PROSPECTS FOR PRACTICAL EXTENDED RANGE ENSEMBLE FORECASTING

A Dickinson, J M Murphy and D Richardson

Meteorological Office
Bracknell, UK

1. INTRODUCTION

The problem of extended-range prediction, on the monthly time scale, has a strong probabilistic element. This is reflected in the ensemble forecast technique, in which a number of numerical model integrations are produced from distinct initial states, each nominally compatible with the uncertainties associated with some initial analysis. The differences grow with time, eventually reaching saturation level, represented by the separation between integrations from random initial states. At some prior stage, the distribution of integrations forms a probability forecast for the atmospheric state.

The potential benefits of ensemble forecasting have been demonstrated using the perfect model approach (e.g. Seidman, 1981; Murphy, 1988), where nature is itself represented by a model integration. Ensemble averaging can improve skill significantly through the elimination of random forecast errors, and the spread of the distribution can be used to predict the forecast skill. In practice the capacity to realise such benefits is limited by the model's deficiencies, which introduce an additional source of forecast error. The more skilful the model, the more useful ensemble forecasting should prove. However, on some occasions skilful extended-range predictions are possible beyond the normal limit of deterministic predictability, using a GCM capable of only modest accuracy at medium range (Mansfield, 1986). In such cases ensemble averaging can improve skill (Murphy, 1988), and geographic variations in local ensemble

spread are found to correlate with corresponding variations in local skill. Kalnay and Dalcher (1987) have found similar evidence of local spread/skill correlation in medium-range forecasts.

In this paper some results are given from a set of eight ensemble forecasts, produced using the UK Meteorological Office (UKMO) global 11-level GCM. The results supplement those from a large sample of individual extended-range integrations of the same model described in Murphy and Dickinson (1988), hereafter denoted by MD. It was found that the model's mean forecast skill, relative to the observed climate, remained significantly positive out to one month over the domain 30-90°N. However, this depended on the availability of an accurate estimate of the model's systematic error, removal of which improved the skill. There were relatively few integrations in which the extended-range skill remained consistently above average on the hemispheric scale. Nevertheless, the frequency of occurrence of high local skill, important in the context of the present paper, was encouragingly large.

The 11-level model ensemble forecast results are presented in section 3. An important concern is the sensitivity of the results to model formulation. Apart from the aforementioned implications for the ensemble technique, it is of particular interest to assess the relative importance of short/medium range skill and climate drift in determining the general level of extended-range skill. The model dependence of geographical, case-by-case and seasonal variations in skill is also important. Comprehensive investigation of these topics is beyond the scope of this paper. However, some parallel results are given from a corresponding set of ensembles produced at ECMWF using the T63 version of their spectral model.

The UKMO model integrations form part of an ongoing project to develop a methodology for extended-range prediction. Future plans in this area are discussed in more detail in section 4. From autumn 1988 we will begin to run production, real-time ensemble forecasts to coincide with the UKMO long-range forecasting (LRF) conferences. Much of the work reported in this paper is therefore directed towards developing tools for interpreting and analysing ensemble forecasts so that they might form a practical and useful input into the long-range forecasting process.

2. MODELS AND EXPERIMENTS

Both the UKMO and ECMWF experiments employed the lagged-average forecast (LAF) technique of Hoffman and Kalnay (1983). Each UKMO ensemble consisted of seven integrations initialised from consecutive UKMO operational analyses at 12 hour intervals, as in the case discussed by Murphy and Palmer (1986). The eight experiments were run at three month intervals from December 1985 to September 1987. Table 1 shows the analysis time of the latest member of each ensemble.

The ECMWF experiments, described in detail by Brankovic (1988), were nine-member ensembles with a 6 hour gap between successive (ECMWF operational) analyses. In one case, March 1987, the ensemble size was reduced to eight as the third member was unavailable for technical reasons. The latest analysis time was always 12 hours beyond that used in the corresponding UKMO experiment (Figure 1).

A global, 11-level grid-point GCM with a regular 2.5 x 3.75° latitude-longitude grid (Slingo, 1985), was used for the UKMO runs. The version of the model, which includes gravity wave drag (Palmer et al, 1986), was as described in MD, except that operationally-analysed SSTs were included in all experiments save the March 1987 case, in which climatological values were used. The SST anomalies, based on an average of the 10 days preceding the initialisation date of the first ensemble member, were held fixed relative to an SST climatology, which was updated every five days during the forecasts. The ECMWF experiments used their operational spectral model at T63 resolution, incorporating an envelope orography (Tibaldi,1986). For the December 1985 and March 1986 cases a 16-level version was used, but the remaining experiments used a 19-level version. After June 1986, gravity wave drag was incorporated. All integrations used ECMWF operationally-analysed SSTs, held constant throughout.

In the following section the experiments are verified as 30 day forecasts, although the UKMO integrations were actually 40 days long. The verifying UKMO observed data is valid at 00Z, which complicates the assessment of the

ECMWF integrations, whose archived forecast fields are all valid at 12Z. In practice each ECMWF field was verified against the observed data for the following midnight (Figure 1). This is liable to penalise the ECMWF scores, relative to the UKMO results. However, the effect may be offset by the later analysis time of the most recent integration. For a time-averaging period of 10 days, to which all of the following results refer, the influence of these factors on the results should be quite small.

Forecast	Initialisation date	
Dec. 1985	00Z	15.12.85
Mar. 1986	00Z	16.03.86
June 1986	00Z	15.06.86
Sept. 1986	00Z	14.09.86
Dec. 1986	00Z	14.12.86
Mar. 1987	00Z	15.03.87
June 1987	00Z	14.06.87
Sept. 1987	00Z	13.09.87

Table 1. Initialisation time of latest member of each UKMO lagged-average forecast.

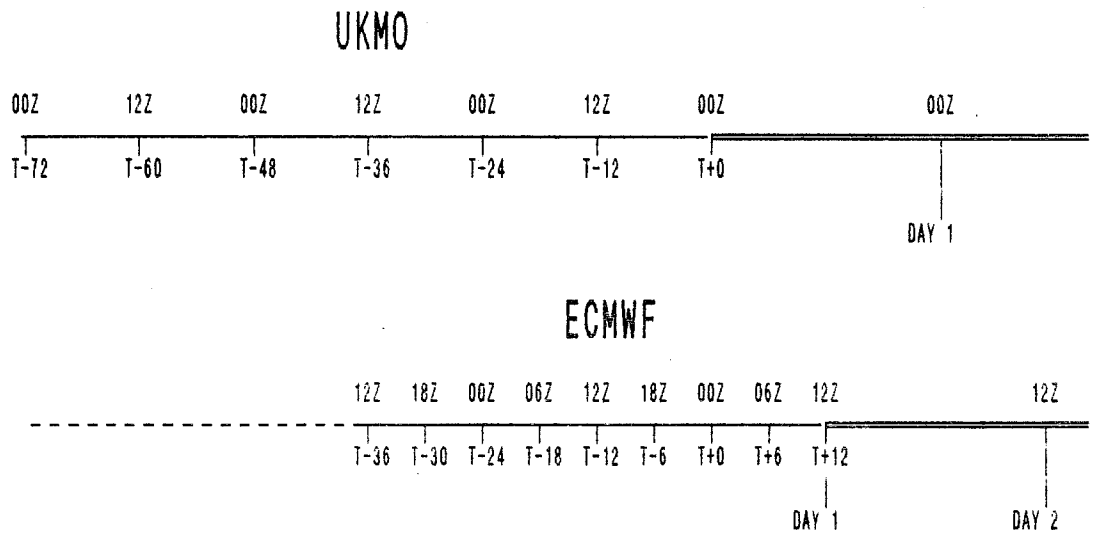


Figure 1. Analysis times of forecasts in UKMO and ECMWF ensembles relative to analysis time of last member in UKMO ensemble (T+0).

3. ENSEMBLE FORECAST RESULTS

All the results quoted in this section refer to 10-day average fields of mean sea-level pressure (MSLP). In sub-sections 3.1 and 3.2 the domain used is 30-90° N. Forecast and observed anomalies are formed relative to normals from the years 1951-80. The basis for this choice was discussed in MD. Anomaly correlation scores for both the UKMO and ECMWF experiments are given in sub-sections 3.2 and 3.3, without any empirical correction for model systematic errors (SE). For the UKMO model, some further results are included to show the effect of removing an estimate of the seasonal SE in the mean flow prior to verification. Each seasonal SE estimate was obtained from a set of 12 integrations taken from the years 1982-85 (see MD).

In the following, $\langle\rangle$ is used to denote an average over a large number of independent experiments.

