RESULTS FROM EXPERIMENTAL EXTENDED RANGE FORECASTS AT THE UK METEOROLOGICAL OFFICE

T.N. Palmer, J.M. Murphy and J.A. Owen
Meteorological Office
Bracknell, U.K.

Summary: The use of ensembles of integrations for extended range prediction has been studied with a hemispheric version of the Meteorological Office 5-layer GCM employing climatological SSTs. In addition to their actual predictive skill, the ensemble forecasts were also verified under 'perfect model' conditions. A comparison of the effects of ensemble averaging with spatial or temporal averaging is made. The ensemble technique is now being used with the Meteorological Office global 11-layer model to produce real-time extended range forecasts for the long range forecasting conference. The first of these forecasts was made in September 1985, and some results will be shown. Extended range forecasts have also been studied to determine the improvement in forecast skill, using observed rather than climatological SST. These include a number of single integrations on the 5-layer model, and two ensembles of integrations from the 1982/3 El Nino winter on the 11-layer model. Results from these forecasts will be discussed.

1. INTRODUCTION

Integrations of numerical weather prediction models can show extended range forecast skill. For example, Mansfield (1986) has documented results from an integration of the Meteorological Office hemispheric 5-layer model (Corby et al, 1977), initialised using data from 14/12/76, where a 15 day average forecast field centred on day 40 has an anomaly correlation of almost 0.6. However, such cases are not typical. In general, the limit of deterministic predictability of the 5-layer model is around 10 days. In order to consider the general problem of extended range skill beyond this limit one must recognise that an individual (deterministic) forecast represents only one of an ensemble of equally likely outcomes given, for example, uncertainties in the initial analysis. The use, then, of ensembles of integrations for extended range prediction is compatible with the philosophy that forecasting beyond the limit of deterministic predictability is inherently probabilistic.

In this paper, ensembles of extended-range integrations will be discussed in three different contexts. In section 2 we consider a number of 7-member wintertime ensembles of integrations of the 5-layer model. Whilst not a state-of-the-art NWP model, it is economical to run and can be used to assess the potential of the technique. Comparison with the effects of spatial and temporal averaging will be made. The results quoted in section 2 are a summary of an extensive study by Murphy (1986) of extended range ensemble forecasting using the 5-layer model.

Beginning in September 1985, ensembles of real-time extended range forecasts have been made using the global 11-layer model (Slingo, 1985), and results from these have been made available for discussion at the long range forecast conferences in the Synoptic Climatology Branch of the Meteorological Office. At present, it is envisaged that a 7-member ensemble forecast will be run in this way every three months. Some results from the first of these real-time forecasts are discussed in section 3.

It has long been recognised that anomalous SST, particularly in the tropics, may be an important component of 'lower boundary forcing' influencing the predictability of extended range forecasts. Results by

Mansfield (1986) using the 5-layer model appear to confirm this. An ensemble of extended range forecasts has been run on the 11-layer model using data from the El Nino winter 1982/3, firstly with observed and secondly with climatological SST. The degradation in predictability with climatological SSTs is shown in section 4, both in the tropics and in the extratropics. Results will be compared with climate sensitivity experiments using seasonally-averaged SST anomalies (Palmer and Mansfield, 1986).

2. POTENTIAL IMPACT OF ENSEMBLE FORECASTING

We present in this section a brief and informal synopsis of a small part of the work carried out to determine the extent to which the extended range predictive skill of a dynamical model (in this case the 5-level model) might be improved by forming an ensemble-mean forecast from a number of individual integrations. A full description of these results is given in Murphy (1986).

Eight 50-day ensemble forecasts, each containing seven individual integrations, were made from winter initial conditions. The ensembles were created by adding spatially-correlated random perturbations to a given observed state to simulate the effects of analysis error. The method used was to take a linear combination of the observed state with an independent analysis such that the difference between the resulting perturbed state and the observed state corresponded to a typical analysis error (30 m rms at 500 mb). Each ensemble forecast was verified both against observations, and against an additional 'nature' integration also produced using the above perturbation technique. The purpose of the latter was to test the ensemble method under the (unrealistic) assumption that the model has no systematic biases. With this so-called 'perfect-model' assumption, the additional integration can be thought of as a possible realisation of the real atmosphere and used to verify the forecasts. The predictability limit under a perfect model assumption can be thought to provide an upper bound to the model's actual predictive skill.

Figs 1 and 2 show results averaged over the eight ensembles under this perfect-model assumption. All the curves refer to results for the 500 mb height field in the area 30-85°N. The skill score is defined by the anomaly correlation coefficient. To calculate this, model fields are calculated relative to an estimate of the model climatology obtained by averaging over the 8 'nature' runs.

Since, under perfect model conditions, the ensemble forecast contains information about the probability distribution of possible evolutions of the atmosphere from the initial state, we expect on average the ensemble-mean forecast to yield an improvement in predictability relative to an individual forecast. However, we may also expect to gain some improvement in skill by removing the least predictable scales of motion by spatial or temporal filtering.

The curves in Fig 1 show the relative effects of spatial filtering and ensemble-averaging on forecast skill. An improvement in skill is obtained if we consider only the long-wave component in an individual

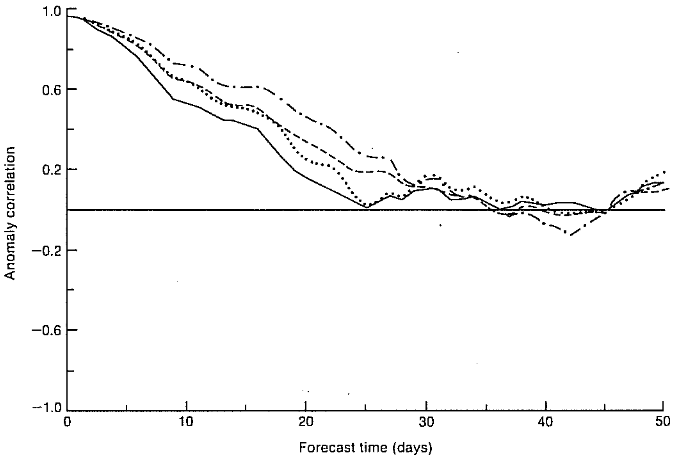


Figure 1 Average 'perfect model' forecast skill for 500 mb height field from 30-85°N for 5-level model integrations.(———) daily unfiltered individual forecast. (– – – – –) daily individual forecast, zonal waves 0-3 only. (••••••) daily ensemble-mean forecast. (–•–•–) daily ensemble-mean forecast, zonal waves 0-3 only.

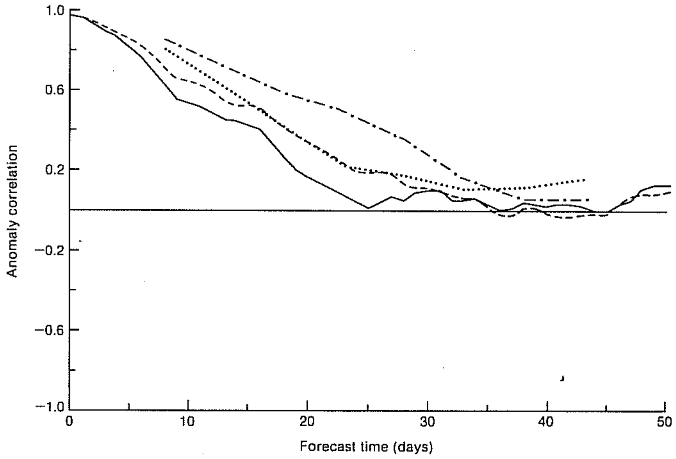


Figure 2 Average 'perfect model' forecast skill for 500 mb height field from 30-85°N for 5-level model integrations.(———) daily unfiltered individual forecast. (– – – –) 15-day mean individual forecast. (••••••) daily ensemble-mean forecast. (–•–•–) 15-day mean ensemble-mean forecast.

forecast. However, if we consider the long-wave component of the ensemble-mean forecast, the skill is improved still further. In terms of the predictability limit, the spatially-filtered ensemble-mean forecast offers an improvement of almost 50% compared with the unfiltered individual forecast; 19 days for an individual forecast and 28 days for the spatially-filtered ensemble-mean forecast. Fig 2 compares the effects of time- and ensemble-averaging. Again we observe a substantial improvement in skill in the time-averaged ensemble-mean forecast compared with the time-averaged individual forecast.

