ON THE PREDICTION OF FORECAST SKILL

T.N. Palmer and S. Tibaldi*
European Centre for Medium Range Weather Forecasts
Shinfield Park, Reading, Berkshire

1. INTRODUCTION

Despite impressive improvements in numerical weather prediction (NWP) over the last few years, forecast models show considerable variability in predictive skill on many different timescales. This has prompted Tennekes et al. (1987) to assert that "no forecast is complete without a forecast of forecast skill." Some results on operational skill forecasting have recently appeared in the literature (e.g. Gronaas, 1985; Branstator, 1986; Kalnay and Dalcher, 1987; Palmer and Tibaldi, 1986 and 1987). An example of forecast skill variability (taken from Simmons, 1987) is illustrated in Fig 1, showing the day 3, 7 and 10 hemispheric skill scores for the operational model for November 1983. It can be seen, for example, that the later the model validity time, the larger the variability in skill, on the large scale, from forecast to forecast. This suggests that, even though it is more difficult to forecast beyond the short range, it may not be more difficult to predict the skill of forecasts beyond the short range. Moreover, even though the mean skill of dynamical models in the extended range is typically very small, case to case variability is sufficiently large that a few are of genuine practical use (e.g. Hollingsworth et al., 1988) and might possibly be relatively easily identified a priori (Palmer, 1988).

It is apparent, therefore, that a scheme to predict forecast skill will have substantial benefit in the medium range, and is an essential requirement for dynamical extended range forecasting. The possible impact of such a scheme has been quantified in a simple way by Palmer and Tibaldi (1987). For example, based on ECMWF data for the last seven winters, if it was possible to devise a scheme which only discriminated between forecasts of above average and below average skill, the effective usefulness of the forecast system during the above-average periods could be enhanced overall by about 2 days at the end of the forecast period. This compares favourably to the impact of any recent change in model resolution or model physics.

Returning to Fig 1, one gets a clear impression that, superimposed on the day to day variability in forecast skill, there is a lower frequency fluctuation

* Department of Physics, University of Bologna, Italy

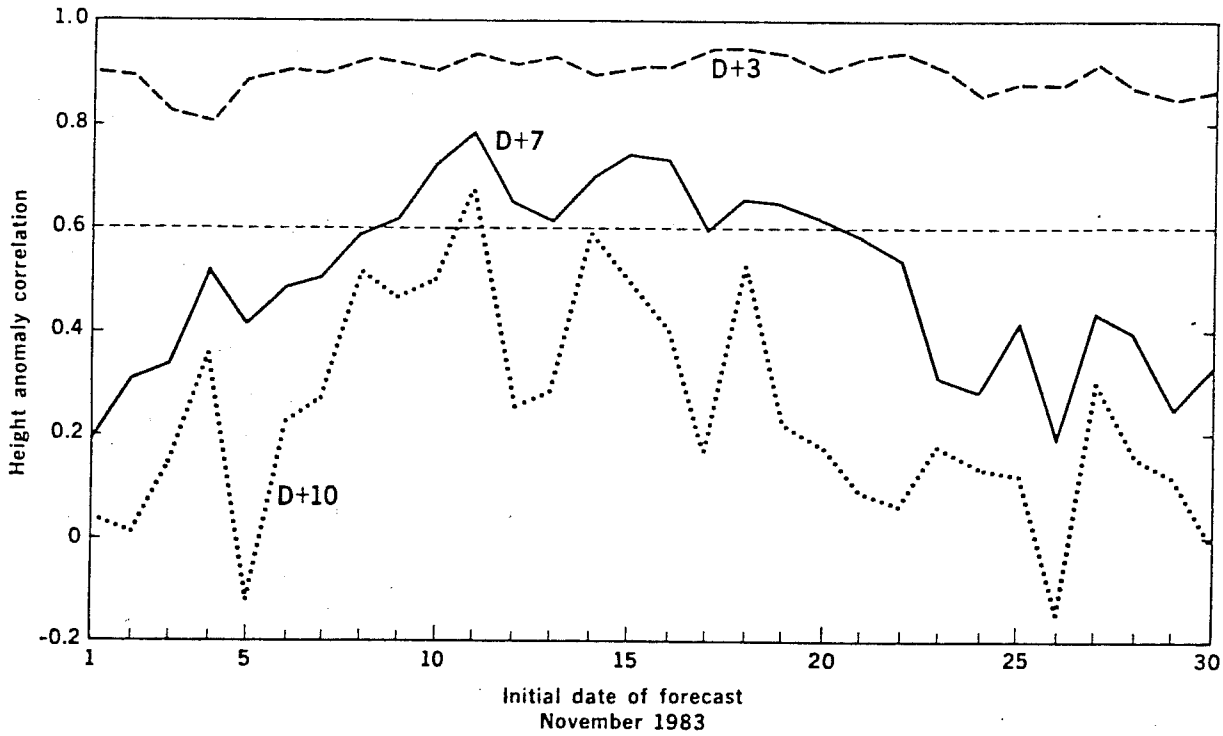

Fig. 1 Anomaly correlations of height for 1000-200 mb and the extratropical Northern Hemisphere for 3-, 7-, and 10-day forecasts performed from initial dates within the month of November 1983.

with timescale of about a week or so. Even if the prognosis of the day to day fluctuations proved difficult, there would still be value in forecasting this lower frequency variability.

In this paper we shall concentrate exclusively on the 10-day forecast timescale. This study is possible because a large enough statistical sample of forecasts is now available. Since the winter of 1979/80, ECMWF has archived into one dataset the 500mb height for each day of its 10-day forecast, including the initial and verifying analyses. This dataset forms the basis of our investigation into predictability and forecast skill of the ECMWF forecast system , preliminary results from which were discussed by Palmer and Tibaldi (1986,1987). The data is discussed more fully in section 2.

The predictors we have studied to forecast the skill of the 10-day forecasts can be thought of as falling into four categories. Motivated by studies of the dispersion within ensembles of forecasts, we have first investigated the consistency of forecasts verifying at the same time but initialised at different times. Secondly, we have studied in an objective manner, large-scale flow patterns associated with skilful and unskilful forecasts; the main tool used here was a statistical regression analysis with empirical orthogonal function coefficients of observed and forecast flow as predictors. Thirdly we have investigated the skill of earlier short range forecasts as predictors for the skill of the current forecast. Finally, motivated by a diagnostic study of the correlation between persistence errors and forecast errors, we have investigated the RMS difference between 500mb height of the forecast and of the initial conditions. Each of these investigations yields a set of potential predictors of forecast skill.

In section 3, the properties and skill of these predictor sets are derived from six winters of data. A seventh winter (1986/87) is used as an independent test of the predictors. Forecasts of forecast skill for 1986/87 are discussed in section 4.

Broadly speaking, one can distinguish two sources of forecast error. Firstly, loss of forecast skill can be associated with analysis errors, or errors in model formulation. These are related to deficiencies in the way we choose to

observe and simulate the atmosphere; they are essentially 'man made'. Secondly, there are errors associated with amplification of unavoidable uncertainties in the analysis or model formulation due to the intrinsic instabilities and nonlinearity of the atmosphere. This process could be thought of as 'God given'.

Some sources of variability of forecast error can be ascribed to one or other of these categories. For example, day-to-day changes in data coverage will result in day-to-day variability in forecast skill; changes in model formulation will result in much lower frequency variability. However, much of the observed variability of forecast skill is influenced by interactions between these two categories of forecast error. For instance, the propagation and amplification of analysis errors depend strongly on the structure and intrinsic stability of the large scale flow. Similarly, the way in which systematic deficiencies in model formulation influence forecast skill will depend on the detailed structure of the flow. For example, one might argue that deficiencies in the model's treatment of tropical convection will most strongly influence the extratropics when the large-scale flow is conducive to the meridional propagation of Rossby-wave activity.

These remarks serve to illustrate the fact that whilst it may be possible to find statistical relationships between the large-scale flow and forecast skill, the mechanisms underlying such relationships may be neither straightforward nor unique. In section 5 we discuss mechanisms associated with amplification of errors due to both barotropic and baroclinic instability, and also the influence of systematic errors in the model climate, in an attempt to understand the results of the statistical analyses given in earlier sections.