3.1 Time variation of ensemble spread

The growth with time of the LAF ensemble spread reflects the increasing uncertainty in the forecast arising from analysis errors, and prediction errors incurred during the extrapolation of lagged analyses up to the start of the forecast period. If s_M is the variance of an ensemble of size M in some experiment, and w_o is the corresponding observed climate variance, then the point at which

$$\langle F \rangle = (M/(M-1))\langle s_M/w_o \rangle \qquad (1)$$

becomes equal to unity represents, in principle, the upper limit of potential predictability (Murphy, 1988). Note that normalisation by w_o *before* averaging over experiments avoids weighting the results towards seasons of high variability. In practice, $\langle F \rangle$ must be modified to account for the effects of model SE.

Let α^2 represent the mean square SE in the model's mean flow, normalised by the mean observed climate variance. Similarly, β^2 may be defined as the model's variance about its own climatology, normalised in the same way. An unbiased measure of intrinsic predictability is then given by

$$\langle F' \rangle = \langle F \rangle / \beta^2, \qquad (2)$$

since the ensemble variance is then compared to the *model* variability. To determine β^2, first note that

$$\langle w_f / w_o \rangle = \alpha^2 + \beta^2, \qquad (3)$$

where w_f is the variance of a model forecast relative to the observed climatology. Next consider the anomaly correlation, c_e, between two independent model integrations initialised, say, on the same date in two different years. It may be shown that

$$\alpha^2/\beta^2 = \langle c_e \rangle / (1 - \langle c_e \rangle), \qquad (4)$$

assuming that any difference between the observed normals used in forming the anomalies, and the true atmospheric climate appropriate to the experiment years, is small (see MD).

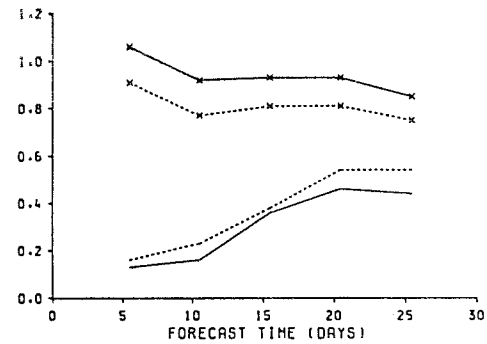


Figure 2a. Normalised mean square systematic error α^2 in the mean flow for UKMO (--------) and ECMWF (————) models for 10-day mean MSLP field over region 30-90° N. (x--x--x) and (x—x—x) show corresponding systematic errors β^2 in model variability.

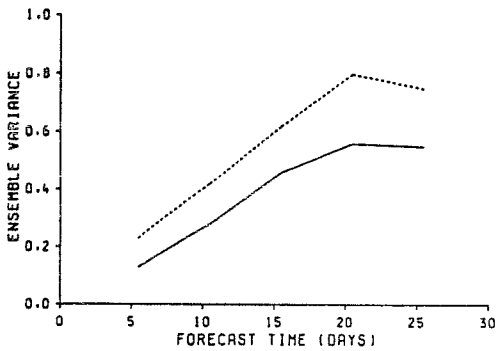


Figure 2b. Corrected ensemble variance $\langle F' \rangle$ for UKMO (--------) and ECMWF (————) models for 10-day mean MSLP field over region 30-90° N .

In practice, a mean value of c_e was found for the two experiments in a given season, by averaging the value for each pair of corresponding ensemble members. The results were then averaged over the four available pairs of experiments to obtain $\langle c_e \rangle$. Estimates of α^2 and β^2 were then deduced by calculating the mean value of w_f/w_o over all ensemble members and experiments.

Figure 2a shows the time variation of α^2 and β^2 for 10-day mean fields, for each model. Note that the time origin is taken as the analysis time of the central ensemble member for the model in question, and all ensemble members are treated as if they possessed the same analysis time. The climate drift in the mean flow grows to significant proportions in both models, reaching just over 50% of the observed climate variance by 30 days for the UKMO model, and just under 50% for the ECMWF model. At this stage the drift is apparently levelling off in both cases. Both models also underestimate low frequency variability, but the discrepancy is greater for the UKMO model. This is an important factor, given the requirement that a model should be able to simulate the full range of possible atmospheric evolutions, in order to produce realistic probabilistic predictions at extended range. For practical forecasting purposes, the seasonal and spatial variation of α^2 and β^2 is important. To determine this accurately, a very large number of experiments is required.

On average the UKMO ensembles show a greater value of the corrected spread $\langle F' \rangle$ at days 1-10 (Figure 2b), but this is probably due simply to the greater spread of the initial analyses. The subsequent rate of growth of $\langle F' \rangle$ is similar in both models. However, both models may underestimate the idealised rate of divergence occurring in a hypothetical ensemble of identical real atmospheres, evolving from a closely-grouped set of initial states (Lorenz, 1982). Therefore, whilst the ensemble spread remains well short of saturation level for both models at a range of one month, the implied degree of remaining potential predictability may be optimistic.

3.2 Large-scale forecast skill

Figure 3 shows the ensemble-mean forecast anomaly correlation score in each experiment for both models, for the domain 30-90° N. Note that for this and subsequent figures the time origin reverts to that shown in Figure 1.

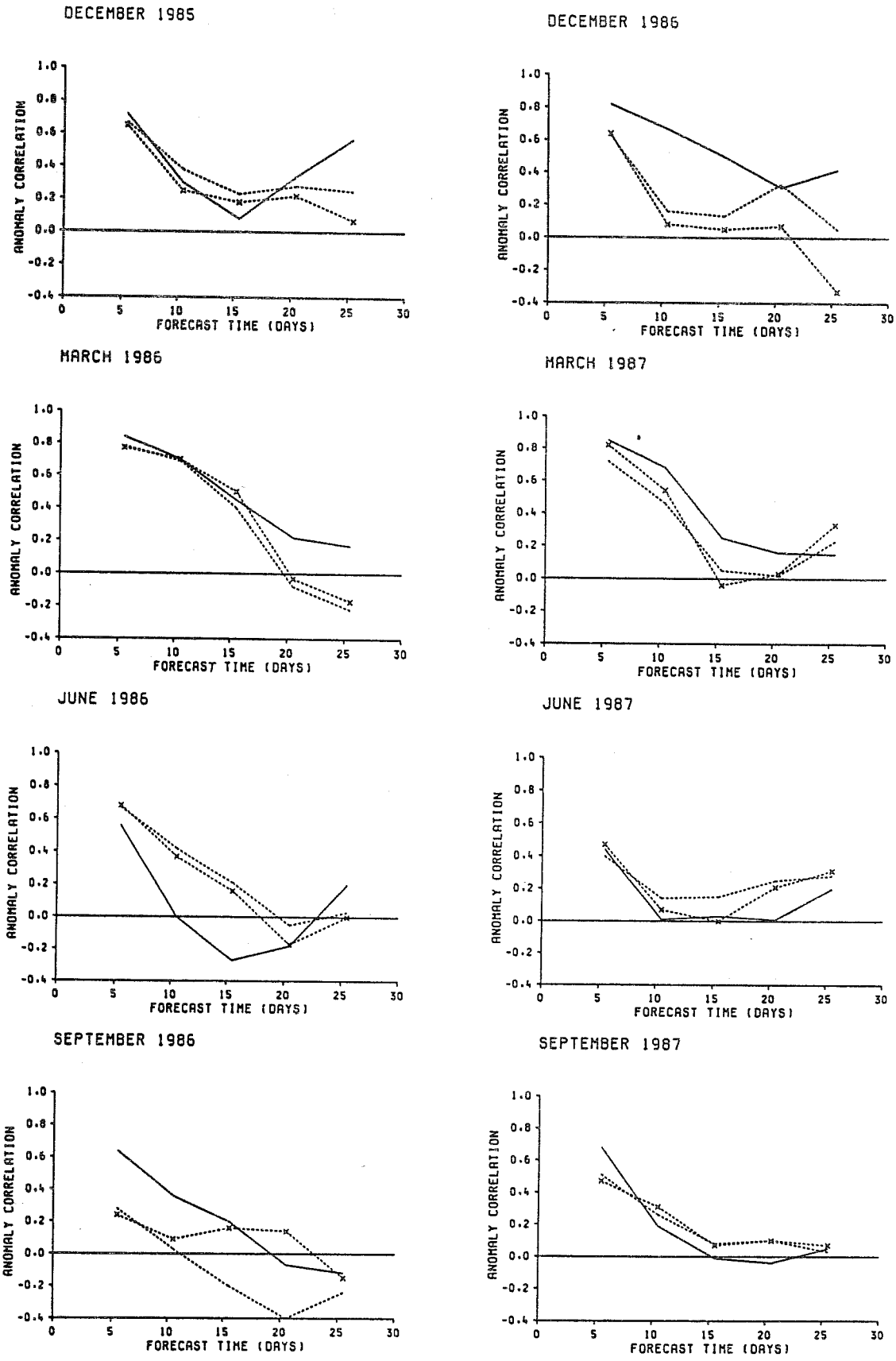


Figure 3. Skill of ensemble mean forecasts for 10-day mean MSLP field over region 30-90° N for each experiment for ECMWF model (————), UKMO model (-------) and UKMO model with SE removed (x--x--x).

