving the PNA mode is provided by two ensembles from winter 1990/91 (figure 21). The ensemble initialised on 15.01.91 completely failed to capture a transition to PNA-(H) occurring in the verifying analysis at days 11-15, but the ensemble initialised two weeks later correctly predicted the transition back to PNA+(H) and the subsequent evolution. Although only one

		Days 11-15				
		PNA+		PNA-		
		H	L	L	H	
Day 0, PNA +	H	10	2	2	5	19
Day 0, PNA +	L	10	10	0	5	25
Day 0, PNA -	L	10	5	10	5	30
Day 0, PNA -	H	5	5	7	10	27
		35	22	19	25	

Table 6 - Transition probability matrix for PNA mode in the verifying analyses for transitions or persistence in the PNA mode between day 0 and days 11-15. All figures are expressed as a % of the total number of entries, which is 41 winter verifying analyses covering 1985-1991. All figures have been rounded to the nearest whole number.

		Days 11-15				
		PNA+		PNA-		
		H	L	L	H	
Day 0, PNA +	H	7 (83)	7 (26)	5 (00)	1 (91)	20
Day 0, PNA +	L	7 (20)	11 (34)	4 (00)	3 (00)	25
Day 0, PNA -	L	4 (65)	10 (19)	9 (62)	5 (00)	28
Day 0, PNA -	H	4 (61)	10 (30)	11 (32)	2 (83)	27
		20	38	29	11	

Table 7 - As table 6 but for the corresponding 364 individual dynamical forecasts comprising the 41winter ensemble forecasts 1985-91. All entries are expressed as a % of the total number of individual dynamical forecasts (364) contained in the 41 ensembles. The additional figures in brackets represent the % of occasions the prediction of that particular transition or persistence was correct.

mode of low frequency variability has been examined in detail, clearly an accurate description of model regime behaviour is a neccessary step in elucidating possible dynamical mechanisms which underly the variations in dynamical extended-range forecast skill.

5. CONCLUSIONS

This paper has been concerned with the practical predictability of UKMO dynamical model prediction methods for the extended range. Nine member lagged average forecast (LAF) ensembles run with a version of the UKMO NWP model were a regular feature of the UKMO operational long-range forecast (LRF) between December 1988 and June 1991, and the impact of the ensemble technique on extended-range skill has been described, as well as the ability of the ensembles to provide an *a priori* indication of ensemble-mean skill using the ensemble spread.

Estimates of potential predictability based upon the ensemble spread suggest a predictable signal remains out to days 21-30 on average in all seasons. The practical levels of average hemispheric skill as measured by anomaly correlation coefficient (ACC) are positive but low beyond days 1-10, although considerable variability about the average values does exist. Winter is the most skilful season with ensemble-mean forecast ACC values at days 11-20 and 16-25 greater than those at days 6-15 in the other three seasons. In general the ensemble-mean forecast provides a more skilful prediction than persistence , most noteably in winter. The ensemble-mean forecast skill is greater than the average skill of the individual ensemble members in all seasons, with the largest improvements beyond 10 days during winter. However the improvements are smaller than predicted by perfect model theory (Murphy 1988). Beyond day 10 the ensemble members become equiprobale and the ensemble-mean forecast is more skilful on average than the latest member (ODF) in winter and autumn, and comparable to the ODF in spring and summer. On a regional and case-by-case basis the ensemble-mean forecast is often a considerable improvement over the ODF at days 11-20. This 11-20 day forecast range has been identified in previous studies as the optimum forecast range in which to benefit from LAF ensemble forecasts, and our results bear this out, particularly for the relatively skilful winter season. Additional gains in the ensemble-mean forecast ACC can be achieved in most seasons out to at least day 20 by correcting for the models systematic error *a posteriori*. A question not fully addressed in these results is whether the small values of average ACC beyond day 10 have any practical utility when converted to elements of interest to the user such as temperature and rainfall. Preliminary results suggested that ACC may not always be an appropriate measure of the practical utility of the ensembles and more work is needed in this area.

A variety of predictors of forecast skill were investigated. Correlations between forecast agreement and

ACC are largest at days 1-10 with a value of 0.60 measured across all seasons for MSLP . As with ACC itself this agreement-ACC correlation falls of rapidly with forecast range, but is still a significant 0.33 (0.29) for H500 (MSLP) at days 11-20 over all seasons. On a seasonal basis only the correlation of 0.53 for MSLP at days 11-20 during spring is significant at the 95% level, and for later time ranges no significant correlations are found. Although weak, the agreement-ACC relationship may have some practical value when expressed in a more probabilistic framework. This was demonstrated for prediction of regional skill at days 6-15 and 11-20 and confirms the conclusions of Murphy (1990) regarding the practical prediction of regional skill. Undoubtedly, the poor performance of the ensemble spread measures as predictors of forecast skill is due the class of ensembles where agreement is high but ACC is low, representing observed regime transitions poorly predicted by the model. Alternative predictors based upon circulation regimes such as the PNA mode of variability also failed to produce a reliable predictor of skill. However an analysis of the models ability to effect transitions involving the PNA mode revealed important assymetries in the model's performance, implying that the model has difficulty in persisting a large negative PNA anomaly or in undergoing a regime transition to this anomaly at the correct time, preferring persistence in or transitions to the positive phase of the PNA mode.

Since June 1991 the new UKMO Unified model (UM) has been used for operational LAF ensembles and preliminary results have shown higher levels of variability in the UM compared to the old NWP model. Future work will concentrate on elucidating the dynamical mechanisms involved in skilful (and unskilful) regime transitions in the UM model, as well as exploring techniques for interpreting the ensembles as probability forecasts, particularly for the weather related elements such as temperature and rainfall.