2. DATA

The basic data used in this study are taken from seven years of ECMWF forecasts. For each extended winter season from 1980/81 to 1986/87, the northern hemisphere 500mb geopotential height field of 100 day 1 to 10 daily forecasts and verifying analyses, from 1 December, were extracted from the global archives. The data for the first six years were then concatenated to form the 'training data' for the correlation and regression studies described below. The 100 forecasts and verifying analyses from the seventh available

winter, 1986/87, were used as an independent sample with which to test the reliability of the statistical results.

Amplitude and phase measures of forecast skill and forecast spread were calculated for the hemisphere north of 20°N, and for 12 regions (see Fig 2). Each region is bounded by 60°N and 30°N and extends from longitudes 0-30°E, 30-60°E.....30°W-0. Fig 2 also shows the climatological band-pass filtered standard deviation of 500mb geopotential height (from Lau et al., 1981), a good indicator of storm track activity. For future reference it should be noted that regions 12 and 1 are positioned at the end of the Atlantic storm track, and that regions 8 and 9 are positioned at the end of the Pacific storm track. Since low-frequency phenomena such as blocking are known to have preferred positions at the end of the storm tracks, regions 12, 1, 8, and 9 are distinguished from the others by a relatively high ratio of low-frequency to band-pass frequency variability.

The amplitude measure of skill is given straightforwardly by the root mean square difference (RMS) between forecast and verifying analysis. In order to calculate the phase measure of skill, first the 700-day mean of forecasts and verifying analyses were made. Each day n forecast was then expressed as an anomaly from the day n model climate (i.e. the anomaly was corrected for the mean model error). The correlation, ACC, between forecast anomaly and observed anomaly was then calculated. The unbounded phase measure ZAC, defined as the z-transform

$$ZAC = \frac{1}{2} \ln \left(\frac{1+ACC}{1-ACC}\right)$$

was also calculated. As is well known, the z-transform of the correlation coefficient between two such statistical distributions has an approximately Gaussian distribution (e.g. Fisher, 1921). It should be mentioned that an anomaly correlation score was also calculated defining the forecast anomaly in terms of deviation from the observed climate. However, in this paper we shall concentrate on results from the model 'bias-corrected' measure.

In addition, the RMS, ACC and ZAC measures of the difference between 500mb height of all overlapping forecasts were also calculated, as was the RMS

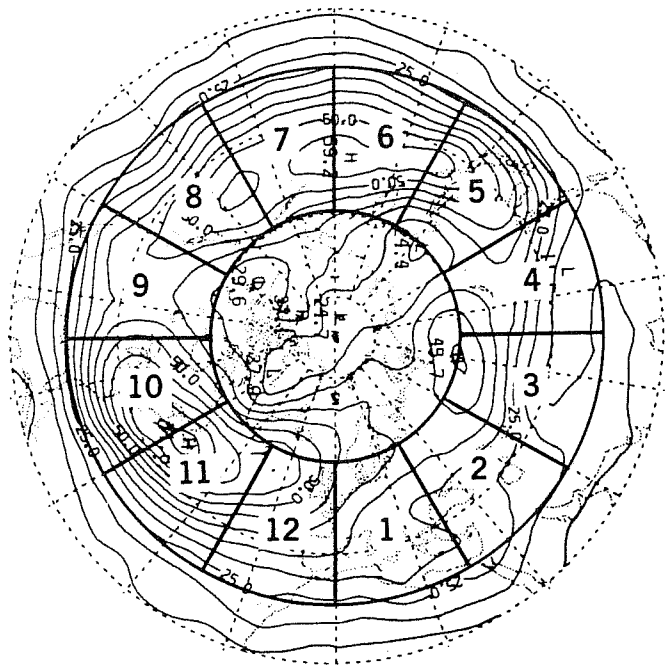

Fig. 2 Northern Hemisphere wintertime standard deviation of 500 mb geopotential height (m). Band-pass filtered retaining periods between 2.5 and 6 days (from Lau et al., 1981). Limited areas where skill scores have been computed are superimposed (1 to 12, 30° longitude width).

difference of each forecast and analysis field from the appropriate 700-day climate (the magnitude of the forecast and observed anomaly).

In order to study possible relationships between forecast skill and objectively defined configurations of the large scale flow, it is necessary to project the 500mb height data onto a suitable set of basis functions. One could use, for example, a spherical harmonic basis. However, in an attempt to minimise the number of basis functions required to describe any significant relationships between large-scale flow variability and skill variability, it was decided to project the data onto a set of N empirical orthogonal functions (EOFs) defined from pentad mean fields from 32 years of wintertime analyses (1952-84 , from NMC and ECMWF archives). These EOFs were calculated separately on the zonal mean and on deviations from zonal symmetry. The first of these EOFs corresponds to variation in the hemispheric mean height. Then five 'zonal' EOFs were retained, explaining 99.96% of the total variance of the pentad zonal mean fields, and 17 eddy EOFs, explaining 87% of the variance corresponding to fluctuations in the zonally varying component of the flow (N=1 + 5 + 17 = 23). For future reference, if a severe truncation is taken, retaining 3 zonal EOFs and 5 eddy EOFs (N=9), then the percentage of variance explained in the zonal mean flow and the zonally varying flow is 89% and 44% respectively. The first eddy EOF corresponds to the stationary wave pattern; higher eddy EOFs have some similarities with the teleconnection patterns of atmospheric low frequency variability (e.g. Wallace and Gutzler, 1981). For reasons of space, these EOFs are not illustrated here, but are discussed in Molteni (1987) and Palmer and Tibaldi (1986).

The results described below are derived either directly from this set of 700 forecasts, or from a temporally filtered version of it. By applying a 5-day running mean to the data, we attempt to extract from the forecasts and analyses, the low-frequency component of skill, which, as mentioned in the introduction, may be more easy to forecast.

3. PREDICTORS OF FORECAST SKILL

In this section, we shall consider four different sets which, given the present data, could be used to predict forecast skill. All of these predictor sets can be defined from forecast or initial analysis data, available at the time the forecast, whose skill we wish to predict, has been

integrated. We will test the skill and discuss properties of each predictor set on the training data for 1980/86.

3.1 Spread

Following the work of Epstein (1969), Leith (1974), and others, there has been much interest in recent years in the analysis of ensembles of integrations of NWP models, generated either through a Monte Carlo approach (see, for example, Murphy and Palmer, 1986) or a lagged average forecast approach (Hoffman and Kalnay, 1983; Molteni et al., 1987). One motivation for such studies is that the dispersion or spread of the ensemble can give an a priori estimate of forecast skill.

In the present study, our database allows us to analyse what might be termed a "poor man's lagged average ensemble" comprising two forecasts verifying at the same time, but initialised from consecutive analyses. However, we are only interested in predicting the skill of an individual deterministic forecast ("today's" forecast) and shall do this using the spread between this given forecast and the forecast initialised one day earlier (today's day n forecast and yesterday's day n+1 forecast).

Just as forecast skill can be measured in terms of phase or amplitude measures, so can forecast spread. The anomaly correlation coefficient between adjacent forecasts (or the z transform of it) provides the phase measure of spread; the RMS difference between adjacent forecasts provides the amplitude measure. Fig 3 shows a graph of the correlation between hemispheric average skill and spread using both ACC and RMS measures, based on daily data for 1980/86. It can be seen that there is a higher correlation between spread and skill using phase measure rather than amplitude measure. This confirms previous results by Branstator (1986) and is consistent with the results of Kalnay and Dalcher (1987) who found that, using anomaly correlation, the dispersion of ensembles of 5-day forecasts started from a number of different analyses was an excellent predictor of skill. Using RMS error, Kalnay and Dalcher were, in their own words, "less successful".