The improvement of the ensemble-mean forecast is to be expected if the spatial variance of each individual forecast anomaly field is greater than the spatial variance of the corresponding ensemble-mean anomaly field. When this occurs one can readily show that, if the anomaly correlation of each individual forecast field is positive, the anomaly correlation of the ensemble-mean field will be greater than or equal to the mean anomaly correlation of the individual fields comprising the ensemble.

This effect appears to be shown when the ensemble-mean fields are verified against real data. In particular, in two of the eight cases discussed above, the individual forecasts were found to show skill well beyond the model's average limit of predictability (Mansfield, 1986). In these two cases the ensemble-mean forecast showed a substantial improvement in skill. To investigate this further, another three random perturbation ensemble forecasts were produced for three independent situations where an individual forecast, run previously, had shown an unusually high degree of skill. In two of these three cases, the average skill of individual forecasts within the ensemble remained positive throughout, and the ensemble-mean forecast again showed a significant increase in skill. Fig 3 illustrates the average improvement in skill given by the ensemble-mean forecast in these four unusually predictable cases compared with the effect in the other seven cases where the model did not show any skill beyond the normal predictability limit. When verified against real data, Mansfield (1986) has shown that time-averaging also improves predictability of the 5-layer model, by comparing the rms error of daily and 15-day average fields. He found that error growth rates of these two fields are identical. However, because climate-mean variability of daily error fields is larger than that of 15 day mean fields, the time at which a daily forecast field ceases to be significantly skilful (defined by a statistical significance test) is less than the time at which a 15 day mean field ceases to be skilful.

Since some weather situations are more predictable than others, forecast skill will vary from case to case. The extent to which these variations in skill may be predicted by corresponding variations in ensemble spread is clearly an important issue. In Murphy (1986), the correlation between spread and skill has been studied for the 8 ensembles using both amplitude and phase measures of spread and skill. In the perfect model case there was a significant correlation between spread and skill, but only for the first half of the period for which the ensemble-mean forecast retained significant skill. When verified

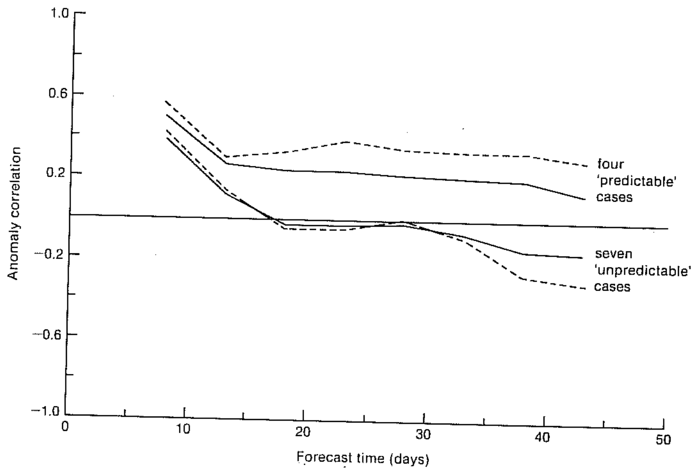


Figure 3 Practical forecast skill for four 'predictable' 5-level model forecasts and seven 'unpredictable' forecasts, for the 500 mb height field from 30-85°N. (———) 15-mean individual forecast. (– – – – –) 15-day mean ensemble-mean forecast.

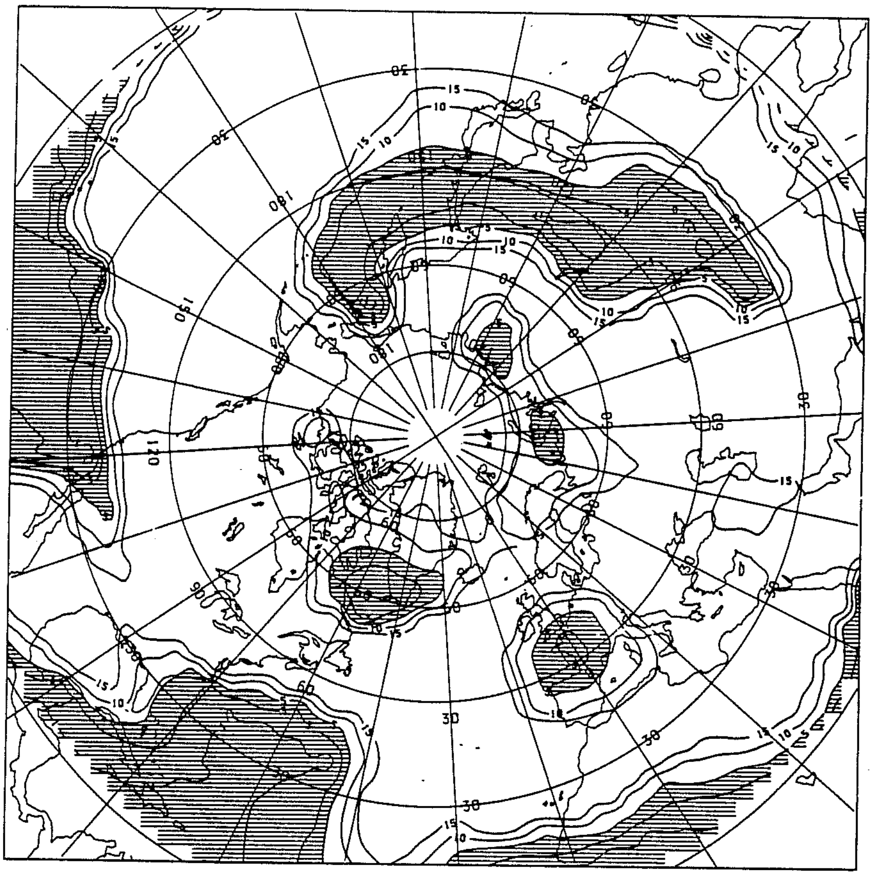


Figure 4 Contours of significance level (%) based on point-by-point t-tests on the days 16-30 ensemble-mean 500 mb height anomaly field from the 5-level model ensemble forecast from 14.12.76. Areas shaded are significant at the 5% level or better.

against real data, no significant correlation existed during any period of the forecast.

However, the ensemble distribution may also be used to provide information about the geographical variation of forecast skill. The idea is to pick out areas where the ensemble-mean anomaly is statistically significant and ask whether those areas, on average, show more skill than the field as a whole. This is testing the notion that where the ensemble 'clusters together' it is likely to be skilful.

To identify significant areas, a statistical t-test was applied to 15-day mean ensemble-average forecast fields, and areas significant at levels greater than 5% were picked out. Fig 4 shows an example of this. Anomaly correlation scores were then calculated for each area, and compared with the full field anomaly correlation score. It was found that, averaging over all the ensemble forecasts, the skill score for the limited areas was greater than the full field score by an amount statistically significant at the 10% level in those cases where the full field score was positive. In the example given in Fig 4, the full field score is 0.57, whereas the average score within the shaded regions is 0.90.