6. References

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

Bell, R.S., and A. Dickinson, 1987: The Meteorological Office operational numerical weather prediction system. Sci. Paper Meteorol. Off., No. 41.

Bourke, W., P. Mullenmeister, K. Arpe, D. Baumhefner, P. Caplan, J.L. Kinter, S.F. Milton, W.F. Stern, and M. Sugi, 1991: Systematic errors in extended range predictions. CAS/JSC working group on numerical experimentation Report. (In preparation).

Brankovic, C., T.N. Palmer, F. Molteni, S. Tibaldi, and U. Cubasch, 1990: Extended-range predictions with the ECMWF models: time lagged ensemble forecasting. Quart. J. Roy. Meteorol. Soc., 116, 867-912.

Corby, G.A., A. Gilchrist, and P.R. Rowntree, 1977: United Kingdom Meteorological five-level general circulation model. Methods in Comp. Phys., 17, 67-166.

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

Dalcher, A., E. Kalnay, and R.N. Hoffman, 1988: Medium range lagged average forecasts. Mon. Weath. Rev., 116, 402-416.

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

Déqué, M., 1991a: 44-day ensemble forecasts with the T42-L20 French spectral model. (This volume)

Déqué, M., 1991b: Removing the model systematic error in extended-range forecasting. Annales Geophysicae, 9, 242-251

Folland, C.K., and A. Woodcock, 1986: Experimental monthly long-range forecasts for the United Kingdom. Part I- Description of the forecasting system. Meteorol. Mag., 115, 301-317.

Folland, C.K., A. Woodcock, and L.D. Varah: Experimental monthly long-range forecasts for the United Kingdom. Part III - Skill of the monthly forecasts. Meteorol. Mag., 115, 377-395

Harrison, M.S.J., T. Legg, A.W. Colman, and A.P. Woodcock, 1991: Techniques of operational long-range forecasts at the U.K. Met Office. In Extended abstracts submitted to the ICTP/WMO international technical conference on long range weather forecasting research, Trieste, 8-10 April 1991, 109-112. WMO/TD No. 395.

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

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

Livezey, R.E., 1990: Variability of skill of long-range forecasts and implications for their use and value. Bull. Amer. Meteor. Soc., 71, 300-309.

Maryon, R.H, and A.M. Storey, 1985: A multivariate statistical model for forecasting anomalies of half-monthly mean surfave pressure. J. Climatol., 5, 561-578.

Milton, S.F., 1990: Practical extended-range forecasting using dynamical models.Meteorol. Mag., 119, 221-233.

Molteni, F., and S. Tibaldi, 1990: Regimes in wintertime circulation over northern extratropics. II: Consequences for dynamical predictability. Quart. J. Roy. Meteorol. Soc., 116, 1263-1288.

Molteni,F., and T.N. Palmer, 1991: A real time scheme for the prediction of forecast skill. Mon. Weath. Rev., 119, 1088-1097.

Miyakoda, K., J. Sirutis, and J. Ploshay, 1986: One month forecast experiments - without anomaly boundary forcings. Mon. Weath. Rev., 114, 2363-2401.

Murphy, A.H., 1971: A note on the ranked probability score. J. Appl. Meteorol., 10, 155-156.

Murphy, A.H., and E. S. Epstein, 1989: Skill scores and correlation coefficients in model verification. Mon. Weath. Rev., 117, 572-581.

Murphy, J.M., 1988: The impact of ensemble forecasts on predictability. Quart. J. Roy. Meteorol. Soc., 114, 463-493.

Murphy, J.M., 1990: Assessment of the practical utility of extended-range ensemble forecasts. Quart. J. Roy. Meteorol. Soc., 116, 89-125.

Murphy, J.M., and A. Dickinson, 1989: Extended-range prediction experiments using an 11-level GCM. Meteorol. Atmos. Phys., 40, 61-83.

O'Lenic, E. and R. Livezey, 1989: Relationships between low frequency anomalies and systematic errors in medium and extended-range forecasts. Mon. Weath. Rev., 117, 1262-1280.

Palmer, T.N., 1988: Medium and extended-range predictability and stability of the Pacific\North American mode.Quart. J. Roy. Meteorol. Soc., 114, 691-713.

Palmer, T.N., and S. Tibaldi, 1988: On the prediction of forecast skill. Mon. Weath. Rev.,116, 2453-2480

Tibaldi, S., and F. Molteni, 1990: On the operational predictability of blocking. Tellus, 42A, 343-365

Tracton, M.S., K. Mo, W. Chen, E. Kalnay, R. Kistler, and G. White, 1989: Dynmical Extended Range Forecast (DERF) at the National Meteorological Center. Mon. Weath. Rev., 117, 1604-1635.

Tracton, M.S., 1990: Predictability and its relationship to scale ineraction processes in blocking. Mon. Weath. Rev., 118, 1666-1695.

Tribbia, J.J, and D.P. Baumhefner, 1988: Estimates of the predictability of low frequency variability with a spectral general circulation model. J.A.S., 16, 2306-2317.

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

Wilson, C.A., and staff, 1990: First results from a unified forecast/climate model. WGNE report No. 14, 6.7-6.9.

Yamada, S., S. Maeda, T. Kudo, T. Iwasaki, and T. Tsuyuki, 1991: Dynamical one-month forecast experiments with the JMA global prediction model. Jour. Met. Soc. Japan., 69, 153-159.