Kalnay and Dalcher speculated that the relative success of the spread predictor using anomaly correlation coefficient was due to its boundedness. However, in Fig 3 we also show the spread skill correlations using the

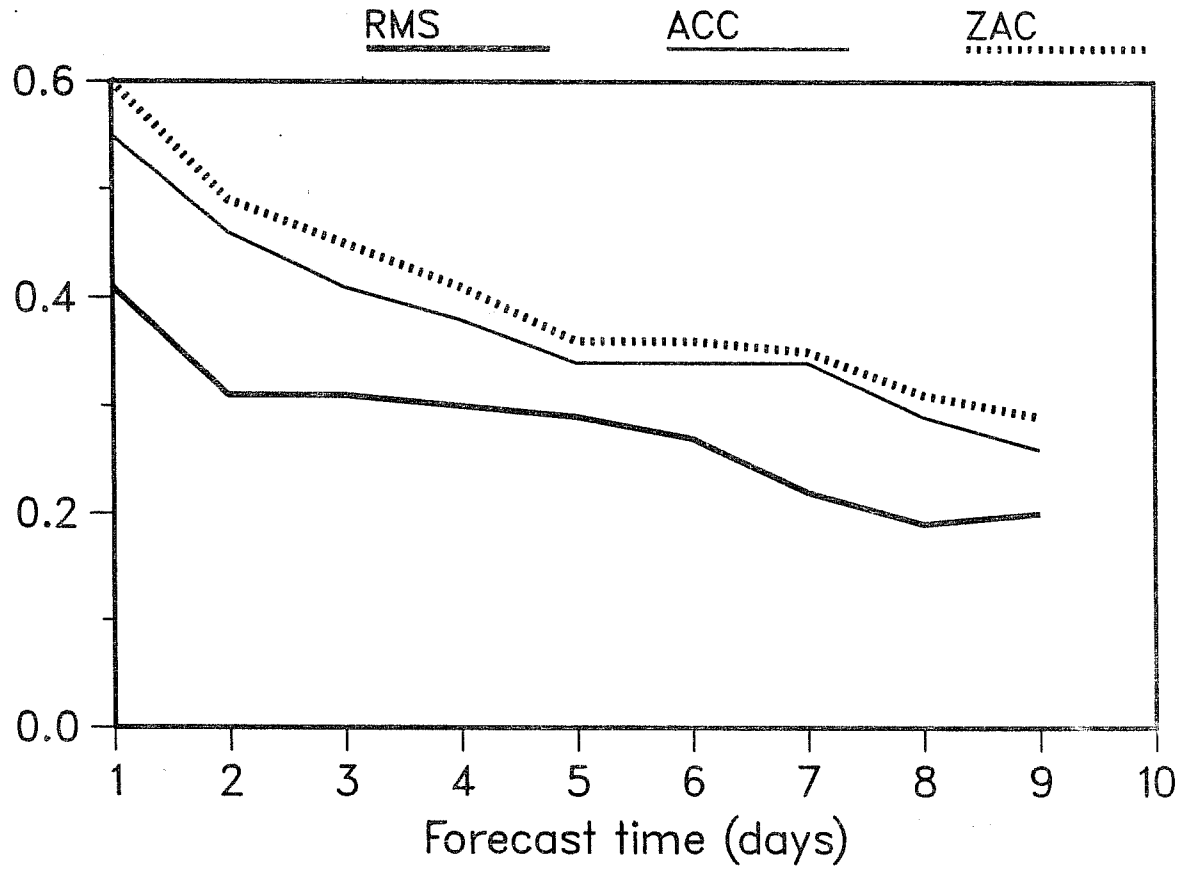

Fig. 3 Correlation between the hemispheric skill of a day n forecast, and the hemispheric difference between a day n and a day (n+1) forecast, both verifying at the same time. RMS, ACC and ZAC measures of spread and skill shown. For each kind of measure of skill the corresponding measure of spread has been used.

(unbounded) z-transform measure ZAC. It can be seen that far from reducing the correlations, they have been enhanced. We must therefore seek other explanations for the apparent enhanced skill of the phase measure of forecast spread.

It is known that, as a measure of short and medium range skill, the anomaly correlation coefficient is strongly dependent on the anomaly amplitude (Arpe et al., 1985; Branstator, 1987). A corollary of this is that the enhanced correlation between spread and skill using anomaly correlation may, in part at least, arise because spread and skill are mutually correlated with the magnitude of the forecast (or observed) anomaly. Fig 4 shows a schematic diagram illustrating two hypothetical ensembles of forecasts whose dispersal in phase space relative to their initial conditions, is essentially identical. A typical one dimensional cut through this phase space is shown. Relative to its RMS dispersion, the magnitudes of the anomalies of the members of the first ensemble are large. Hence, in terms of anomaly correlation, the spread between the forecasts themselves, and between forecasts and verifying analysis, will be small (i.e. high skill, low spread). For the second ensemble, whose members straddle the climatology line, the anomaly correlation between some of the members within the ensemble could even be negative, as could the anomaly correlation between some members of the ensemble and the verifying analysis. Hence for this ensemble, skill will be low and spread high. Comparing the two ensembles, we have a clear correlation between ACC spread and skill.

Fig 5 shows the (600 day) correlation between hemispheric spread (defined in terms of ZAC) and the RMS magnitude of the forecast anomaly (full line), and the correlation between the hemispheric skill (also defined in terms of ZAC) and the RMS magnitude of the forecast anomaly (dashed line). As anticipated by the above argument, both curves show significant correlation, particularly at the beginning of the forecast period. Note that the RMS magnitude of the forecast anomaly is available once the forecast has been produced, and is therefore a viable predictor of forecast skill. Indeed, comparing Figs 3 and 5, it can be seen that the magnitude of the forecast anomaly is a more skilful predictor (of phase measure skill) than the spread between adjacent forecasts. Essentially similar results were obtained (see Palmer and Tibaldi, 1986) when ACC was calculated defining the forecast anomaly in terms of deviation from observed climate.

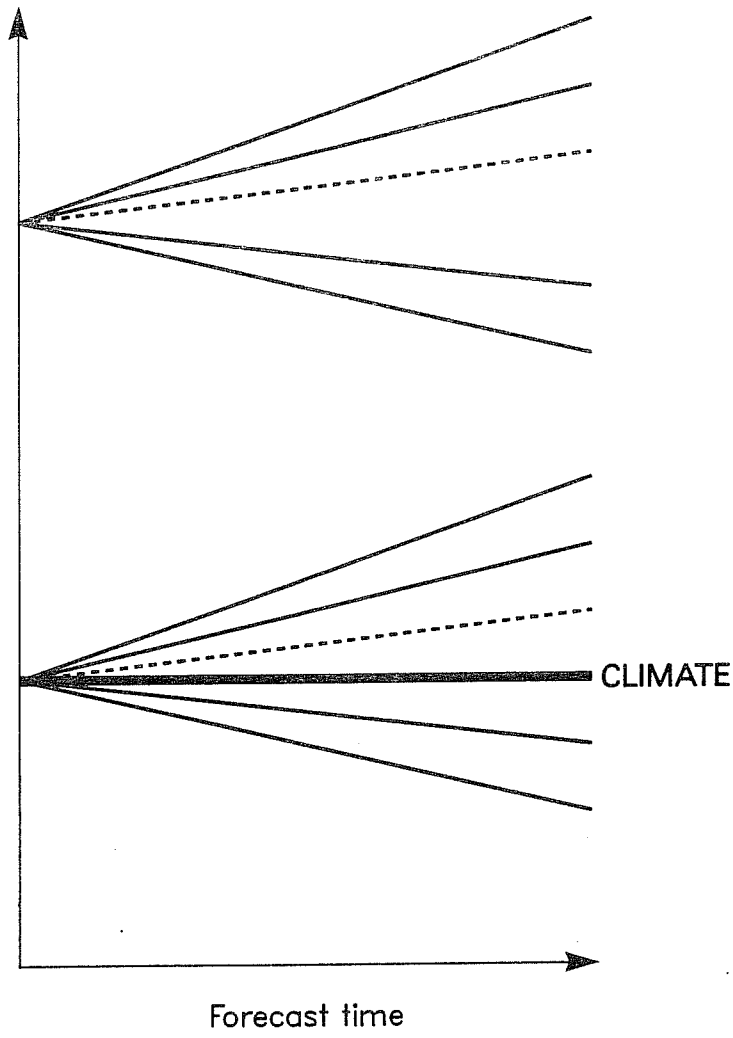

Fig. 4 Schematic illustration of the dispersion of two ensembles of forecasts (shown by the solid line) relative to their verifying analyses (dashed line). From an RMS point of view the ensembles have identical dispersion. It is argued in the text that using ACC, there will be a correlation between the skill and spreads of such ensembles on account of their dependence on the magnitude of the forecast anomaly.