There are no cases of high skill at extended-range, although the ECMWF forecast score remains above 0.3 throughout in December 1986. The ECMWF model is consistently more skilful in December 1986 and March 1987, and is vastly superior at medium range in September 1986. However, the UKMO model is better in the two summer cases, particularly June 1986. Both models tend to be more skilful at medium range in winter and spring, compared to summer and autumn. For the UKMO model this feature was clearly apparent in the runs studied by MD.

The effect of SE removal in the UKMO model is variable, being large and positive in the September 1986 case, but negative at all time levels in the two winter experiments. Overall the average effect is very small for the 10-day mean fields. (The mean effect of applying the correction to independent forecasts was greater in MD). In the current experiments, nevertheless, the average score for the mean of days 1-30 is improved from 0.26 to 0.35 by removal of the SE. For the longer period the SE will typically form a greater proportion of the uncorrected forecast climate variance.

The distribution of individual forecast scores in each experiment is shown in Figure 4, for days 1-10, 6-15 and 11-20, for the ECMWF and uncorrected UKMO models. Some corresponding 500mb geopotential height results are shown for the ECMWF model in Hollingsworth et al (1987). At days 1-10, the UKMO model tends to show a greater trend across the analysis times. This is particularly apparent in the two autumn experiments. Whilst in part due simply to the greater spread of analysis times, this also reflects a more rapid decay of medium-range skill. By days 11-20 there is no evidence of any systematic variation of skill with analysis time for either model. As remarked by Hollingsworth et al (1987), the distribution of scores generally becomes wider at the later forecast times. This is a natural result of the spreading of the ensemble. However, the apparent effect may be accentuated by the non-linearity of the anomaly correlation statistic. Note that there is little evidence of clustering in the individual forecast scores, although this does not necessarily rule out such behaviour in the distribution of forecast fields (Murphy and Palmer, 1986).

In cases where one model was noted to be superior from Figure 3, the difference is found to be quite consistent across the individual ensemble

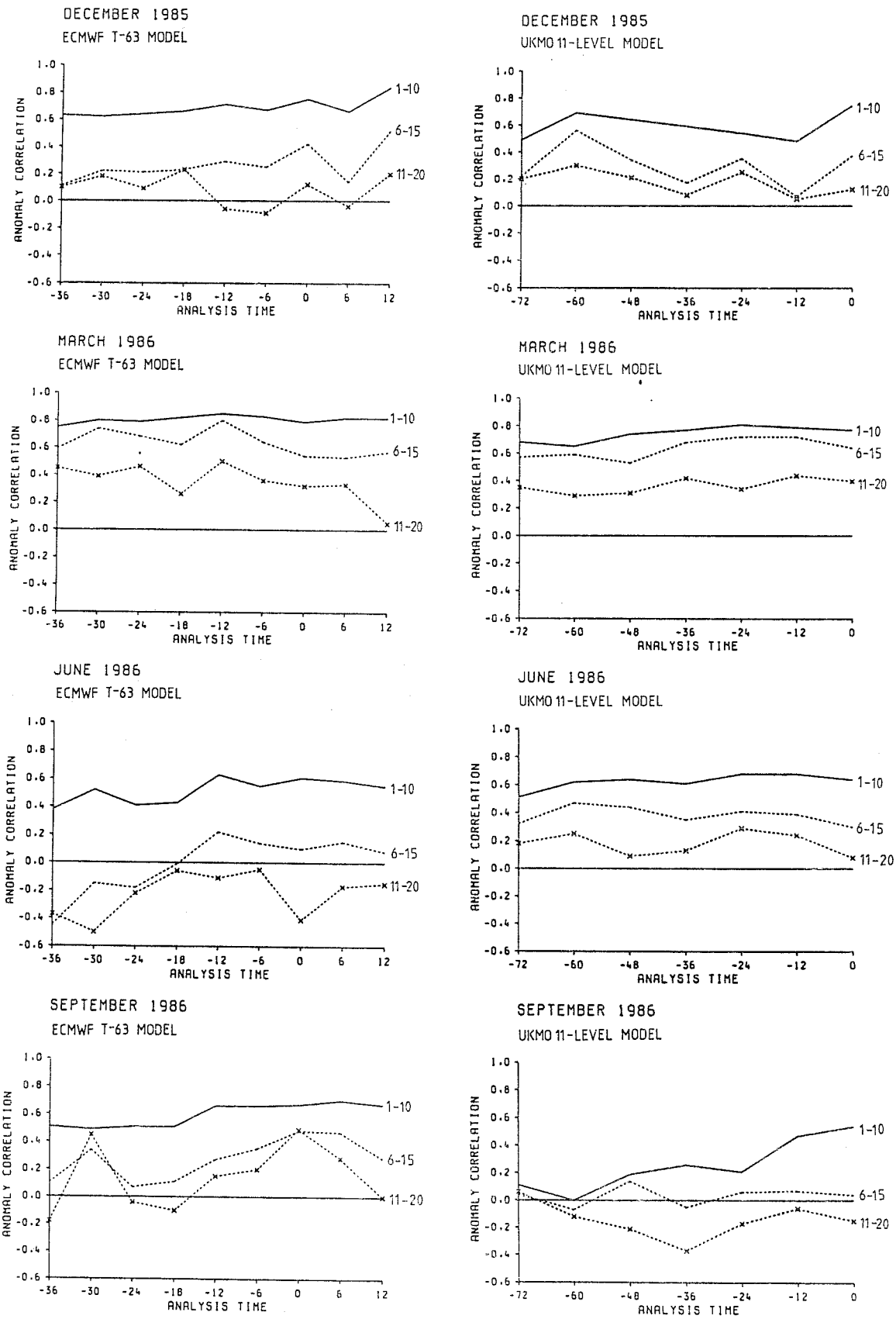


Figure 4. Skill of ensemble members for 10-day mean MSLP field over region 30-90° N for each experiment for ECMWF model and uncorrected UKMO model. The analysis time of each forecast (in hours) is given relative to the analysis time of last member of UKMO ensemble.