3. REAL-TIME EXTENDED-RANGE ENSEMBLE FORECASTS

A global 11-layer climate model was used for the real-time integrations. The model, similar in many respects to the 5-layer model, has a regular latitude/longitude grid with $2^1/_2$ x $3^3/_4$° resolution, and is described in detail by Slingo (1985). The 11-layer model was designed for long period integrations and has an energy conserving finite difference scheme. It also has sophisticated physical parametrizations all described in Slingo, op cit. For these reasons, it was thought to be the most appropriate model available to us for long range forecast integrations. Whilst it was not primarily designed as an NWP model, de facto, we treat it as such in this paper.

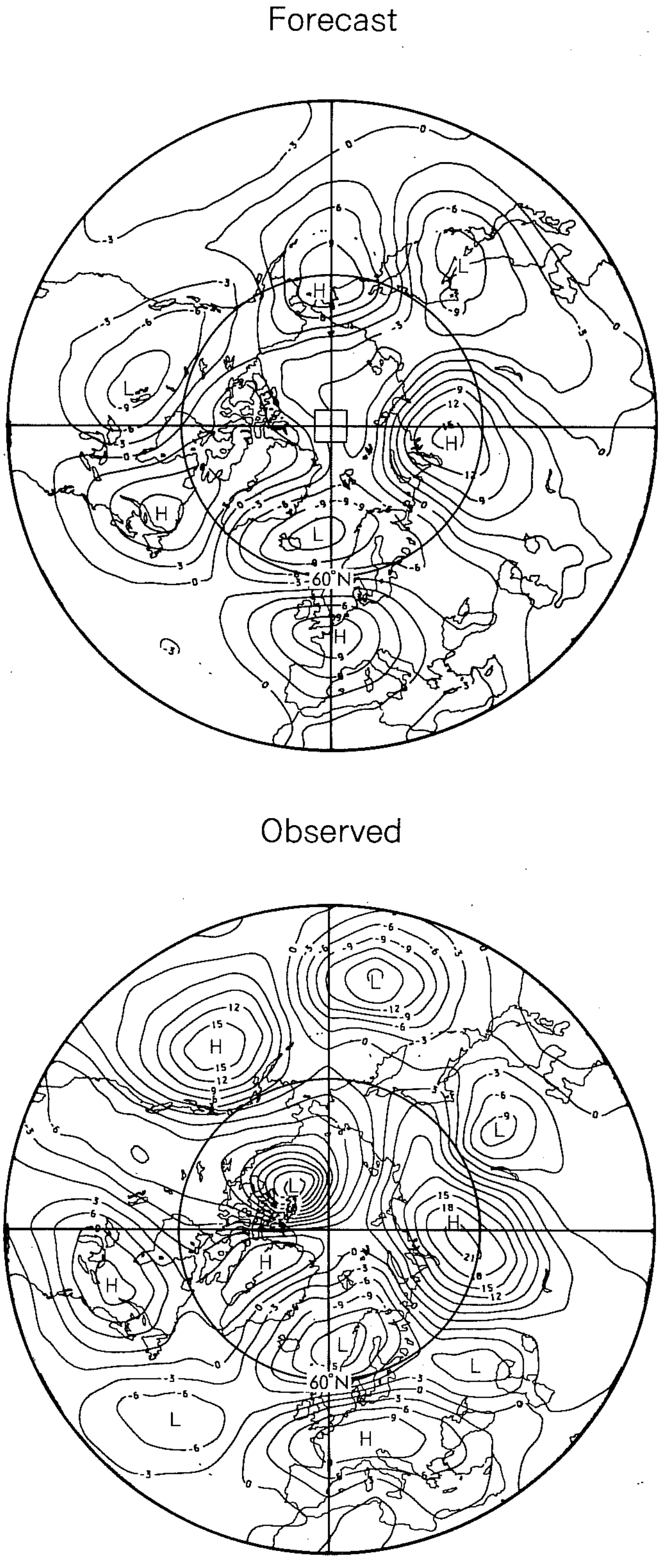
Instead of the spatially correlated random perturbation technique used to generate the 5-layer model ensembles, seven consecutive operational analyses at 12 hour intervals between 00Z 12.9.85 and 00Z 15.9.85 inclusive were used for initial conditions for the real-time forecast. The correspondence between this 'time-lagged' technique for generating initial conditions for an ensemble and the 'random perturbation' method depends on the forecast skill of the model, and it is still an open question as to which technique is more appropriate in practice. The initialisation dates were chosen to be as close as practicably possible to the date of the long range forecast conference on 17 September.

As in section 2, results are expressed as anomalies with respect to an estimate of the model's autumn climatology. The latter was formed from a set of eight integrations employing a selection of initial conditions at least 10 days apart from September and October of 1983 and 1984. These integrations used climatological sea surface temperatures (SSTs). For the forecast integrations, fixed SST anomalies based on operational SST analyses averaged over the 10 days preceding the initialisation date of the first of the forecasts were added on to a SST climatology which was updated every 5 days during the integrations.

From each forecast, time-mean fields were produced corresponding to the 5-day period immediately following the long range forecast conference (18.9.85-22.9.85), the next 10 day period (23.9.85-2.10.85) and the following 15 day period (3.10.85-17.10.85). (Clearly these fixed verification periods correspond to different model forecast times in each of the seven integrations.) Results from the ensemble of forecasts were considered, together with the multivariate statistical model (Maryon and Storey, 1985), in producing the mid September long range forecast.

The five days 18-22 September represent a medium-range period of the forecast during which there was a relatively small spread between the ensemble members and there was a reasonable degree of skill in their predictions. Fig 5 shows the ensemble-mean 500 mb height anomaly field for this period compared with the verifying actual anomaly field. Most of the major centres have a counterpart in the ensemble-mean, the most notable exception being the deep low of -30 decametres centred just off the pole. There is a fairly close correspondence between the individual forecasts in most areas at this stage (not shown).

The correlation between the ensemble-mean and observed anomaly patterns, for the northern hemisphere north of 15°N, is 0.49 (table 1). At this range we could of course improve the skill somewhat by weighting the more recent individual integrations more highly when



Forecast

Observed

Figure 5 Ensemble-mean forecast and observed 500 mb height anomaly patterns for 18-22 September. Contour interval 3 dam.

forecast period integration	18-22 SEPT	23 SEPT-2 OCT	3-17 OCT
1 (00Z 12.9.85)	0.09	0.16	-0.01
2 (12Z 12.9.85)	0.21	0.10	0.08
3 (00Z 13.9.85)	0.14	0.01	-0.05
4 (12Z 13.9.85)	0.46	0.18	-0.05
5 (00Z 14.9.85)	0.51	0.34	0.09
6 (12Z 14.9.85)	0.65	0.21	0.02
7 (00Z 15.9.85)	0.60	0.23	-0.07
1-7 average individual forecast score	0.38	0.18	0.00
1-7 ensemble-mean forecast score	0.49	0.25	-0.12

TABLE 1. Anomaly correlation scores for forecast 500 mb height anomaly fields from 15°N-90°N for the three periods of the long-range forecast discussed in section 3.

forming the ensemble-mean as they are likely on average to be more skilful than the earlier ones (Table 1 confirms this to be so in the present case). However, since we are mainly interested in the extended range forecast periods, for which we can treat the integrations essentially as equally likely realisations, such a procedure was not used.

The ensemble-mean forecast for the 10-day period 23 September-2 October (days 6-15 of the ensemble forecast) and the actual anomaly map are not shown for reasons of space. There is a reduction in the intensity of the ensemble-mean anomaly field compared with observations, which reflects the spreading of the ensemble towards loss of predictability. Nevertheless a degree of coherence still appears to exist between the individual patterns. To determine objectively whether the ensemble-mean anomaly field represented anything more than random noise would require a series of point-by-point statistical t-tests on the ensemble-mean anomaly field as, for example, shown in Fig 4 to ascertain whether the number of points at which the anomaly was significant was greater than that expected by chance. Table 1 reveals that all the individual forecasts retain a positive anomaly correlation at this stage, although the level of correlation is low, with an average score of 0.18 compared with a score of 0.25 for the ensemble-mean. This difference in skill, although modest, illustrates the principle of increasing the signal-to-noise ratio through ensemble averaging.