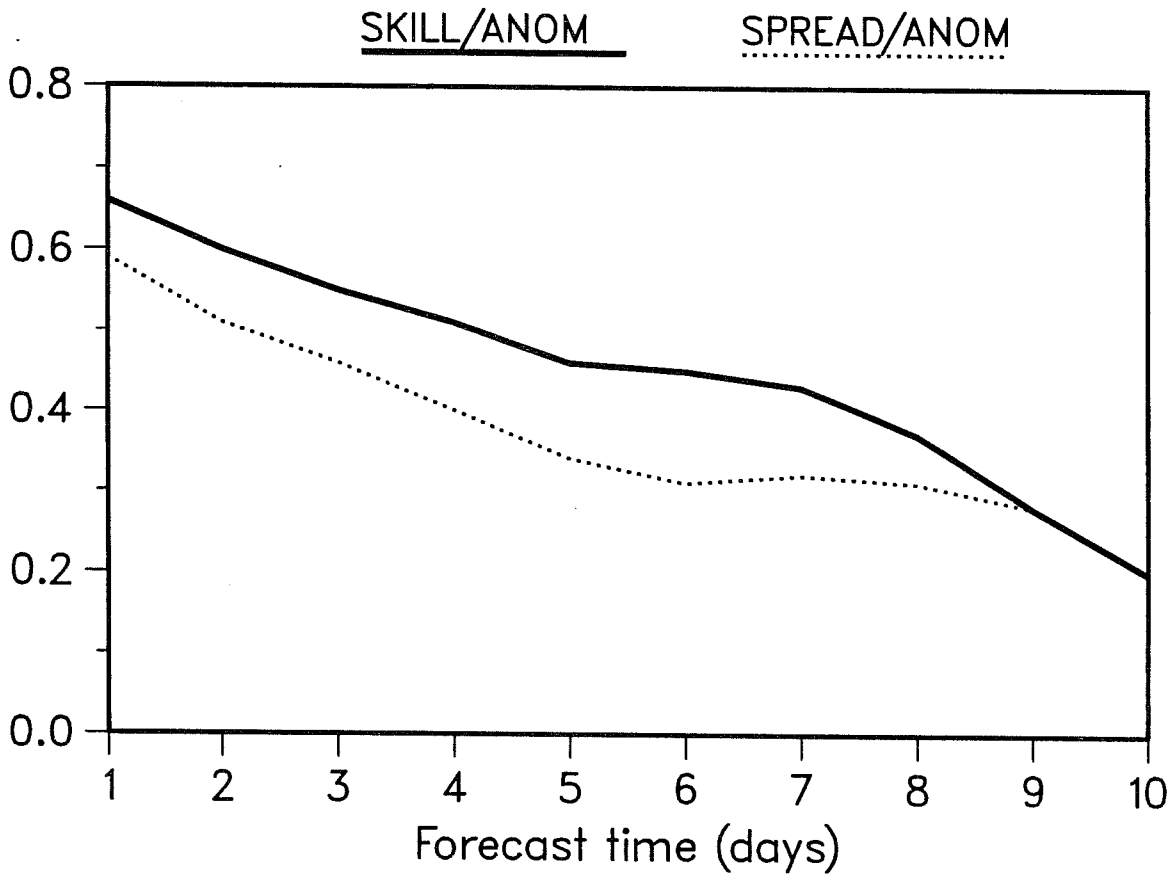

Fig. 5 Correlation between a) the day n hemispheric RMS error, and the RMS hemispheric difference between the day n forecast and climate (curve marked skill), and b) between hemispheric forecast spread and RMS magnitude of forecast anomaly (curve marked spread).

Of course it might be queried whether RMS spread and RMS skill are similarly correlated to some simple bulk measure of the amplitude of the flow. On the seasonal timescale they certainly are correlated with the spatial standard deviation of geopotential height. (So, for example, RMS spread and RMS skill will tend to be relatively large in winter, and relatively small in summer). To test whether this correlation is important on daily data within the winter season, we have calculated the 600 day correlation between daily hemispheric forecast standard deviation of 500 mb height and firstly RMS error, and secondly RMS spread. For day 1 the correlation with RMS error is 0.16, and with RMS spread 0.09. For later forecast days the correlations are smaller. Similarly (see Palmer and Tibaldi, 1987) the magnitude of the forecast anomaly is very poorly correlated with RMS spread and (particularly at the beginning of the forecast period) poorly correlated with RMS error. We therefore conclude that the RMS measure is more suitable to study spread/skill relationships, and we restrict ourselves to this measure throughout the rest of the paper.

(It should be mentioned that we have found that the magnitude of the spread skill correlations using ZAC are not significantly larger using regional rather than hemispheric estimates. This would appear to be in contradiction to Kalnay and Dalcher's results. However, if a much smaller sample is chosen, just one year for example, then the correlation between skill and the magnitude of the anomaly is somewhat larger for limited areas than for the hemisphere. For example, using data from only the first year of our sample, the day 1 hemispheric skill correlates 54% with the magnitude of the hemispheric anomaly, whereas in region 1 the regional skill correlates 71% with the magnitude of the regional anomaly. We agree with Kalnay and Dalcher's opinion that this simply reflects the fact that, over a limited time series, the magnitude of regional anomalies will fluctuate much more strongly than the magnitude of hemispheric anomalies. Kalnay and Dalcher's forecasts were all taken during a one month period during 1979, therefore their weaker results for hemispheric measures reflects the relatively small variability in the magnitude of hemispheric anomalies during that period.)

In conclusion, therefore, it appears that the spread estimator based on anomaly correlation coefficient, is not an entirely appropriate estimate of ensemble dispersion in the sense envisaged above. Its skill in predicting

anomaly correlation is significantly influenced by the relation between such a measure and the departure of the forecast from climatology. Moreover, if anomaly correlation is nevertheless the required measure of skill to be predicted, the magnitude of the forecast anomaly itself would be a more skilful predictor.

Fig 6 shows four curves describing the correlation between (RMS) spread and skill throughout the forecast period: the thin lines show calculations performed with daily unfiltered data, the bold lines show calculations on data which had been filtered by the 5-day running mean. The full lines show the correlation between the skill of a day n forecast and the spread between that day n forecast and the day n+1 forecast verifying at the same time. We call these 'prognostic correlations' since this spread indicator is available 'operationally ' as a predictor of forecast skill. The dashed lines show the correlation between the skill of the same day n forecast and the spread between that day n forecast, and the day n-1 forecast verifying at the same time. We call these 'diagnostic correlations', since they are clearly not available operationally.

It can be seen that diagnostic correlations are larger than prognostic correlations. One way to understand why this should be is to note that 'tomorrow's' day n-1 forecast is, in general, a more skilful representation of reality than 'yesterday's' day n+1 forecast. Hence the diagnostic RMS spread should, in general, be a more faithful representation of the RMS forecast error than the prognostic spread. Put another way, given the spread between today's and yesterday's forecast, it is easier to predict the skill of yesterday's forecast than today's forecast.

It could be argued that it may be worth waiting a day in order to have the more reliable information supplied by the prognostic spread; however, it can be seen that the only substantial improvement in skill of the diagnostic spread over the prognostic spread is in the first few days of the forecast period. The loss of forecast skill incurred in the short range by 'waiting a day' may be unacceptable.

There is an interesting corollary to this result relevant to lagged average forecasting. In the short and medium range, the optimal lagged average

ensemble would be obtained by some weighted combination of its members. Forecasts from the most recent initial conditions would have the largest weights. Hence for an optimally weighted forecast, the ensemble mean initial conditions are weighted towards the more recent initial conditions. Consider for simplicity, a two member ensemble. By the above argument, the ensemble spread at any verification day, will correlate more strongly with the skill of the earlier forecast than with the skill of the later forecast. Hence we would expect that the spread will correlate more strongly with the skill of the uniformly weighted ensemble forecast, than with the skill of the optimally weighted ensemble forecast. Again this demonstrates the principle that, in order to improve ones a priori estimate of forecast skill using the spread indicator, it is necessary to operate in a forecasting environment in which forecast skill itself is not optimised. Whether anything is gained in the long run by taking such an approach would need a careful 'trade-off' analysis.

It is interesting to note that (see Fig. 6) the 5-day mean filter has a comparatively larger effect on prognostic correlations than on diagnostic correlations. This suggests that the high frequency (day-to-day) fluctuations in skill and diagnostic spread are better correlated than the high frequency fluctuations in skill and prognostic spread. This can be better understood if we think of the comparison between today's forecast and tomorrow's forecast as a 'perfect model' predictability experiment (Lorenz, 1982), where the diagnostic spread (RMS difference) between today's day n forecast and tomorrow's day n-1 forecast (that verify on the same day) is to be interpreted as the consequence of the growth of the day 1 error of today's forecast. (Remember that the day 1 diagnostic spread is nothing but today's day 1 forecast RMS error). If we take the day 1 forecast error as a proxy for the analysis error, it would be reasonable to interpret the high frequency fluctuations in diagnostic spread as indicative of fluctuations in the quality of today's analysis. On the contrary, correlations in the high frequency component of today's analysis error and yesterday's analysis error will in general, be small, so that high frequency fluctuations in prognostic spread would not be expected to correlate well with today's analysis error. This is another way to understand why diagnostic correlations are generally larger than prognostic correlations (see above).