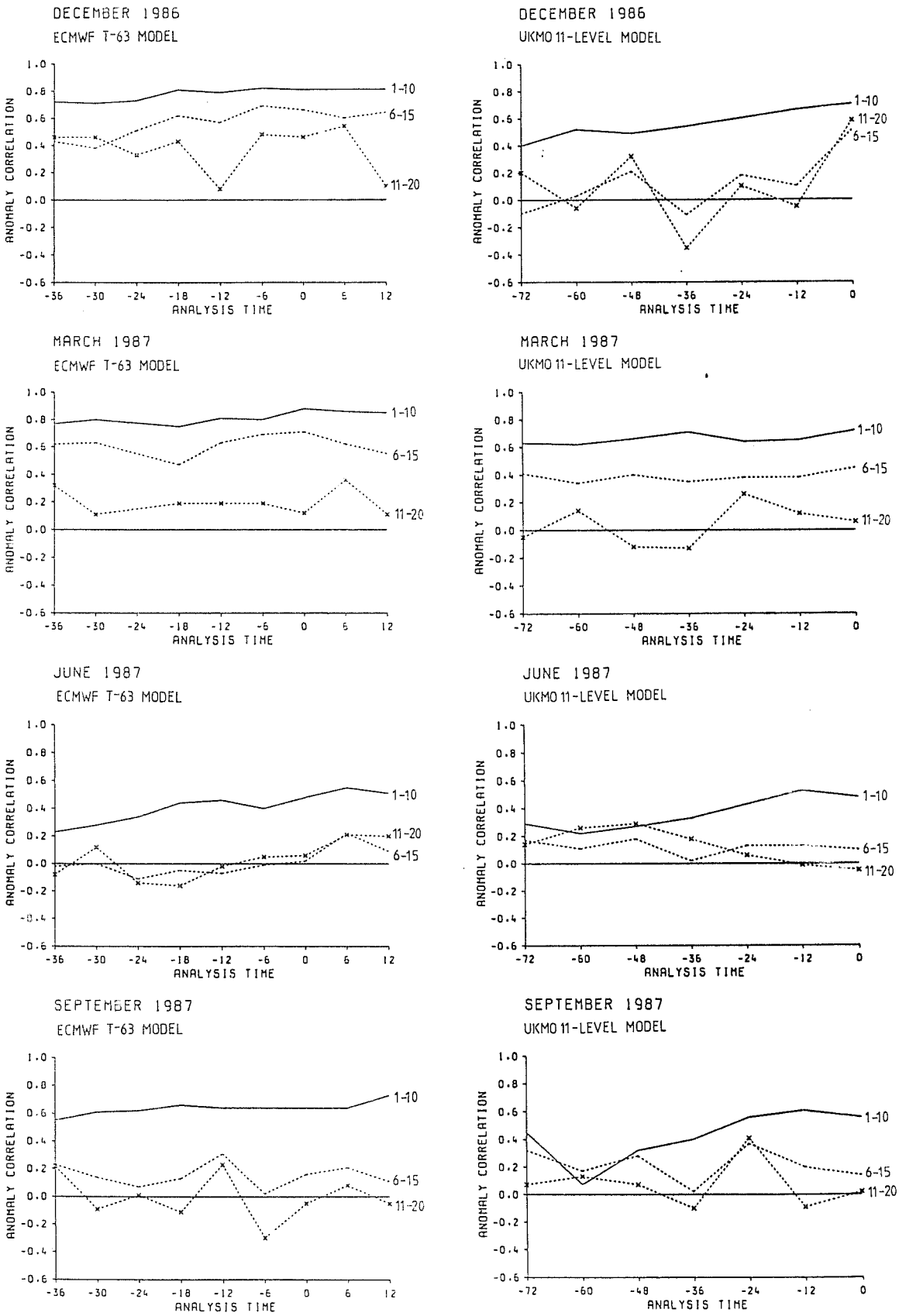


Figure 4. Continued.

members. For example, at days 6-15 every ECMWF integration shows higher skill than every UKMO integration in the December 1986 and March 1987 cases. However, in the June 1986 experiment every UKMO integration shows positive skill at days 11-20, whereas every ECMWF integration shows negative skill. In MD it was suggested that monthly predictions of significantly above-average skill might be possible on around 25% of occasions. An important topic for future consideration is whether different models will perform unusually well on the same occasions. The current results suggest, tentatively, that this may not be the case. If the skill of extended-range predictions depends strongly on medium-range skill, this would be one possible reason why case-by-case variations in predictability should be model dependent.

Figure 5 shows that the average skill of the ECMWF ensemble-mean forecast is considerably larger than that of the uncorrected UKMO model at days 1-10. This confirms the additional medium-range skill of the higher resolution model alluded to above. However, the difference is accentuated by the use of older analyses in the UKMO experiments. Each pair of ensembles have four analysis times in common. The average individual forecast score for these cases is 0.64(0.59) for the ECMWF(UKMO) model, compared with 0.66(0.53) for the average score over all ensemble members. The ECMWF score remains slightly superior at all subsequent time levels. This may reflect the benefit of maximising medium-range skill. Alternatively, the fact that the ECMWF model possesses slightly smaller climate drift, and simulates the observed variability somewhat more realistically (Figure 2), may be important.

Following on from the discussion of Figures 3 and 4, the divergence between models in a given experiment may yield additional information on predictability. This can be investigated by considering a single 'super-ensemble' containing all individual integrations from both models. In general the impact of such a super-ensemble depends on whether the two constituent ensemble distributions show significant differences. Alternatively, the super-ensemble distribution may be indistinguishable from a single ensemble of increased size from one of the models. If the two models do contribute independent information of approximately equal quality, the skill of the super-ensemble mean should exceed the average

skill of the constituent ensemble means. Figure 5 shows that this is indeed the case at all time levels. In fact, up to days 11-20, the super-ensemble score exceeds that of both models taken individually. This suggests that such an approach may be worthwhile, although it remains to be shown that the skill of the super-ensemble would outstrip that of an ensemble of increased size from a single model.

Forecast period (days)	Mean individual forecast score	Ensemble-mean forecast score	Last individual forecast score
1-10	0.53(0.66)	0.58(0.69)	0.64(0.72)
6-15	0.27(0.31)	0.32(0.36)	0.32(0.36)
11-20	0.12(0.12)	0.13(0.15)	0.13(0.06)
16-25	0.05(0.07)	0.06(0.09)	0.05(0.04)
21-30	0.04(0.15)	0.05(0.21)	0.12(0.24)

Table 2. Mean anomaly correlations over all experiments for forecasts of 10-day mean MSLP over 30-90° N for uncorrected UKMO model and ECMWF model (in brackets). The score for the ensemble mean is compared to the average score for all individual ensemble members, and to that for the individual forecast from the most recent analysis.

Returning to separate consideration of the two models, Table 2 shows the effect of ensemble averaging on the practical forecast skill in each case. Compared with the average individual forecast score, the ensemble mean gives improved skill for both models at each time level, but the size of the increase is fairly small. The ensemble-mean score is lower than that of the integration from the most recent analysis at days 1-10, and equal to it at days 6-15. Production of an optimally weighted ensemble-mean may yield further benefit at medium range (Dalcher et al, 1988). However, attempts along these lines with the UKMO model (not shown) suggest that in terms of anomaly correlation, the impact is very small for time-mean fields.

It is instructive to compare the practical impact of ensemble averaging with an idealised upper limit obtained using a perfect model approach. Figure 6 shows some appropriate results for the UKMO model with SE removed. The perfect model approach was implemented by producing an extra integration from the analysis time 12 hours later than that of the latest member of each UKMO ensemble. The ensemble was then 'verified' against this integration in each case. Figure 6 shows that the gap between the

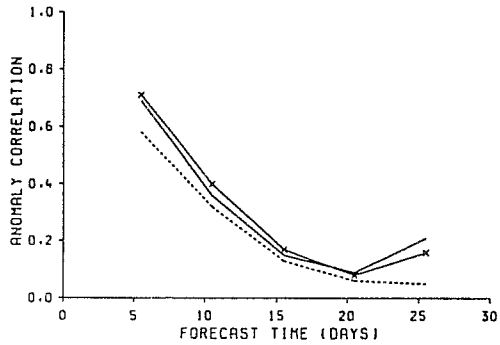


Figure 5. Average skill of ensemble-mean forecast for 10-day mean MSLP over region 30-90° N for uncorrected UKMO (------) and ECMWF (————) models. (x—x—x) shows skill for 'super-ensemble mean' created by averaging all individual forecasts from both models.

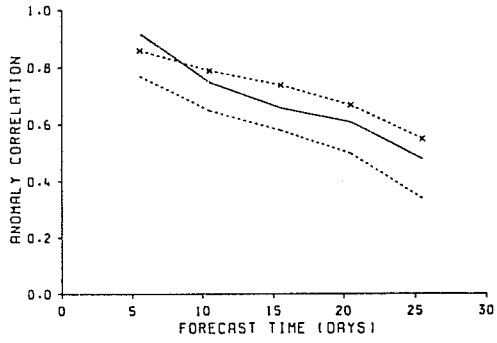


Figure 6. Average 'perfect model' skill for 10-day mean MSLP field over region 30-90° N for ensemble-mean forecast (x--x--x) , and individual forecast from the most recent analysis (————) , for UKMO experiments with SE removed. (--------) shows the mean score for all individual ensemble members.

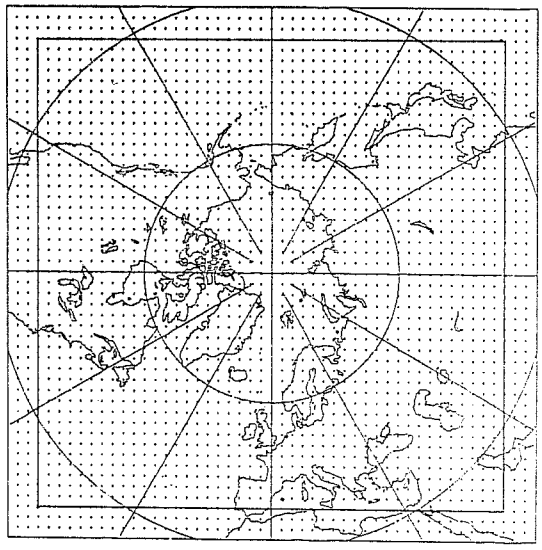


Figure 7. Polar stereographic 51x51 grid used for analysis of local ensemble spread and skill.