Fig 6 shows the forecast patterns from each integration for the 15 day period 3-17 October (days 16-30 of the ensemble forecast). The ensemble-mean pattern has a featureless 'washed-out' nature suggesting initially that the ensemble distribution has become essentially random by this point. It is certainly true that, as measured by anomaly correlation, the forecast has completely lost skill at this stage (table 1).

In section 2 we considered only the full ensemble-mean of all seven integrations. If an ensemble forecast is always normally distributed about its mean, this is an appropriate quantity to consider. However if, as postulated in the introduction, ensemble distributions tend to cluster into a small number of distinct groups, the full ensemble-mean may become less meaningful and we should alternatively form 'sub-ensemble-means' from the members of each separate cluster, presenting the final forecast as a series of probabilities based on each of the sub-ensemble-means.

Interestingly there does appear to be some evidence of such clustering among the individual integrations in Fig 6. In integrations 5-7 there is a pattern showing areas of low anomaly centred in the Pacific and near Hudson Bay with an area of high anomaly between, somewhat similar to the wintertime Pacific/North American (PNA) pattern of Wallace and Gutzler (1981). In contrast integrations 1-4 show no sign of this pattern, but all show a low anomaly centred near Alaska. Thus we could consider that, at least in the Pacific/North American region, the

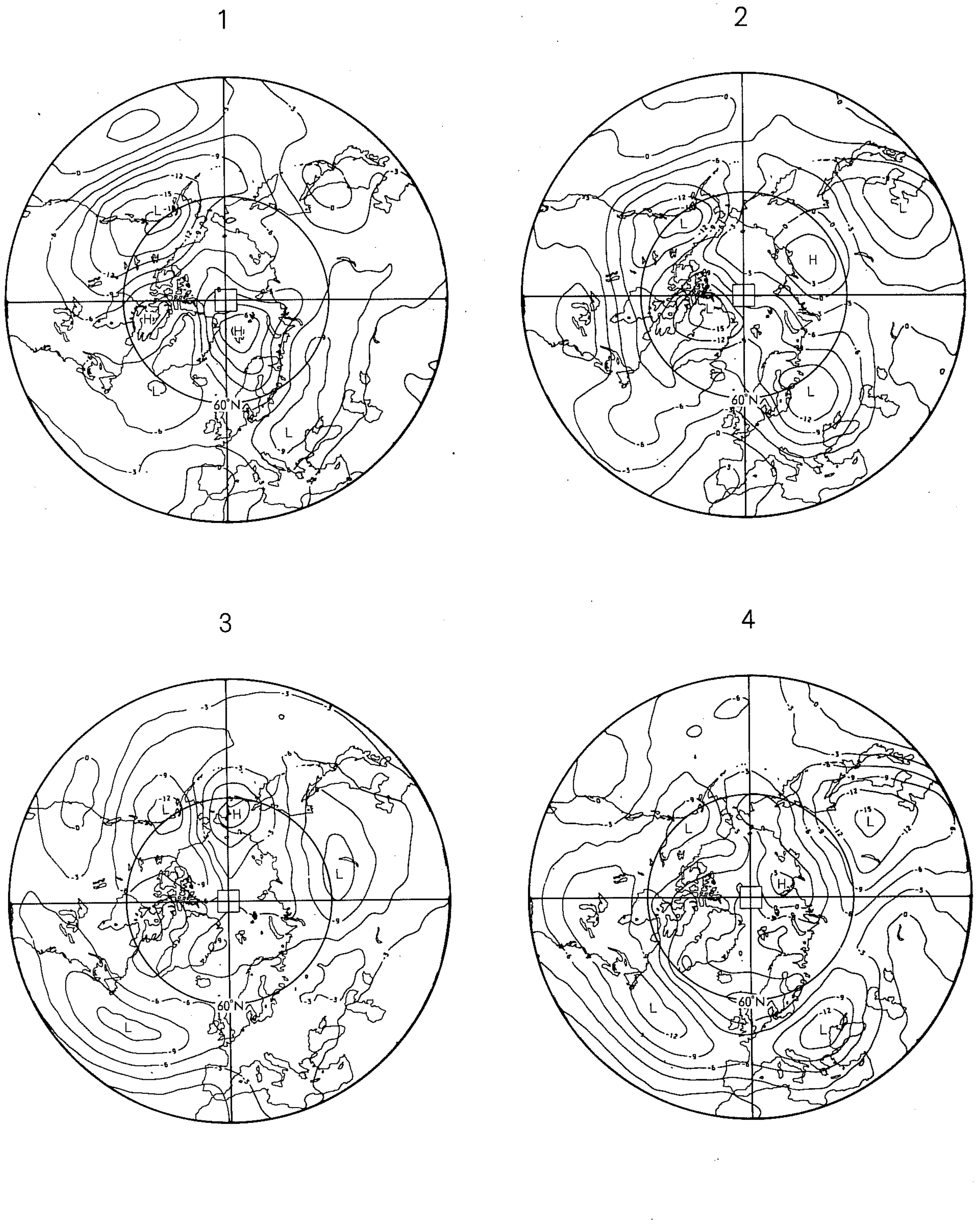
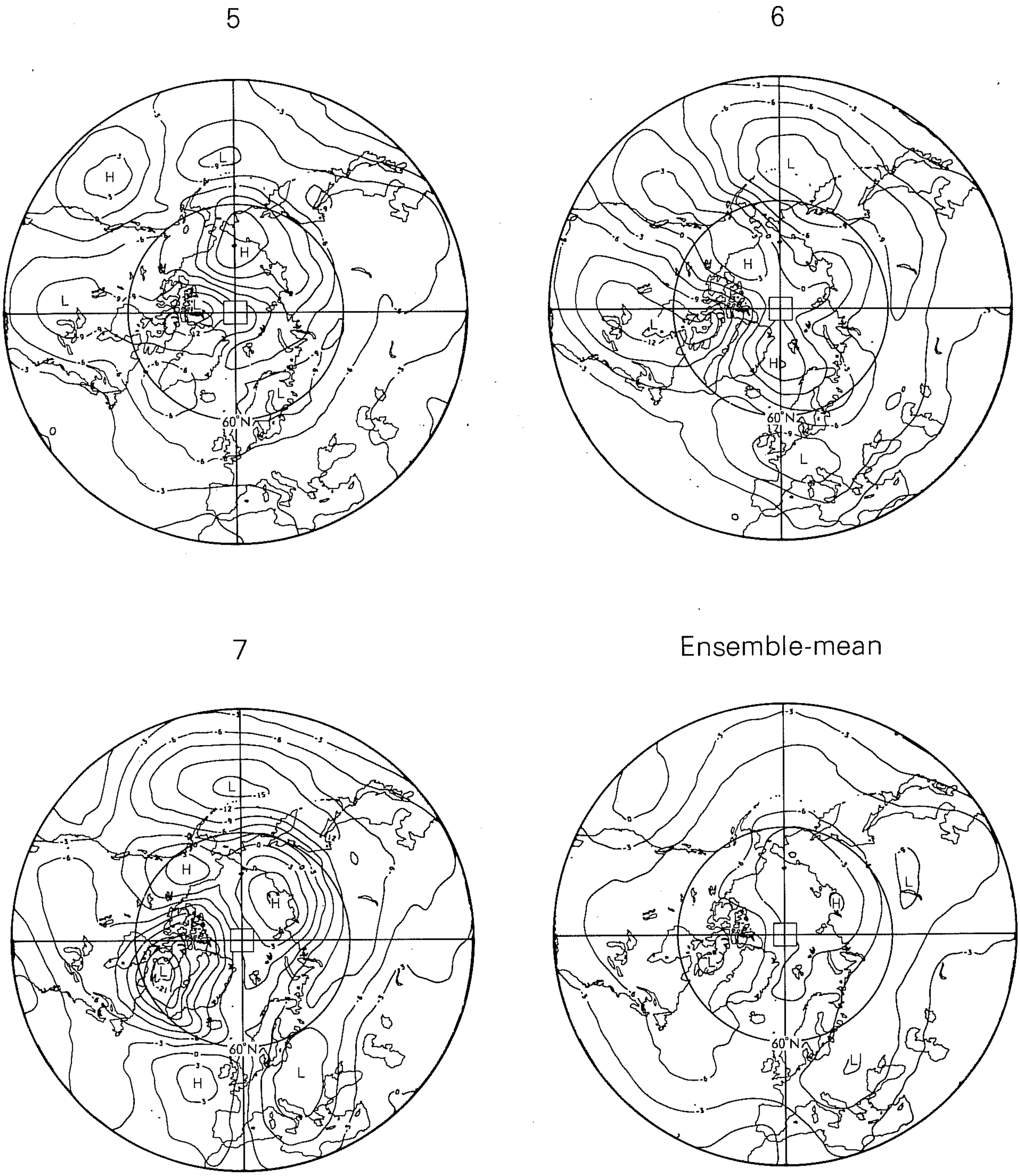