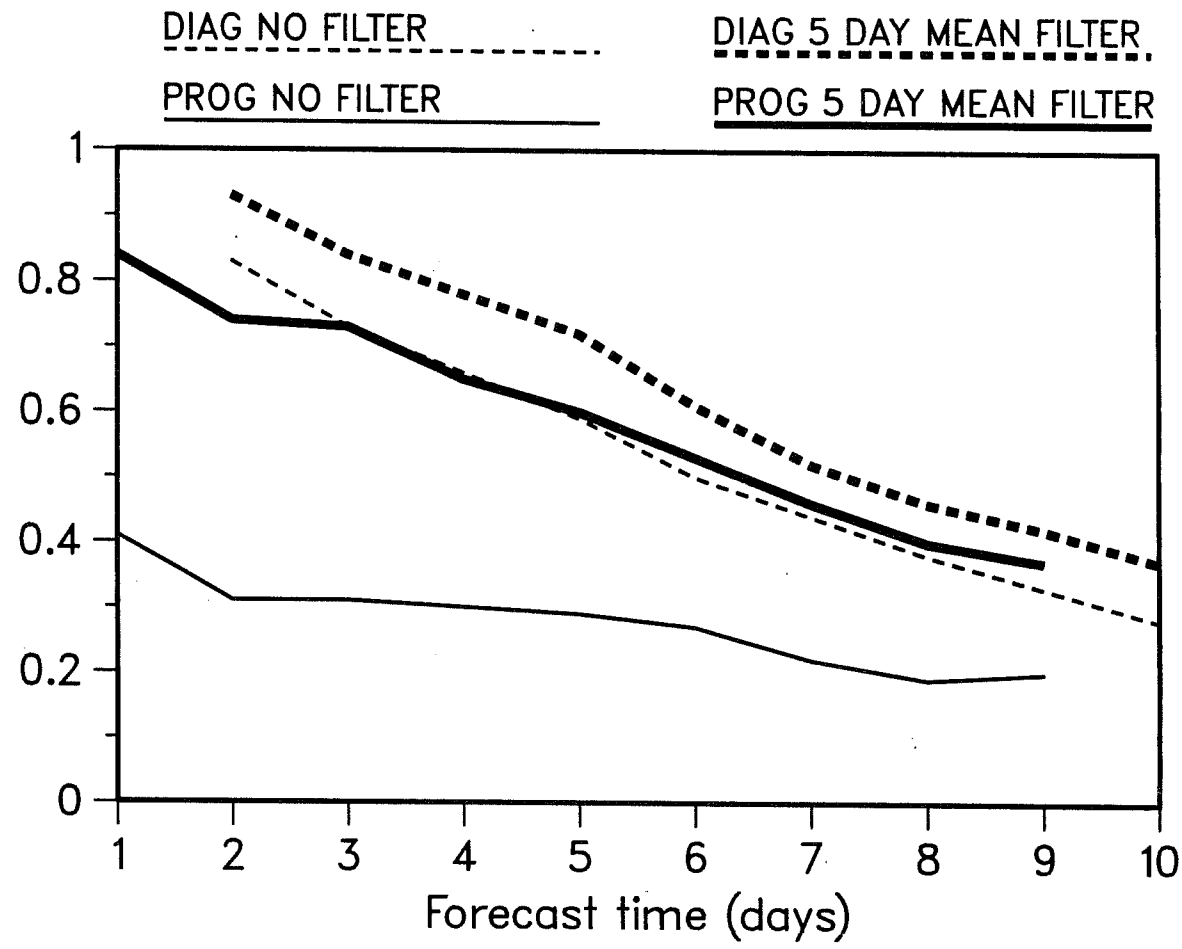

Fig. 6 Correlation between RMS skill and RMS spread. Full lines show prognostic correlations, dashed lines show diagnostic correlations (see text for definition of these terms). Bold lines show correlations after a five day running mean temporal filter has been passed through the skill and spread data.

So far we have considered mainly hemispheric relationships. Fig 7 shows the spread-skill relationships for the twelve regions at day 6 of the forecast. In fact, it should be mentioned that correlations were calculated between (day 6/day 7) spread in some region i and (day 6) skill in all other regions j (i,j=1...12), both contemporaneously, and with all possible time lags. The results showed invariably that the highest correlations are to be found between quantities computed on the same limited area and at the same time, (however, see section 3.3 for possible upstream effects). Therefore, all results below are for contemporaneous correlations between spread and skill in the same area. In Fig 7 it can again be seen that the prognostic correlations are sensitive to the temporal filter, whereas the diagnostic correlations are not. Neither prognostic nor diagnostic correlations are strongly sensitive to region, though the prognostic spread shows somewhat higher correlations for regions near the end of storm tracks.

3.2 Flow pattern predictors

A conceptually different predictor to forecast spread is given by the synoptic features of flow patterns associated with the forecast height field. Forecast quality may depend on the current flow regime. Gronaas (1985) studied the skill of the ECMWF model during subjectively defined blocking and non-blocking spells. The question we wish to address is how to describe objectively the modes of atmospheric flow variability associated with forecast skill fluctuations. To answer this we consider a linear regression of forecast skill against the EOF coefficients of either the 500mb height field of either the forecast or initial conditions, or both. Some aspects of this study have been given by Palmer (1988) in relation to forecast skill dependence on the PNA mode, but a more extensive analysis is given below. In the following we use either the N=23 truncation of EOF coefficients or the reduced (N=9) set comprising the first 3 zonal EOFs and the first 5 eddy EOFs.

In order to fix notation, we give below a basic description of the algebra of linear regression analysis applied to the EOF coefficient predictor sets described above.

Let $1 \leq \alpha \leq 600$ denote a daily index over the sample of training data (600 wintertime forecasts from 1980/86). In addition, let M denote the number of predictors used in the regression, and N the truncation of the EOF expansion.