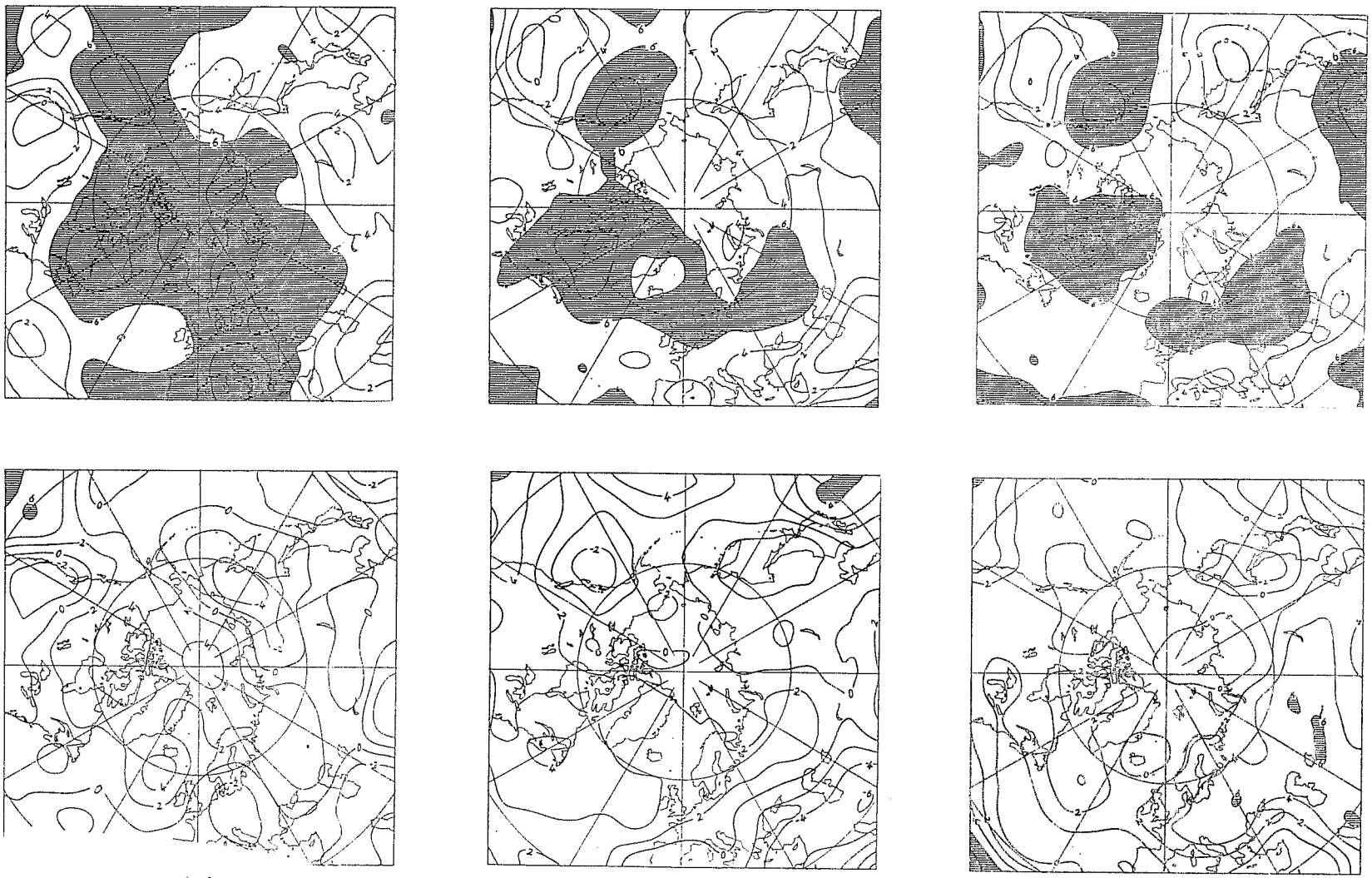
ensemble mean and individual forecast scores is considerably enhanced compared with the practical results of Table 2. The perfect model gap is in close agreement with the theoretical prediction of equation (9) in Murphy (1988). The most recent individual forecast is inferior to the ensemble mean at days 6-15, but remains superior at days 1-10. However, this may be because the spread of forecasts at time zero is overestimated for perfect model purposes, due to the presence of external prediction errors in the lagged ensemble members. Despite this, optimal weighting using the minimum error variance criterion also has a positive effect. For example, the perfect model anomaly correlation scores of the weighted ensemble mean at days 6-15 and 11-20 are 0.82 and 0.78 respectively, compared with 0.79 and 0.74 for the unweighted ensemble mean.

These results demonstrate the large gulf which currently exists between the practical and theoretical impact of ensemble forecasting. Even for the more skilful ECMWF model the same comment applies. Clearly there remains great scope for improvement, in this sense, through the development of more skilful forecast models.

3.3 Local forecast skill and spread

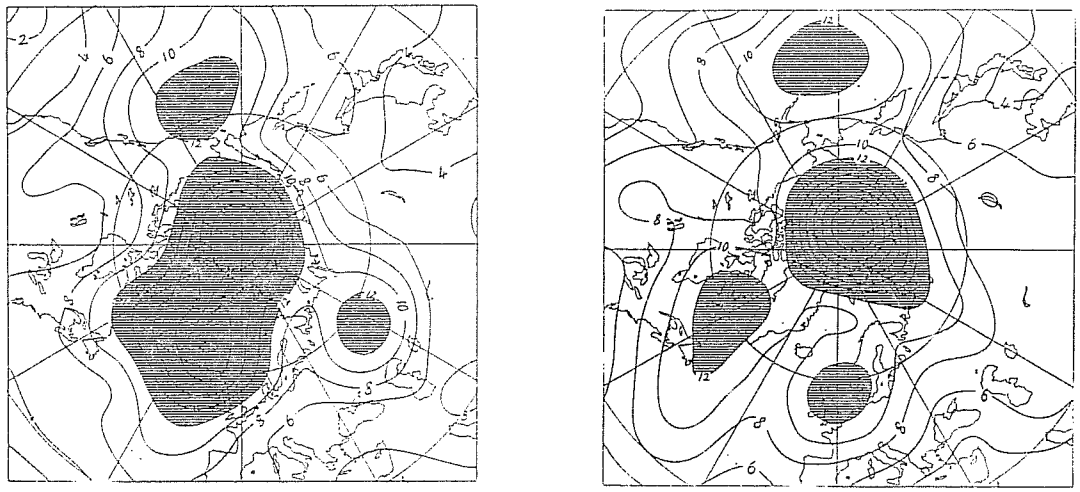
The present experiments do not include a case showing exceptional skill on the hemispheric scale at extended range (see previous section). Although occasional examples of such cases have been captured (Hollingsworth et al, 1987), the prospects for more regular extraction of useful information at extended range may depend crucially on the ability to identify locally skilful areas in advance (eg Molteni et al, 1986; Murphy, 1988).

In this section local forecast anomaly correlation, and its relationship to local ensemble spread, is considered using the polar stereographic grid of 51x51 points shown in Figure 7. Spread and skill values were calculated over boxes of dimension 7x7 points, centred on each point within the inset region. Figure 8 shows, for both models, maps of the local skill of the ensemble-mean forecast, averaged over all experiments for days 1-10 and 11-20. Considerable geographic variations are apparent, values ranging from below zero to 0.8 or greater at days 1-10. For this period the mean skill of the ECMWF model is superior (Table 3). The fractional area for which the score exceeds the 0.6 level, often taken as a baseline of useful skill



(a) (b) (c)

Figure 8. Geographical distribution of local area anomaly correlation (x10) averaged over all experiments for (a) ECMWF model, (b) UKMO model and (c) UKMO model with SE removed. The upper panel shows days 1-10 and the lower panel days 11-20. Areas with correlation greater than 0.6 are shaded.



(a) (b)

Figure 9. Observed local r.m.s. anomaly magnitude (x10) for 10-day mean MSLP for (a) days 1-10 and (b) days 11-20. Local values were normalised by the space averaged value in each case. The normalised values were then averaged over all experiments. Areas where the mean value exceeds 1.2 are shaded.

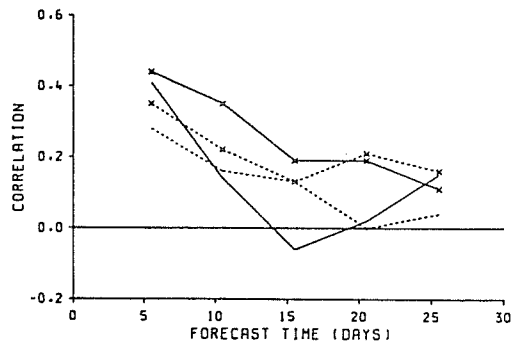


Figure 10. Average point-by-point correlation, ρ_E, between experimental variations in local ensemble spread and skill for uncorrected UKMO (------) and ECMWF (————) models. (x--x--x) and (x—x—x) show corresponding correlations between skill and the square root of the mean climate variance among individual ensemble members.

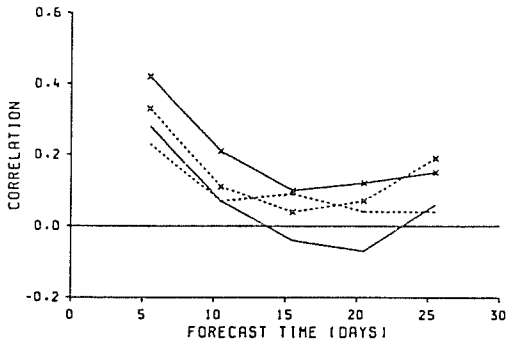


Figure 11. Average correlation, ρ_G, between geographical variations in local ensemble spread and skill for uncorrected UKMO (------) and ECMWF (————) models. (x--x--x) and (x—x—x) are corresponding correlations between skill and the square root of the mean climate variance among individual ensemble members.