Figure 6 Individual and ensemble-mean forecast 500 mb height anomaly patterns for 3-17 October. Contour interval 3 dam.



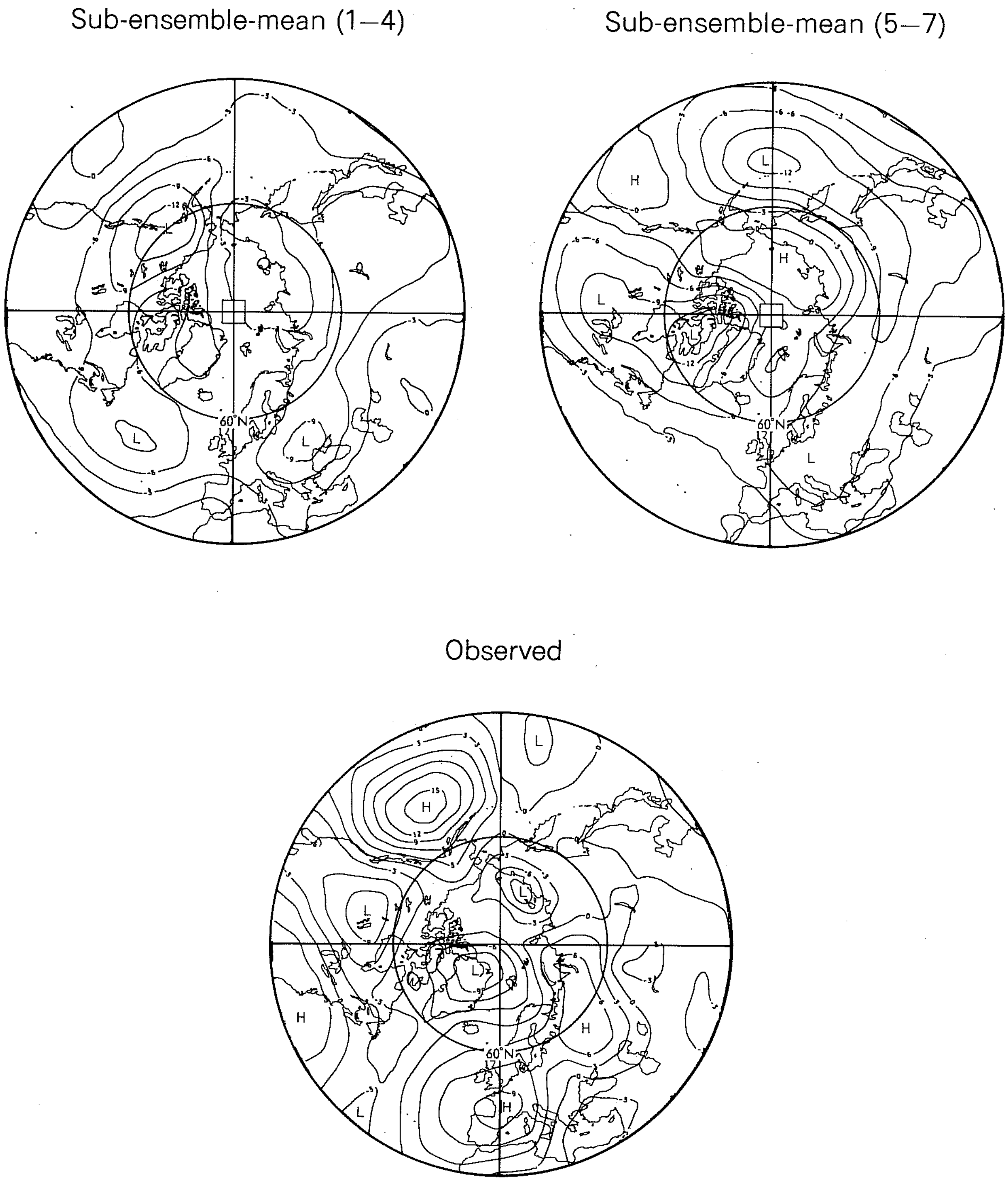
5

6

7

Ensemble-mean

Figure 6 (Cont.)



Sub-ensemble-mean (1—4)

Sub-ensemble-mean (5—7)

Observed

Figure 7 500 mb height anomaly for 3-17 October. Contour interval 3 dam.

(a) sub-ensemble-mean forecast formed from integrations (1-4)

(b) sub-ensemble-mean forecast formed from integrations (5-7)

(c) observed pattern.

integrations have split into two definite groups. The two relevant sub-ensemble-means are shown in figure 7a,b. Despite the slack pattern of the full ensemble-mean this clustering suggests that the ensemble distribution has not yet become randomly distributed. Furthermore the actual atmospheric anomaly pattern (Fig 7c) in the Pacific/North American area does show a structure similar to that of the 5-7 sub-ensemble-mean. However, since the latter incorrectly predicts high anomalies in polar regions and also underestimates the broadness of the high part of the pattern at lower latitudes, the correspondence in type does not show up in terms of objective skill. Nevertheless it is certainly encouraging that the actual pattern seems to bear a notable subjective resemblance to that which defines one of the two subsets predicted by the ensemble forecast.

To demonstrate such clustering behaviour objectively would require a method of cluster analysis, possibly based on a criterion of maximising the phase correlation between cluster members rather than the more conventional one of minimising rms difference. Such a method will be developed to aid the investigation of this intriguing phenomenon in future long-range ensemble forecast experiments.

4. INFLUENCE OF SEA SURFACE TEMPERATURE ANOMALIES ON LONG RANGE FORECASTS

From a set of 9 50-day 5-layer model forecasts, over 5 separate winters, Mansfield (1986) concluded that there is evidence of increased extratropical extended range forecast skill when observed as opposed to climatological SSTs are used in the integrations. In spite of the hemispheric domain of the model, Mansfield notes that, on occasion, this increase in skill is quite substantial. (Though, averaged over all 9 cases, improvement was modest).

In order to test this conclusion further, two ensembles of 90 day forecast experiments were run from 12Z ECMWF analyses from the 15th, 16th and 17th December 1982. One ensemble had seasonally varying climatological SSTs (Alexander and Mobley, 1976); the other had observed SST anomalies (from the Climate Analysis Center) added to these climatological values. In both ensembles, SSTs were updated every 5 days. Similar experiments have been performed by other centres in a coordinated modelling effort (Shukla, 1986). In the following, day 1 is taken to be 18 December for all integrations.