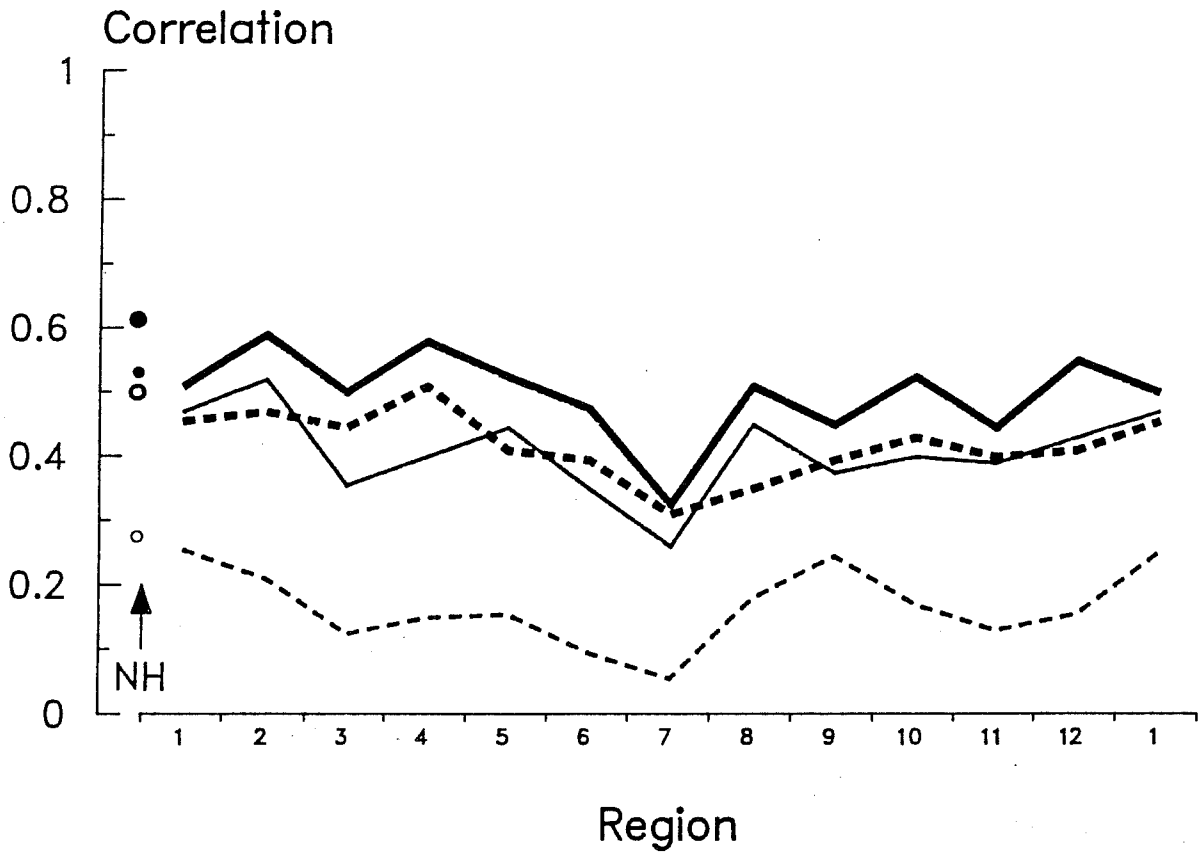

Fig. 7 Correlation between day 6 forecast skill and both prognostic (day 6-7) and diagnostic (day 5-6) forecast spread. Bold lines and large dots: diagnostic spread. Thin lines and smaller dots: prognostic spread. Full lines and full dots: 5 day running mean filtered data. Dashed lines and circles: unfiltered data. The dots on the left represent hemispheric values.

Furthermore, for day α, let $e_{j\alpha}$ denote the j^{th} EOF coefficient, $1 \leqslant j \leqslant M$, of either the forecast flow (M=N), initial flow (M=N) or both forecast and initial flows (M=2N). The hemispheric or regional score for that day will be denoted by S_α. We shall assume in the following that for fixed j $\{e_{j\alpha}\}$, $\{S_\alpha\}$ have been normalised and standardised, i.e. if $\langle\ \rangle$ denotes a mean over α,

$$\langle S_\alpha \rangle = \langle e_{j\alpha} \rangle = 0$$

$$\langle S_\alpha^2 \rangle = \langle e_{j\alpha}^2 \rangle = 1$$

In linear regression theory one finds a set of weights a_j such that

$$Q = \langle (\sum_j a_j\, e_{j\alpha} - S_\alpha)^2 \rangle$$

is minimised. The condition $\partial Q / \partial a_j = 0$ straightforwardly gives these weights as

$$a_j = \rho_i C_{ij}^{-1}$$

where

$$\rho_j = \langle S_\alpha e_{j\alpha} \rangle$$

$$C_{ij} = \langle e_{i\alpha}\, e_{j\alpha} \rangle$$

The quantity

$$r_\alpha = \sum_j a_j\, e_{j\alpha}$$

is the regressed or 'predicted' skill score given a forecast with EOF coefficients $e_{j\alpha}$. The sample-mean correlation between the real and regressed skill score is

$$\langle r_\alpha S_\alpha \rangle = \sum \rho_i C_{ij}^{-1}\, \rho_j \equiv \rho$$

The so-called factor structure constants ρ_j give the correlation between the scores S_α and each EOF coefficient $e_{j\alpha}$. Forecasts with EOF coefficients proportional to ρ_j are most strongly correlated with variations in forecast skill. Alternatively, writing

$$\rho_j = a_j\, C_{ij}$$

then a forecast field with EOF coefficients equal to ρ_j can be thought of as representative of the synoptic situations associated with the regression weights, taking climatological correlation between EOF coefficients into account.

A number of different regressions have been calculated as outlined below. The regressions with the largest number of predictors (M=46) used the full set of EOF coefficients on both the forecast flow (for different forecast days) and on the initial conditions. In addition, regressions were performed using the EOF coefficients of the forecast flow only, or the coefficients of the initial conditions only. Regressions were also run with the reduced set (N=9) of coefficients, using forecast and initial conditions, or forecast or initial conditions separately. Finally, some regressions were performed on the temporally filtered training data.

For reasons of space, it will not be possible to give full details from all these calculations. However, some results from the various regressions are recorded below. For example, regression correlations for day 9 forecasts with different values of N and M are recorded in Table 1 for the twelve regions and the whole hemisphere using RMS error as the measure of skill. With forecast and initial data (column 1) and no temporal filtering, there are three regions with correlation coefficient larger than .5: regions 1, 8, and 12. The fourth highest limited-area correlation is for region 9. As noted in the introduction, these 4 regions are distinguished dynamically in having relatively large low-frequency atmospheric variability, and relatively small band-pass variability. Using only data from the initial analysis (column 2), the correlations are considerably smaller, generally around .3. Correlations using only the 23 EOF coefficients of the forecast flow (column 3) are larger than those using only the initial data, and in region 1 only 6% smaller than the correlation using the combined forecast and initial analysis data. The fourth column shows correlations using 9 EOF coefficients from the forecast data only. In region 1, the correlation is only 3% smaller than that for the regression using the full 23 EOF coefficients. However, in the adjacent region 12, there is a 12% drop in the magnitude of the correlation, and in general the drop in correlation in different regions is significant. In column 5 we show correlations using forecast data only, with 23 EOF coefficients, but where the 5-day running mean filter has been applied to the data before the regression analysis. Correlations are not only higher than those using 23 predictors on daily unfiltered data, but also higher than those in the first column with 46 predictors (with the exception of region 8 where the correlation coefficients are equal). Correlations with 5-day filter and M=46 (column 6) are larger again, though generally the increase in correlation

Table 1

Region	1 day 9 M=46 N=23	2 day 9 M=N=23 (initial EOFs)	3 day 9 M=N=23 (forecast EOFs)	4 day 9 M=N=9 (forecast EOFs)	5 day 9 M=N=23 (forecast EOFs 5-day filter)	6 day 9 M=46 N=23 (5-day filter)	7 day 3 M=46 N=23	8 day 3 M=N=23 (forecast EOFs)	9 day 3 M=N=23 (forecast EOFs 5-day filter)
1	.54	.29	.48	.45	.62	.66	.55	.48	.63
2	.46	.28	.40	.23	.54	.58	.54	.44	.60
3	.43	.21	.33	.20	.51	.53	.46	.37	.50
4	.43	.31	.27	.16	.44	.54	.50	.36	.45
5	.45	.38	.31	.12	.52	.63	.53	.39	.51
6	.40	.27	.29	.23	.47	.54	.44	.35	.49
7	.46	.33	.38	.29	.55	.58	.42	.31	.46
8	.52	.32	.42	.34	.52	.61	.47	.42	.54
9	.47	.37	.37	.29	.58	.63	.44	.36	.49
10	.38	.27	.29	.15	.47	.51	.42	.35	.50
11	.40	.25	.29	.16	.42	.52	.36	.31	.47
12	.54	.32	.43	.32	.57	.62	.39	.34	.51
NH	.49	.33	.34	.29	.42	.59	.56	.45	.56

Table 1 Correlation coefficients between regressed and actual skill scores for 12 regions based on wintertime forecast data from 1980/81 to 1985/86. Columns 1-6 are for day 9 forecasts, column 7-9 are for day 3.

between columns 5 and 6 is not as large as between 3 and 5. That is to say, a larger increase in correlation obtains when the data is temporally filtered, than when the number of predictors is enhanced. Overall, this suggests that the regression analysis is defining flow patterns correlated principally with low-frequency atmospheric variability.

The 500mb height anomaly patterns corresponding to the factor structure constants ρ_i for the four regions 12,1,8 and 9 are shown in Fig 8-12. Such patterns were derived from the formula

$$z\ (x,y) = \frac{1}{\rho} \sum_{1}^{N} \rho_i\ E_i\ (x,y)$$

where $E_i(x,y)$ is the ith EOF. If a particular forecast flow was given by this z(x,y), its EOF coefficients would equal $\frac{1}{\rho}\ \rho_i$ and its regressed skill scores would be one standard deviation worse than average. This sets the amplitude of these plots. Fig 8 shows the anomaly pattern corresponding to regressions with forecast EOF coefficients only (M=23,N=23) for regions 12, 1, 8 and 9. In synoptic terms, the pattern for region 1, for example, can be interpreted synoptically by saying that if in the day 9 forecast there is anomalous north westerly flow over the NW Atlantic, with an anomalous trough over central Europe, then the forecast will be relatively poor over Europe. The pattern is somewhat different for region 12, where a geographically concentrated trough to the west of the British Isles is associated with relatively poor forecasts.