Forecast period (days)	S K I L L B I N				
	1 (0.8)	2 (0.4)	3 (0.0)	4 (-0.4)	5 (-0.8)
1-10	.88(.81) 45%	.55(.42) 30%	.02(.02) 16%	-.49(-.36) 8%	-.78(-.68) 2%
11-20	.86(.75) 12%	.58(.39) 28%	.01(.01) 30%	-.59(-.37) 21%	-.83(-.69) 5%
21-30	.89(.75) 13%	.65(.38) 26%	-.04(.00) 31%	-.66(-.38) 23%	-.86(-.74) 9%

Table 4. Average local ensemble-mean forecast anomaly correlation for 10-day mean MSLP, observed at points where the mean score for individual ensemble members falls within a given range (average value in brackets), for the UKMO model with SE removed. Each skill bin is of width 0.4, centred on the value shown. The area-weighted relative frequency (%) associated with each bin is also given.

Despite this drawback, in all the experiments there are areas, even at days 11-20 and beyond, where the local skill reaches a useful level. Table 4 shows that the wide distribution in skill is partly due to the local effect of ensemble averaging. Where the average score for the individual ensemble members is positive, the ensemble mean shows a significantly improved score. However, the reverse effect is observed in areas of negative skill. This polarisation increases the premium on the *a priori* prediction of skill. For example, the ensemble-mean score exceeds 0.6 over 32% of the total area on average at days 11-20, compared with 12% for the average individual forecast score. Clearly the dependence of such results on the choice of skill score is an important question. Further work is in progress in this regard.

Local ensemble spread was measured using the anomaly correlation between the ensemble mean, and all the individual ensemble members taken together (ie $(a_M)^{\frac{1}{2}}$, where a_M is as defined in Murphy (1988)). The correspondence between the spread and the skill of the ensemble mean can be measured in different ways. One method is to calculate a coefficient, ρ_E, at each grid point, representing the correlation between spread and skill measured over the available experiments. Figure 10 shows the time variation of the space-averaged value of ρ_E for both models. The correlations are greatest for

(Hollingsworth et al, 1980), is markedly greater than in the UKMO model. The general pattern of skill variation is broadly similar for both models. For example, both show maxima centred near either coast of North America, although the UKMO model does not show a maximum near Italy. Removal of the SE improves the mean UKMO model score slightly (Table 3), but does not greatly alter the pattern at days 1-10.

Forecast period (days)	ECMWF model	UKMO model	UKMO model SE removed
1-10	0.56	0.47	0.51
11-20	0.09	0.12	0.15
21-30	0.12	-0.02	0.05

Table 3. Average local ensemble-mean forecast anomaly correlation for 10-day mean MSLP.

For days 11-20, the average skill is much lower. Substantial geographic variations are still observed, and the patterns for the two models show more divergence. However, there is an element of persistence from days 1-10 in both cases. Whilst the overall effect of SE removal is small in the UKMO model, large changes are observed in certain locations. Note particularly the large increase in skill over much of Central Asia.

If extended-range skill is linked to low frequency variability, preferred regions for high local skill may correspond to preferred regions for high variability (Blackmon et al, 1977). The mean observed variability over the current set of experiments is shown in Figure 9 for days 1-10 and 11-20. This was obtained by calculating, for each experiment, the local observed anomaly magnitude at each point, normalised by the space-averaged r.m.s. value for the whole domain. The results were then averaged over all eight cases. In this way equal weight was given to each season. For days 1-10 there is reasonable correspondence with the skill maps of Figure 8, more particularly for the ECMWF model, which has a skill maximum near the pole. This feature is retained at days 11-20, in common with the persistent area of high variability. However, as in the case of the UKMO model, there is little correspondence in other areas. This suggests that the models generally fail to reproduce changes in the observed low frequency patterns beyond the medium range.

days 1-10, but remain positive throughout for the UKMO model. Removal of the SE does not have a consistent positive effect on the correlation levels for this model (Table 5). Interestingly, as with the skill itself, the ECMWF model gives the largest value at days 1-10. This is to be expected, given that its short/medium range intrinsic error growth rate is closer to its external error growth rate, as compared with the UKMO model. Figure 10 also shows the average value of the corresponding correlation, r_E, between skill and the local value of $(\hat{w}_F)^{\frac{1}{2}}$, where w_F is the climate variance of an individual forecast, and $\hat{}$ denotes the average value over all ensemble members. This is generally similar to, or a little greater than, the mean value of ρ_E. However, as shown in Table 5, this correlation is surprisingly reduced when the SE in the UKMO model is removed. With only eight experiments, the results are naturally subject to considerable sampling error.

Forecast period (days)	ρ_E	r_E
1-10	0.28(0.28)	0.35(0.27)
11-20	0.13(0.08)	0.13(0.11)
21-30	0.04(0.10)	0.16(-0.07)

Table 5. Average point-by-point correlation over experiments between local ensemble-mean forecast skill and ensemble spread (ρ_E), for 10-day mean MSLP. Also shown is correlation r_E between skill and the square root of the mean climate variance of individual ensemble members. Values are for the uncorrected UKMO model, compared with values (bracketed) when the SE is removed.

Forecast period (days)	ρ_G	r_G	$(r_G)_{IND}$
1-10	0.23(0.24)	0.33(0.18)	(0.14)
11-20	0.09(0.06)	0.04(-0.09)	(-0.08)
21-30	0.04(0.08)	0.19(-0.05)	(-0.02)

Table 6. As Table 5 for ρ_G and r_G, correlations between geographic variations in the corresponding quantities, averaged over all experiments. Also shown is $(r_G)_{IND}$, the average correlation between geographic variations in the local skill of the individual forecast from the most recent analysis, and its local forecast anomaly magnitude, with SE removed.

The climate variances used in Figure 10 and Table 5 are not normalised, and therefore contain the seasonal cycle. This may partially explain their relatively good performance as skill predictors in comparison to the ensemble spread, which shows no evidence of significant seasonal dependence. To test this, Figure 11 shows average values of p_G, the correlation between geographic variations in skill and spread in a given experiment. The corresponding average of r_G, the correlation between skill and $(\hat{W}_F)^{\frac{1}{2}}$, is also shown. Compared with Figure 10, levels of correlation are generally a little lower, more notably for the ECMWF model. Although geographic correlation excludes seasonal effects, the mean value of r_G still generally exceeds that of p_G, for both models. However, removal of SE reverses this result for the UKMO model, and r_G becomes negative beyond days 11-20 (Table 6). If $(\hat{W}_F)^{\frac{1}{2}}$ is genuinely comparable to ensemble spread as a skill predictor, it might be thought that skill predictions of a similar quality can be obtained using a single forecast. Table 6 shows the correlation $(r_G)_{IND}$ between the anomaly magnitude of the most recent UKMO ensemble member with SE removed, and its own local skill. This is a little lower than r_G for days 1-10, and negative thereafter. For 500mb geopotential height (not shown), the superiority of p_G and r_G over $(r_G)_{IND}$ is considerably more marked.

It is necessary to consider how a given level of ensemble spread/skill correlation translates into a practical skill prediction facility. Tables 7 to 9 show spread/skill contingency tables for the UKMO model with SE removed, for days 1-10, 6-15 and 11-20. The spread bins were chosen to give an approximately equiprobable distribution at days 21-30. All pairs of spread/skill values, for all points and experiments, were included together in the tables. For days 1-10 the overall spread/skill correlation is 0.26. The probability of obtaining a skill score in excess of 0.6 (skill bin 1) is 57%. However, for spread bin 1, containing 37% of observations, the probability rises to 75%. For spread bins 3,4 and 5, containing 28% of observations, the probability drops to 42%. Thus a discriminating approach may yield useful probabilistic statements of likely forecast skill, despite a rather modest overall level of spread/skill correlation. This also clearly apparent at days 6-15 (Table 8). The overall probability of skill bin 1, although still substantial, drops to

SPREAD BIN	SKILL BIN					
	1	2	3	4	5	REL FREQ
1	.75	.11	.06	.04	.03	.37
2	.53	.21	.12	.08	.06	.33
3	.46	.26	.14	.10	.05	.17
4	.36	.31	.14	.12	.07	.09
5	.31	.35	.13	.12	.08	.02
MEAN	.57	.19	.09	.08	.05	

Table 7. Contingency table of probabilities for local ensemble-mean forecast anomaly correlation as a function of $(a_M)^{\frac{1}{2}}$, local ensemble spread, for UKMO model with SE removed for MSLP days 1-10. Skill bins are as in Table 4. Spread bins are : $1-(a_M)^{\frac{1}{2}} \geq 0.9$; $2-0.9 > (a_M)^{\frac{1}{2}} \geq 0.78$; $3-0.78 > (a_M)^{\frac{1}{2}} \geq 0.65$; $4-0.65 > (a_M)^{\frac{1}{2}} \geq 0.50$; $5-0.50 > (a_M)^{\frac{1}{2}}$. Also shown are the overall skill probabilities and the relative frequency of each spread bin.