The difference in 200 mb streamfunction for days 1-30, between each of the three forecasts from 15th, 16th and 17th December with observed and climatological SSTs is shown in Figure 8a)-c) respectively. Apart from the anticyclone pairs over the tropical East Pacific, the westerlies elsewhere in the tropics and the cyclonic centres over the Southern United States, there are considerable differences between each forecast difference field, indicating the strong influence of the initial conditions in the first 30 day mean. The ensemble-mean difference field, shown in Figure 8d, shows a very weak PNA pattern. For days 31-60 (Figure 9) the PNA pattern has stronger amplitude than in Figure 8d). In both Figures 8 and 9, an anomalous tropical westerly band is apparent.

The 200 mb streamfunction difference between the ensemble-mean fields and verifying data is shown in Figures 10-11 for days 1-30, 31-60 respectively. For both periods there is a clear and unambiguous improvement in forecast skill in the tropics with observed rather than climatological SSTs. In the extratropics there is no discernible difference for days 1-30 between the forecast skill with observed and climatological SSTs. For days 31-60 (and 61-90), however, especially over the PNA area, there is a visually discernible improvement in forecast skill with observed SSTs.

An objective measure of skill (30 day mean 200 mb rms wind speed error) is shown in Figure 12a)-c) for the areas 90S-30S, 30S-30N, 30N-90N. Forecast skill is improved throughout with observed SSTs, though improvement in tropical forecast skill is the most spectacular.

The ensemble-mean difference fields shown in Figure 8d and Figure 9 closely resemble the 'equilibrium' response to a composite December 1982-February 1983 SST anomaly discussed by Palmer and Mansfield (1986). In this climate sensitivity study, the 11-layer model was integrated from one set of initial conditions, for 540 days, in perpetual January mode. Interestingly, the t-statistic for the 540 day mean field also gives a good indication of the consistency of the ensemble-mean response from one individual forecast to another. For example, the tropical Pacific anticyclone pairs and the cyclonic anomaly over the southern USA shown in Figures 8 and 9 are all highly

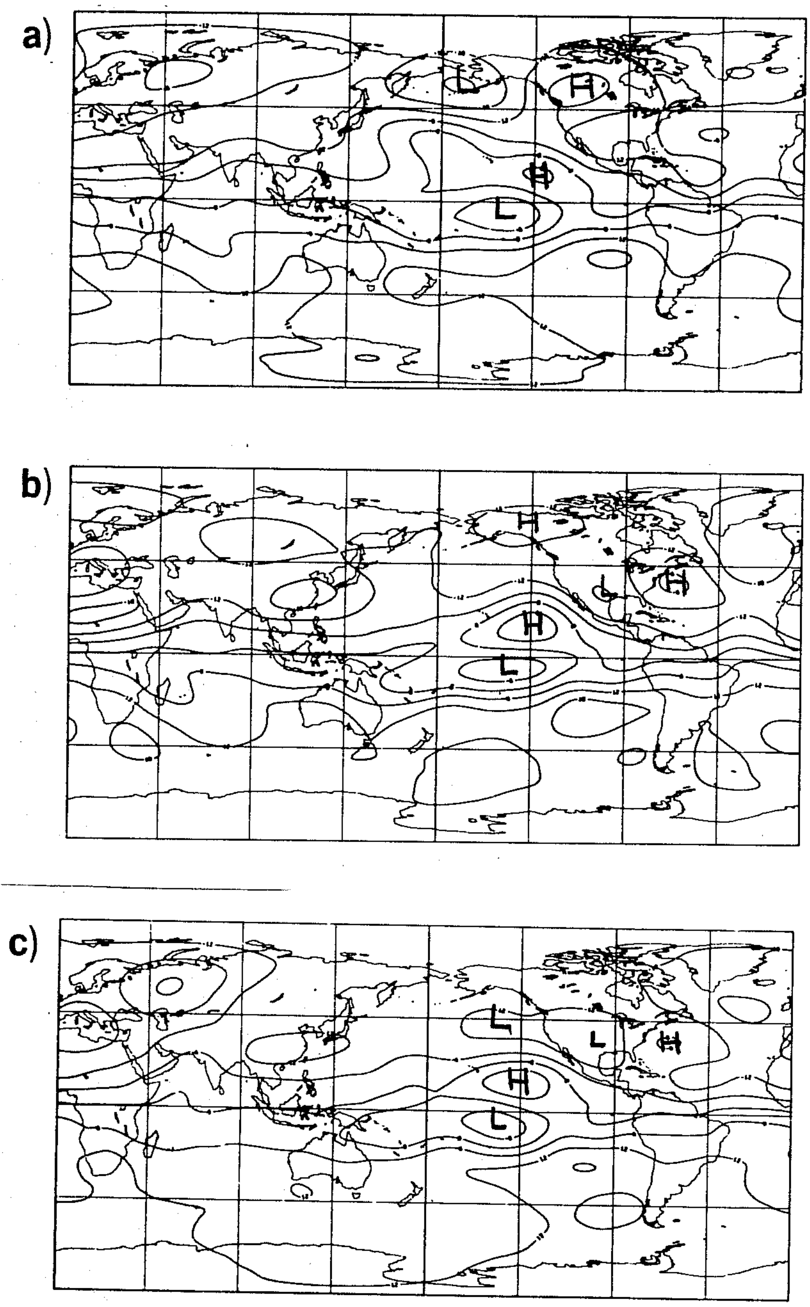


Figure 8 200 mb streamfunction difference field for days 1-30 between integrations with observed SST and climatological SST. Contour interval 6 x $10^6 m^2 s^{-1}$.

(a) Initial date 15th December

(b) Initial date 16th December

(c) Initial date 17th December

(d) Ensemble mean

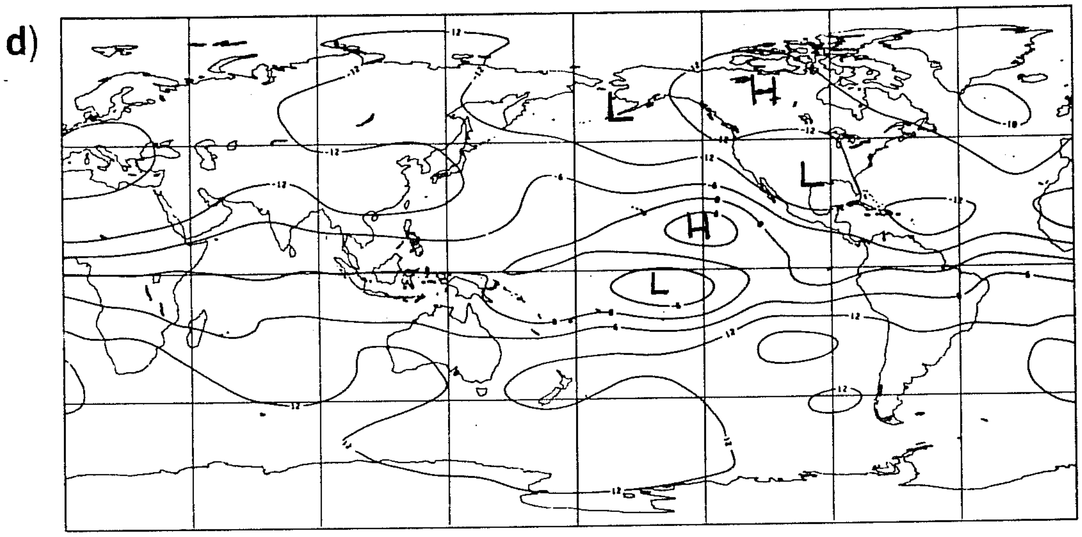


Figure 8 200 mb streamfunction difference field for days 1-30 between integrations with observed SST and climatological SST. Contour interval 6 x $10^6 m^2 s^{-1}$.