The patterns for regions 8 and 9 (Fig. 8c, d), on the other hand, are fairly similar to each other; they indicate that forecast skill is correlated with the signed amplitude of the Pacific/North American (PNA) mode of low-frequency variability of the atmosphere (Wallace and Gutzler, 1981). When there is strong troughing over the Aleutian Islands, and strong ridging over the Rockies, forecast skill will be relatively high. This association of forecast skill with the PNA mode has been discussed in Palmer (1988), where it was shown that forecast skill in the extended range also strongly correlated with the signed amplitude of the PNA mode. It should be mentioned that there is a weak correspondence between the patterns for both regions 12 and 1, and Wallace and Gutzler's teleconnection patterns for the Euro/Atlantic sector. However, these are much less clear than the PNA correspondence with patterns for regions 8 and 9.

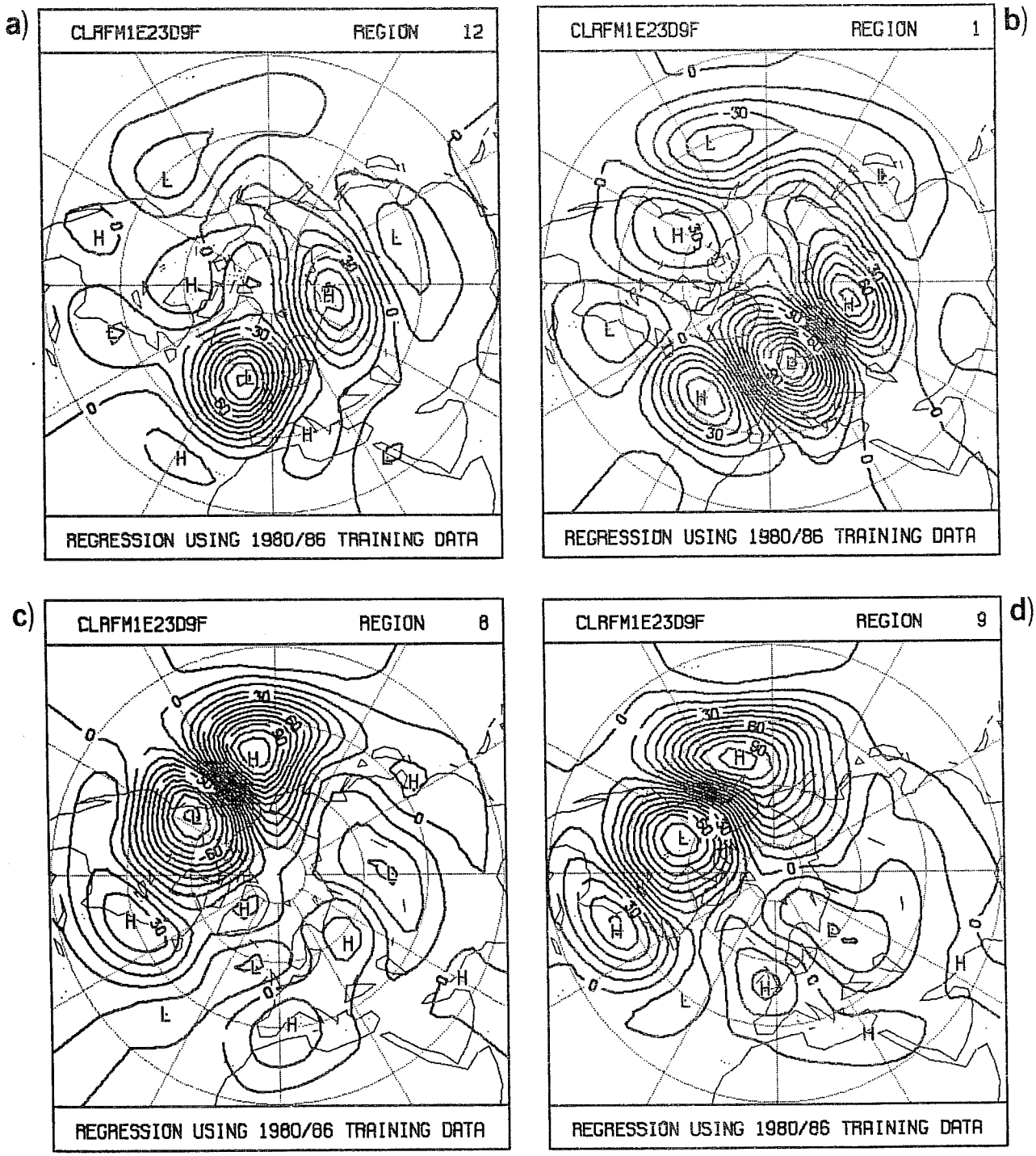

Fig. 8 Factor structure constants expressed as 500 mb height anomaly maps (geopotential metres) from a regression between day 9 RMS error, and day 9 forecast EOF coefficients. a) Region 12, b) Region 1, c) Region 8 and d) Region 9. The amplitude of the pattern is such as to correspond, in the regression analysis, to a skill score one standard deviation worse than average.

For region 1, by far the largest eddy weight in the regression analysis is for forecast eddy EOF number 5. On the other hand for region 12 forecast eddy EOFs 5 and 7 have comparable weights, and for region 8 forecast eddy EOFs 6 and 10 have the largest of the eddy weights. This is consistent with the fact (noted above) that with a reduced set N=9 of predictors comprising 3 zonal and 5 eddy EOFs, the correlations for region 1 are barely reduced over the full set, whereas for the other regions (12, 8 and 9 in particular) there appears to be a loss of correlation.

The 500mb height anomalies associated with the factor structure constants for regions 12, 1, 8 and 9 for the regressions with initial data only (M=23, N=23) are shown in Fig 9. We have already noted that the correlation between regressed and actual scores is smaller using initial rather than forecast EOFs. In addition to this, the factor structure constants have some differences with their counterparts in Fig 8. In particular, the localised anomaly centre to the west of the British Isles in Fig 8a for region 12, is absent in its counterpart in Fig 9a, and overall the pattern is less geographically localised. The pattern for region 1 is qualitatively similar, but the position and magnitude of the anomaly centres differ between Figs 8 and 9b. For regions 8, and 9, there is still evidence of the dependence of forecast skill on the signed amplitude of the PNA pattern in the initial analysis, but again the patterns are less geographically localised than their counterparts with forecast EOFs.

The 500mb height anomalies associated with the factor structure constants for regions 12, 1, 8 and 9 for the regressions with both initial and forecast data (M=46, N=23) are not shown for reasons of space. However, the patterns for the forecast anomalies and initial condition anomalies are very similar to those shown in Figs 8 and 9 respectively, though the amplitude of the patterns corresponding to the initial conditions are much smaller than those shown in Fig 9. This is consistent with the fact that the regression weights for the initial conditions are relatively small compared with the weights for the forecast flow.

The 500mb height anomalies associated with the factor structure constants for regions 12, 1, 8 and 9 for the regressions described in column 4 of Table 1 are shown in Fig 10. With 9 forecast EOF coefficients only (M=9, N=9), the