SPREAD BIN	SKILL BIN					
	1	2	3	4	5	REL FREQ
1	.59	.12	.09	.08	.13	.25
2	.44	.19	.12	.14	.11	.27
3	.33	.23	.15	.17	.11	.20
4	.25	.27	.21	.16	.11	.15
5	.14	.24	.23	.19	.19	.14
MEAN	.39	.20	.14	.14	.13	

Table 8. As Table 7 for days 6-15.

SPREAD BIN	SKILL BIN					
	1	2	3	4	5	REL FREQ
1	.48	.15	.10	.10	.17	.12
2	.40	.16	.12	.15	.17	.25
3	.26	.18	.15	.22	.19	.22
4	.25	.20	.20	.21	.14	.20
5	.26	.27	.20	.16	.11	.19
MEAN	.32	.19	.15	.17	.15	

Table 9. As Table 7 for days 11-20.

39%. However, for spread bin 1 (25% of observations) the probability is 59%, which is greater than the overall value for days 1-10. Encouragingly, the probability of skill bin 1 drops monotonically through spread bins 1-5. Even at days 11-20, when the overall correlation is only 0.05, the probability of skill bin 1 rises from 32% to 48% for the 12% of observations in spread bin 1. Note, however, that the probability of skill bin 5 also increases somewhat for small spread. This is also a feature of the ECMWF results (not shown). It may be explained by the fact that areas of small (correlation based) spread tend to equate to areas of where the ensemble-mean forecast anomaly is of large magnitude, and therefore has more scope to be spectacularly wrong at extended range.

4. FUTURE PLANS AND PROSPECTS

4.1 Impact of new computer

The UKMO is in the process of replacing its Cyber 205 computer by a four processor ETA^{10} system. The new computer will be available sometime in the third quarter of 1988 and will be 8 times more powerful than the Cyber 205. This increase in computer power will enable us to run real-time ensembles of extended range forecasts for consideration at our two-weekly long-range forecasting conferences. Each ensemble will consist of at least 7 forecasts, with initial dates separated by 12 hours as in the UKMO experiments described above. The use of a 6 hourly separation between forecast data times, as used in the ECMWF experiments, is unlikely because of the concentrated demands on computer time it implies. Other integrations, out to a month ahead, will be necessary to provide an estimate of the model's systematic error and climatology. It is hoped that these integrations will form a daily time series, since this is a natural way of generating a large sample of ensembles and will facilitate an investigation of sub-seasonal fluctuations in model skill.

Some of this increase in computer power will also be used to run higher resolution models. Initially we will continue to use the 11-level GCM, but other models will be available on the ETA^{10}. Two models currently being evaluated are a 2x3° version of the 11-level GCM and the global version of the UKMO 15-level numerical weather prediction model which uses a 1.5x 1.875° horizontal grid (Bell and Dickinson, 1987). Both of these models

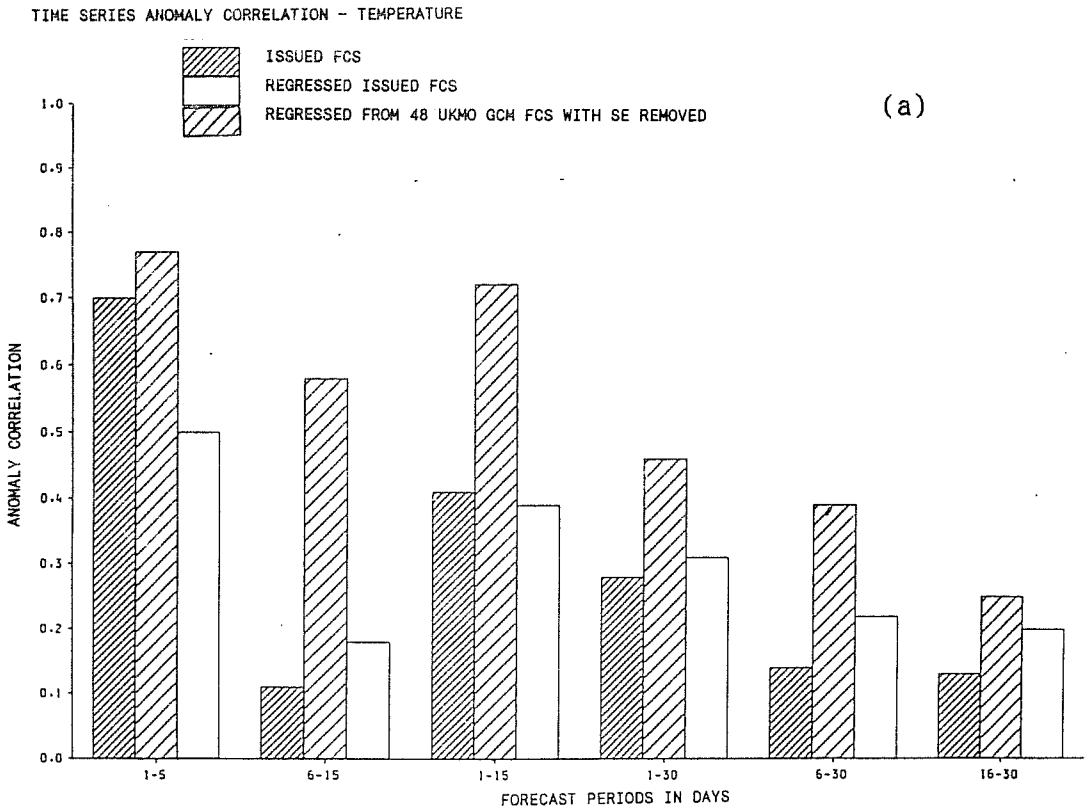
will be available with 20 levels on the ETA[10]. However, the choice of model will still be constrained by the available computer power and by the fact that operational short-range forecasting takes precedence, and so we will continue to use models of a lower resolution than the operational global forecast model, which on the ETA[10] will have a horizontal resolution of 1x1.5° and 20 levels.

4.2 Dynamical forecasts and practical long-range forecasting

Although the assessment of numerical models in terms of the skill of MSLP or 500mb height forecasts is a natural part of the development of dynamical extended-range prediction, practical long-range forecasting ultimately demands the production of rainfall and temperature forecasts. It is therefore important to see if dynamical model output can be used with any degree of success to produce forecasts of these surface weather variables at extended range. It is equally important, if dynamical models are to be used in an operational system, to see how the skill of dynamical forecasting methods compares with more established long-range forecasting techniques.

The UKMO long-range forecasting system uses a mixture of statistically based techniques and medium-range dynamical products (Folland and Woodcock, 1986). Before autumn 1987 temperature and rainfall forecasts were derived through a combination of subjective and analogue techniques from forecast MSLP fields. Now objective methods based on multiple regression equations are used, although experience shows that some subjective modification is still needed. The regression equations have been produced for each half month and for each of the 10 areas into which the UK is divided. They relate the two surface weather variables to functions of MSLP, averaged over 5, 10 and 15 days. The functions are the zonal and meridional geostrophic wind components, the geostrophic vorticity and the mean pressure. In addition, for temperature, sea surface temperature anomalies upwind and 500mb-1000mb thickness are included in the regression equations.