(a) Initial date 15th December

(b) Initial date 16th December

(c) Initial date 17th December

(d) Ensemble mean

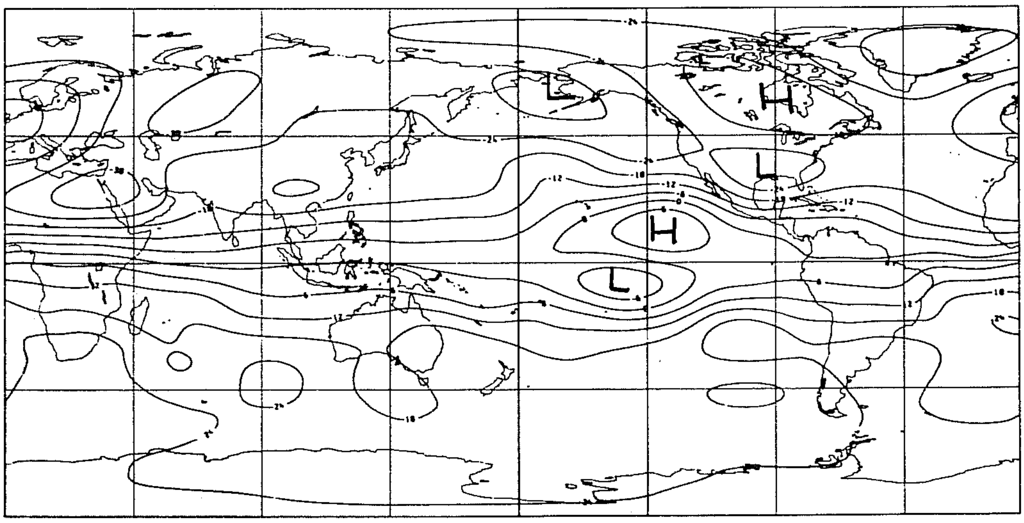


Figure 9 200 mb streamfunction difference field for days 31-60 between integrations with observed SST and climatological SST. Ensemble-mean only. Contour interval 6 x $10^6 m^2 s^{-1}$.

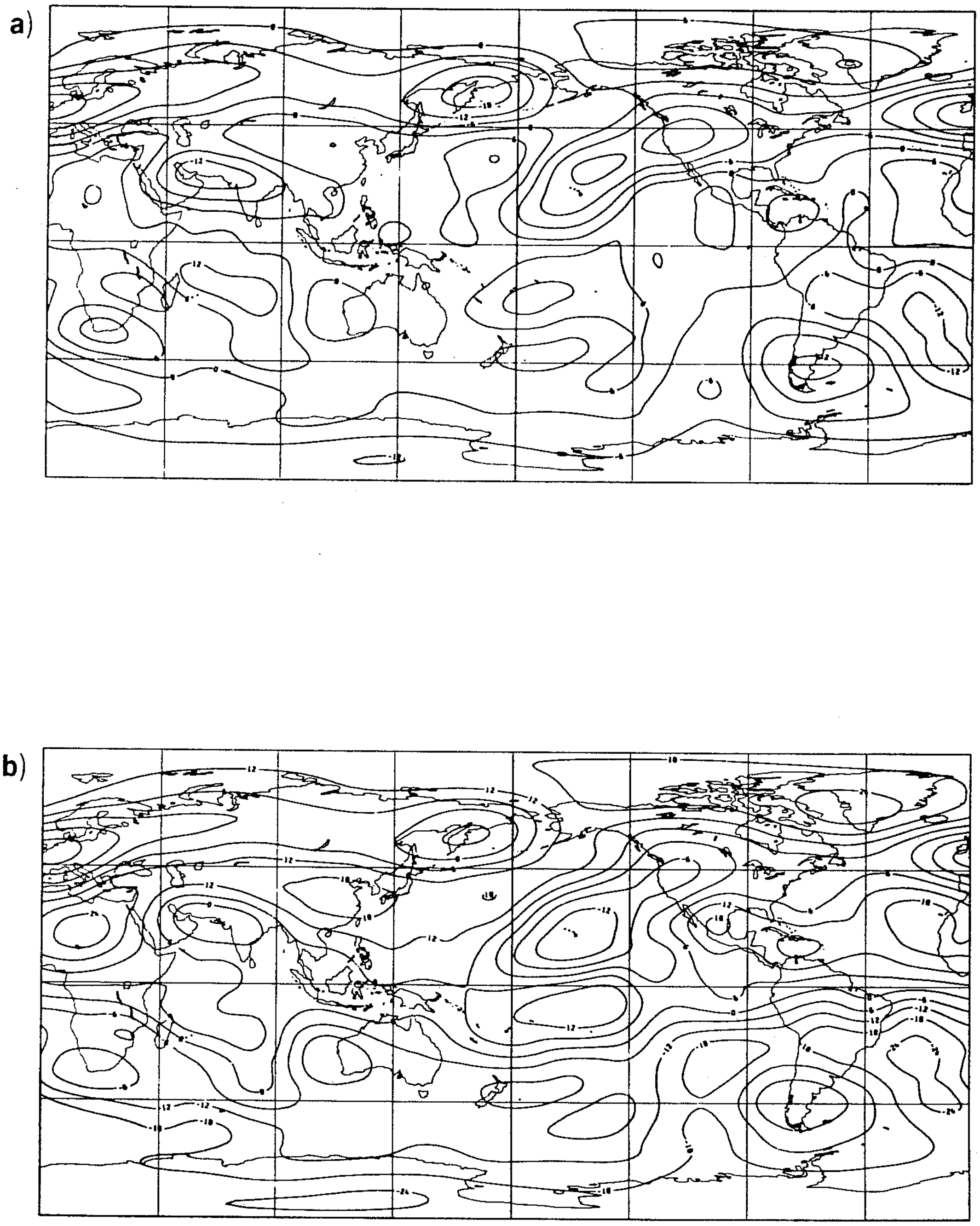


Figure 10 200 mb streamfunction difference field between ensemble-mean fields and verifying data for days 1-30. Contour interval 6 x $10^6 m^2 s^{-1}$.
(a) with observed SST
(b) with climatological SST