The mean multiple correlation obtained on dependent data for a 15-day mean is .78 for rainfall and .91 for temperature when thickness is included. Without thickness the temperature multiple correlation reduces to .77. Rainfall is most closely related to vorticity and mean pressure; for



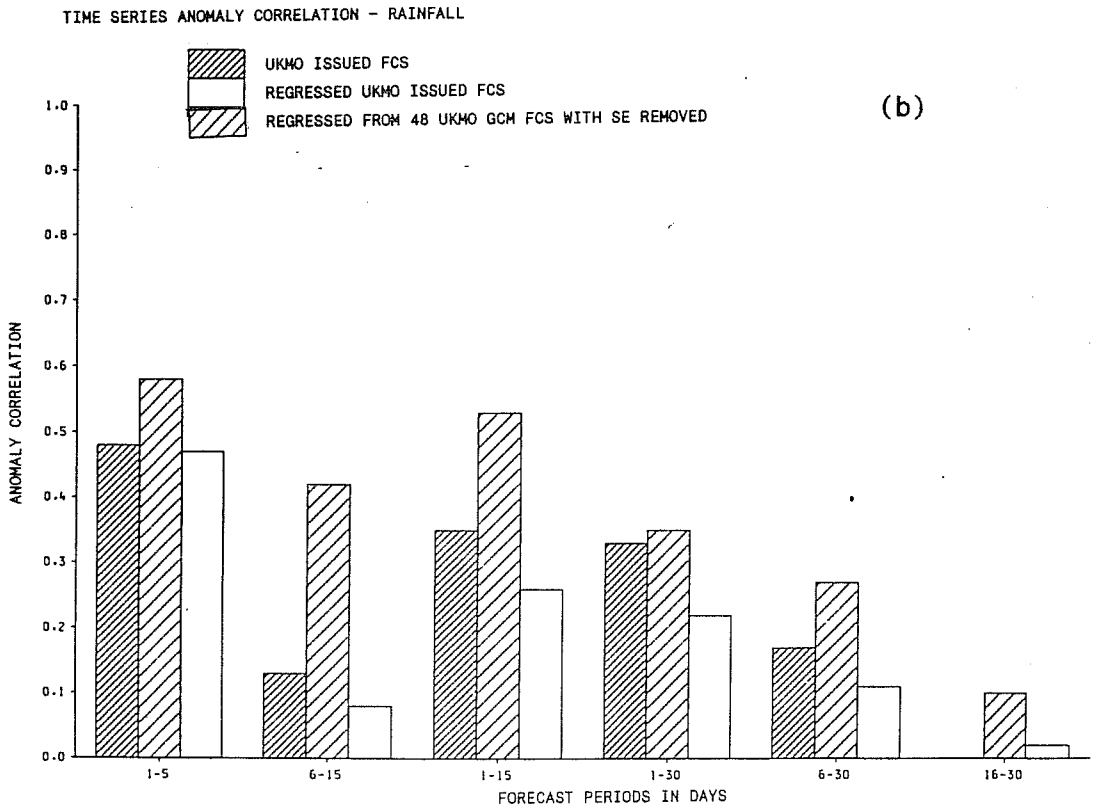


Figure 12. Forecast time series anomaly correlation for (a) screen temperature and (b) rainfall averaged over the 10 UK districts and all forecasts. ▨ as issued at UKMO LRF conferences from Feb 1983 - Jan 1987; □ applying regression equations to MSLP fields derived at LRF conferences from Feb 1983 - Jan 1987; ▨ applying regression equations to MSLP (and 500mb - 1000mb thickness) fields for 48 UKMO GCM forecasts with SE removed covering period April 1982 - Dec 1985.

temperature the largest partial correlation is with the 500mb - 1000mb thickness field.

Figures 12a and 12b show the skill obtained when these regression equations are retrospectively applied to (a) 96 MSLP forecasts produced by the LRF conferences between Feb 1983 and Jan 1987, and (b) to the the MSLP and 500mb-1000mb thickness fields derived from a set of 48 11-level GCM integrations covering the period April 1982 to Oct 1985. The skill is measured in terms of the time series correlation score, calculated from the mean of the individual scores for each of the 10 UK forecast areas. The skill of the MSLP fields from these forecasts in terms of time series correlation has already been discussed in MD. The dynamical forecasts are the same set as used in Section 3 to estimate the model's seasonal SE, although these integrations have this same SE removed. Figures 12a and 12b also give the the skill of the temperature and rainfall forecasts actually issued at the conferences. It can be seen that for all of the forecast periods, the best rainfall and temperature forecasts were produced by the regressed dynamical forecasts, although the statistical significance of these improvements is difficult to assess because of the different selection of cases. The differences between the three systems are more marked for the temperature forecasts, partly because the dynamical products forecast 500mb - 1000mb thickness and the statistical products do not.

These results are consistent with those for MSLP reported in MD. They hold out a hope of a significant improvement in the skill of the issued long-range temperature and rainfall forecasts, especially for the period 6-15 days ahead, once extended-range dynamical forecasts become a regular feature of the conferences.

5. CONCLUSIONS

In this paper we have examined 10-day mean MSLP fields from two corresponding sets of 8 LAF ensemble experiments run at quarterly intervals during the period Dec 1985 to Sept 1987. These form part of a continuing series of integrations being run at ECMWF with a T63 version of their spectral model and at the UKMO with an 11-level GCM.

For both models the mean skill over the region 30-90° N, as measured by anomaly correlation, remains positive out to 30 days, although neither model produces an example of a highly skilful ensemble forecast at extended range. The ECMWF forecasts are more skilful for days 1-10 as might be expected given the T63 model's higher spatial resolution. Beyond this forecast period the ECMWF model remains superior on average, although the difference is small. There is, however, considerable case-by-case variability, with the ECMWF integrations being consistently better in December 1986 and March 1987, but much worse than the UKMO integrations in June 1986. This shows up clearly in the distribution of individual forecast scores (Figure 4). The average skill of a 'super-ensemble mean' of both models exceeds that of either constituent ensemble up to days 11-20. This suggests that the divergence between the models may provide additional predictability information. In either model the ensemble mean score exceeds the corresponding individual forecast score. However, the difference in skill is small compared with perfect model estimates, showing the potential for increased impact as more skilful models become available.

Forecasts from both models show wide geographic variations in local skill at extended range. This is particularly noticeable when the ensemble mean forecasts are considered, since ensemble averaging increases skill in regions where the individual forecast skill is positive, but decreases it in regions of individual negative skill. Even in those forecasts where the large-scale skill is fairly small, useful levels of local skill are achieved over a significant area, thus underlining the importance of local skill prediction. Potentially useful levels of correlation between local skill and ensemble spread, or average individual forecast anomaly magnitude, are observed for days 1-10 for both models. Use of contingency tables show that even where the correlation is rather small, selective skill prediction may still be possible.

The UKMO is now nearing the time when ensembles of dynamical forecasts can be run in real time to coincide with our long-range forecasting conferences. This paper has highlighted the potential of the ensemble forecasting technique for predicting local skill beyond the medium range. It is hoped that these ideas will be further developed as the number of ensembles available for analysis rapidly increases. There is also good

evidence that dynamical models can compete favourably with more established (statistical) techniques for predicting time-averaged values of regional rainfall and near-surface temperature at extended range through the use of highly skilful regression equations. There is distinct possibility, therefore, that the use of these equations within the framework of ensemble forecasting could lead to a real increase in skill over that currently achieved by the UK long-range forecasting system, particularly in the period 6-15 days ahead.

6. ACKNOWLEDGEMENT

The authors wish to thank ECMWF for the use of their T63 lagged-average forecast data.

7. REFERENCES

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

Blackmon, M. L., J. M. Wallace, N. C. Lau and S. L. Mullen, 1977: An observational study of the northern hemisphere wintertime circulation. J. Atmos. Sci., 34, 1040-1053.

Brankovic, C., 1988: Lagged average forecasting with the ECMWF model. Proceedings of the ECMWF workshop on predictability in the medium and extended range, 16-18 May 1988, ECMWF, Shinfield Park, Reading UK.

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

Folland, C. K. and A. Woodcock, 1986: Experimental monthly long-range forecasts for the United Kingdom. Pt.I. Description of the forecasting system. Met. Mag., 115, 301-318.

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

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

Hollingsworth, A., U, Cubasch, S. Tibaldi, C. Brankovic, T. N. Palmer and L. Campbell, 1987: Mid-latitude atmospheric prediction on time scales of 10-30 days. In 'Variability in the atmosphere and oceans', Roy. Met. Soc. Monograph.

Kalnay, E. and A. Dalcher, 1987: Forecasting forecast skill. Mon. Wea. Rev., 115, 349-356.

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

Mansfield, D.A., 1986: The skill of dynamical long-range forecasts, including the effect of sea surface temperature anomalies. Quart. J. Roy. Met. Soc., 112, 1145-1176.

Molteni, F., U. Cubasch and S. Tibaldi, 1986: 30- and 60- day forecast experiments with the ECMWF spectral models. Proceedings of the ECMWF workshop on predictability in the medium and extended range, 17-19 March 1986. ECMWF, Shinfield Park, Reading . UK.

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

Murphy, J. M. and A. Dickinson, 1988: Extended range prediction experiments using an 11-level GCM. To appear in Met. and Atmos. Physics.

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

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.

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

Slingo, A., 1985: Handbook of the Meteorological Office 11-layer atmospheric general circulation model. Vol.1: model descrption. Dynamical Climatology Tech. Note No. 29, Meteorological Office, Bracknell, England.

Tibaldi, S., 1986: Envelope orography and maintenance of the quasi-stationary circulation in the ECMWF global models. Adv. in Geophys., 29, 339-374.