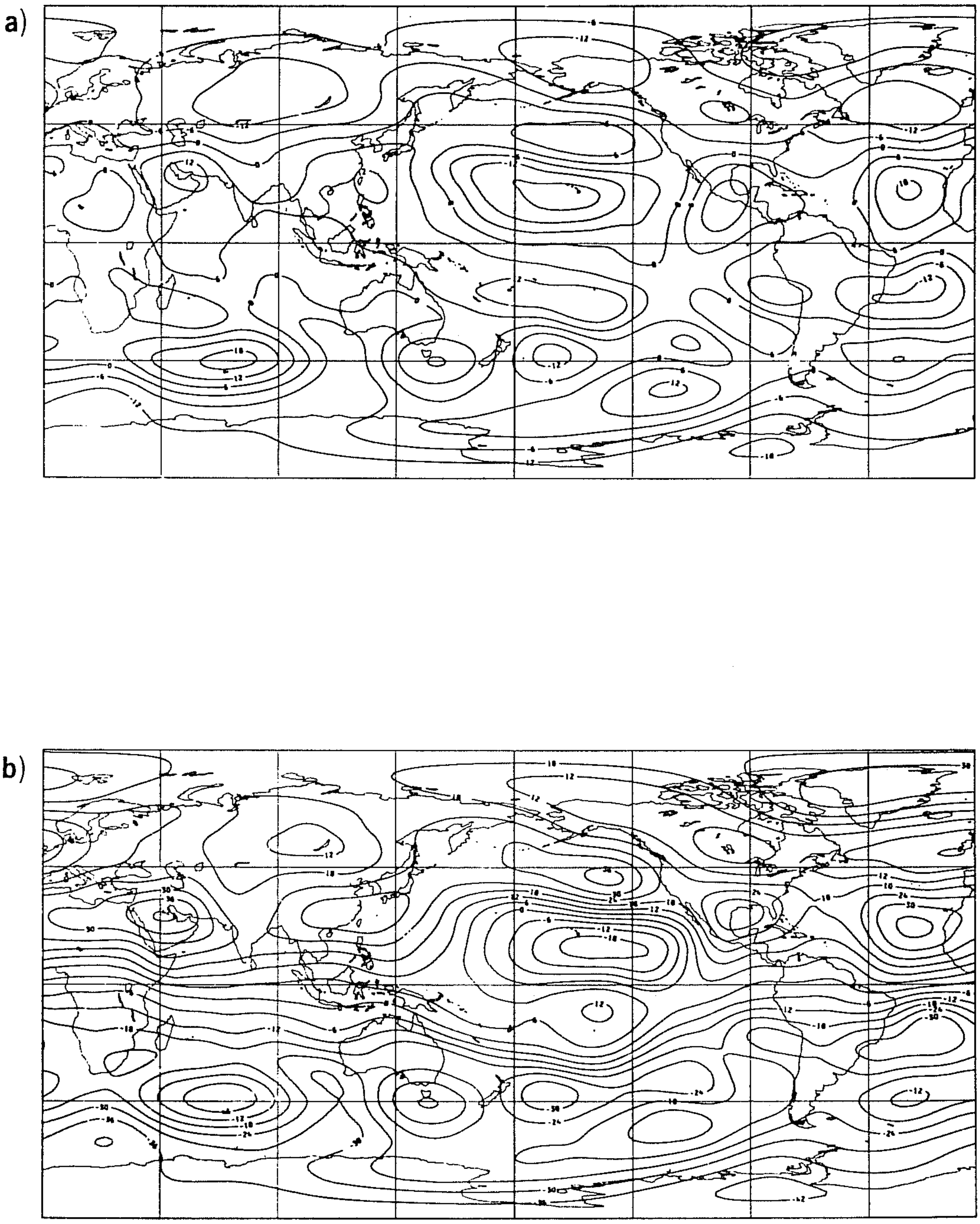


Figure 11 As Figure 10 but for days 31-60.

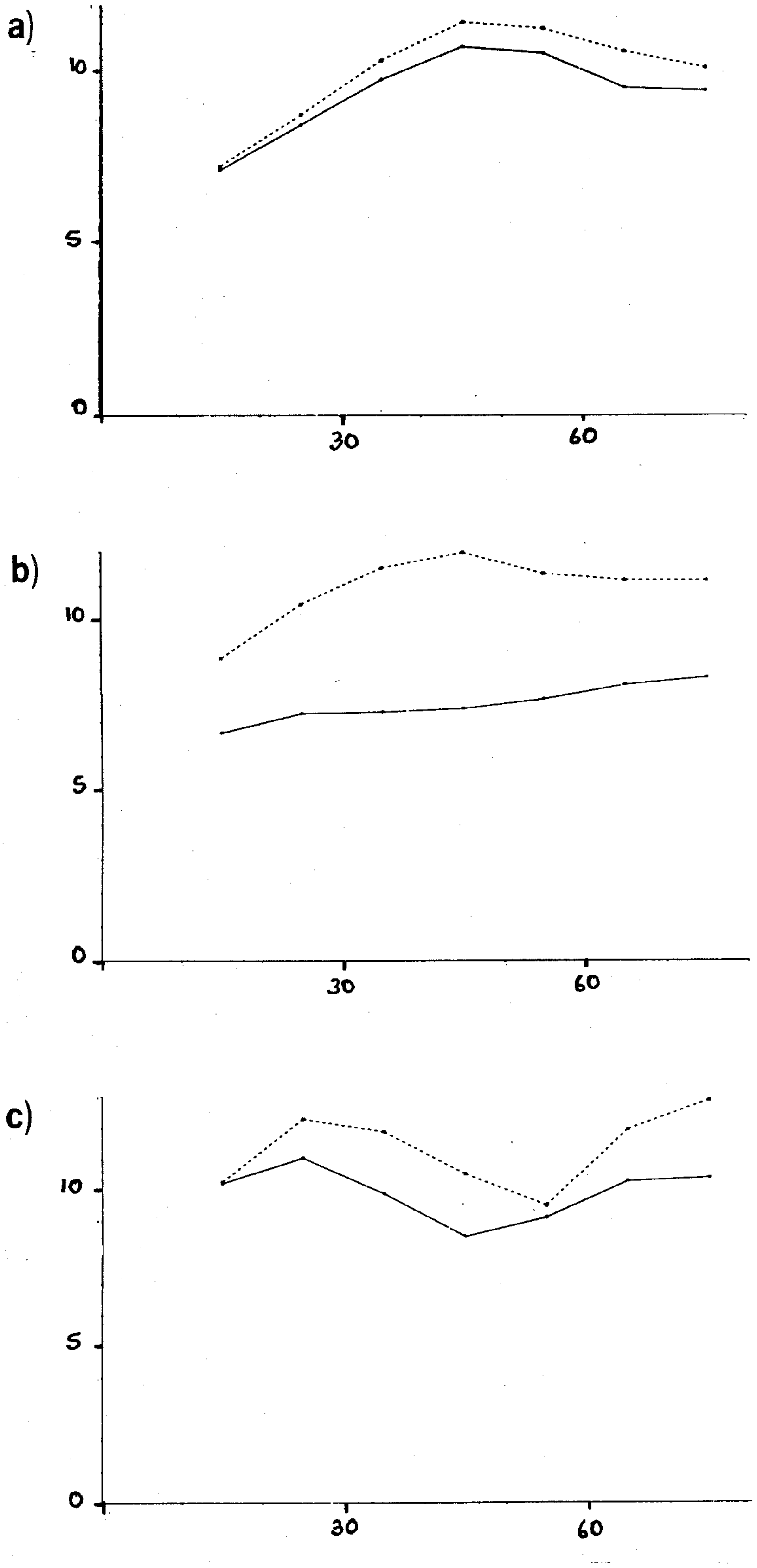


Figure 12 30-day mean 200 mb rms wind speed error (ms^{-1}) with climatological (••••) and observed (——) SST.
(a) 90S-30S;
(b) 30S-30N;
(c) 30N-90N.

significant in the equilibrium response (see Figure 9 of Palmer and Mansfield, 1986, part II).

In summary, while the impact of SST anomalies on an individual forecast is strongly dependent on initial conditions, a lagged-average ensemble of three forecasts shows, even in the first 30 days, and certainly in the next 30 days, the extratropical PNA pattern obtained in a 540 day perpetual January mean. The ensemble-mean extratropical response is weak in the first 30 days and does not noticeably improve forecast skill. In the next 30 days, however, forecast skill is noticeably improved. In the tropics, it is dramatically improved, even in the first 30 days.

References

Alexander, R. C. and Mobley, R. L., 1976: Monthly average sea surface temperature and ice pack limits on a 1° global grid. Mon. Wea. Rev., 104, 143-148.

Corby, G. A., Gilchrist, A. and Rowntree, P. R., 1977: The United Kingdom Meteorological Office 5-level general circulation model. Methods Comp Physics, 17, 67-110.

Folland, C. K. and Woodcock, A., 1986: An updated description of recent developments in monthly long-range forecasting for the United Kingdom. Met Mag. To be submitted.

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. To appear in QJR Meteorol. Soc.

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

Murphy, J. M., 1986: The impact of ensemble forecasts on predictability. In preparation.

Palmer, T. N. and Mansfield, D. A., 1986: A study of wintertime circulation anomalies during past El Nino events, using a high resolution general circulation model. Parts I and II. Quart. J.R. Met. Soc. To appear.

Shukla, J., 1986: Presentation in this volume.

Slingo, A. (Ed), 1985: Handbook of the Meteorological Office 11-layer atmospheric general circulation model, Volume 1: Model description. Dynamical Climatology Technical Note No. 29, Meteorological Office, Bracknell.

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