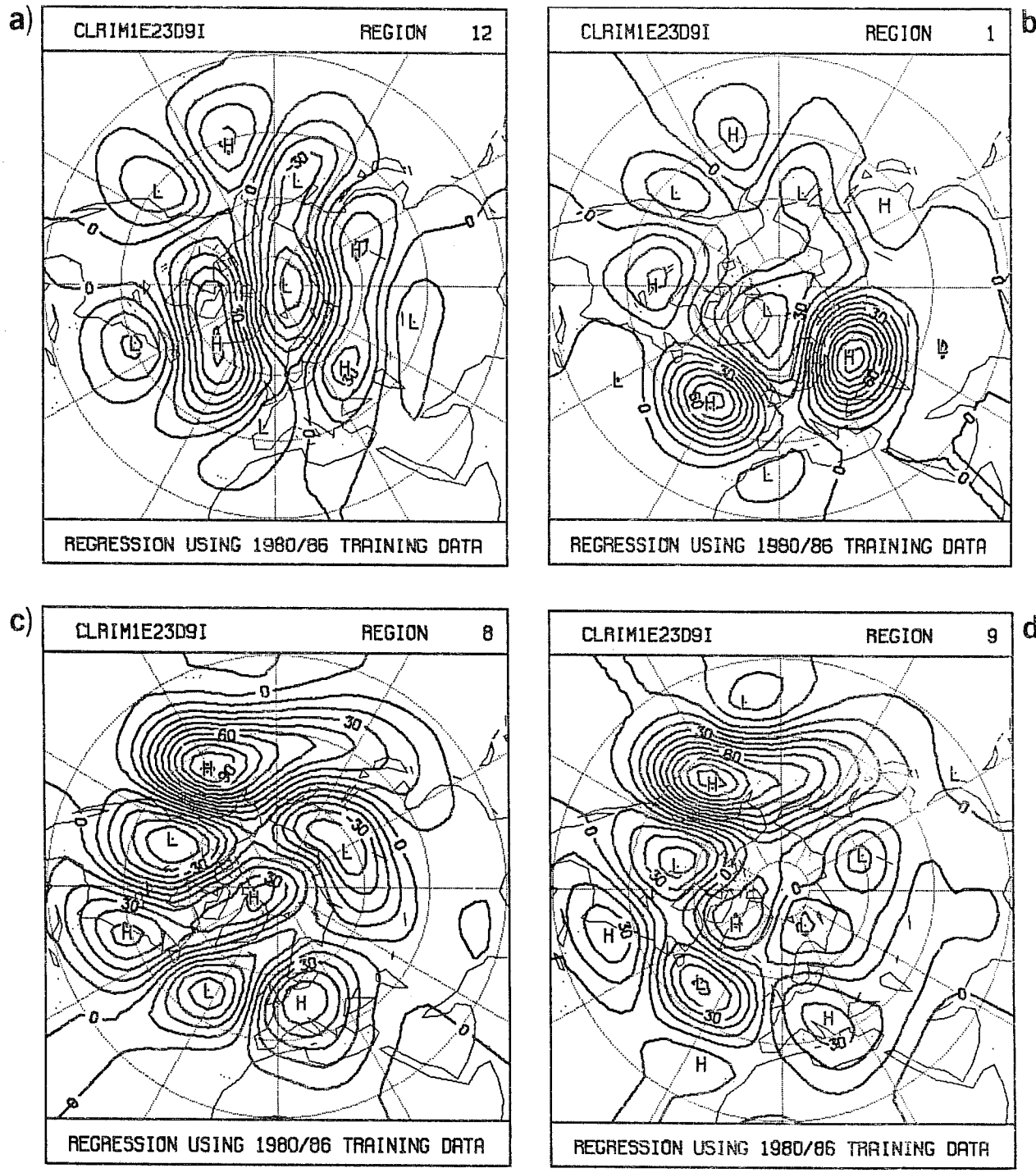

Fig. 9 As Fig. 8 but for a regression between day 9 RMS error and EOF coefficients from the forecast initial conditions.

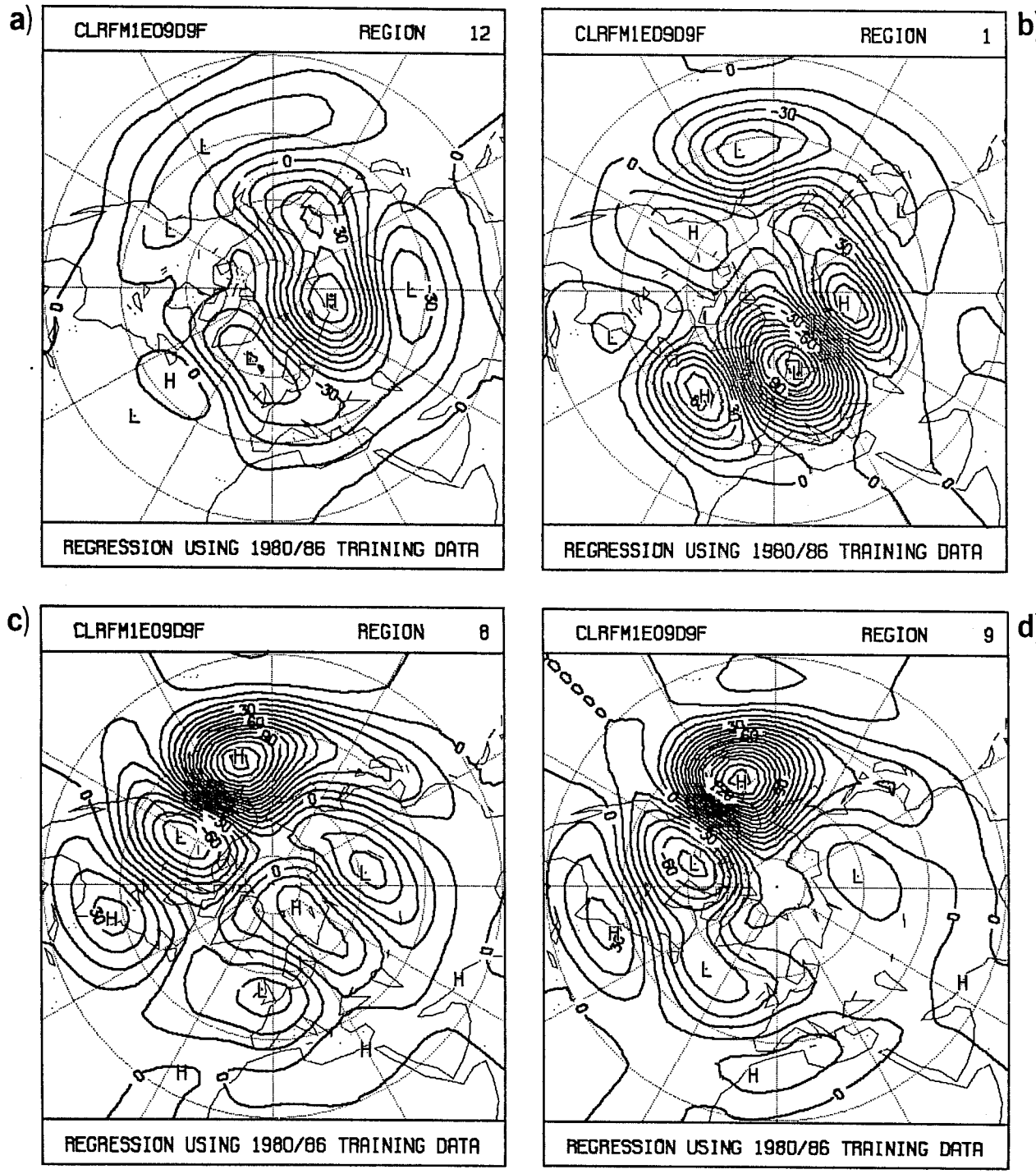

Fig. 10 As Fig. 8 but using an EOF truncation retaining 3 zonal EOFs and 5 Eddy EOFs.

pattern in region 1 is very similar to that shown in Fig 8 using all 23 EOF coefficients. This, of course, is consistent with the fact just mentioned that eddy EOF number 5 has the ovewhelmingly largest weight in the regresion. Similarly consistent is the fact that for region 12 in particular, and 8 and 9 to a lesser extent, the regression with 9 EOFs does not give rise to a geographically well-localised 500mb height pattern.

The forecast flow anomaly patterns using 5-day mean filtered data (not shown) are very similar to those using the daily unfiltered data, except that the amplitude of the patterns are much smaller (since standard deviations of 5 day mean fields are smaller). One might infer from this that the basic regression itself acts as a filter on the daily data, only picking out low-frequency modes of variability of flow. This is consistent with an interpretation of some of the patterns in Fig 8, for example, as modes of model-atmosphere low-frequency variability.

We can make a number of conclusions based on these results. Firstly, the relative smallness of the correlations using initial data only, and the relative smallness of the amplitude of the anomaly patterns for the initial data in the M=46 regression, together with their relatively poor geographical localisation, suggests that the initial conditions may be of marginal use in the regression analysis. Indeed, with the full set of 46 EOFs we run the risk of over-fitting when the regression weights are applied to an independent set of data. We shall test this when we attempt to predict the forecast skill for the independent winter 1986/87. Secondly, with the exception of region 1, the 'delocalisation' of the anomaly patterns using 9 EOFs, suggests that this small number may give a poor prediction of skill when the regression is applied to independent data. Hence for a given number of predictors, it may be preferable to increase the detail of the representation of the forecast flow, than include aspects of both forecast and initial conditions. However, results have shown that there is no obvious gain to be had by pre-filtering the data temporally, rather than temporally filtering the output of the regression analysis (that is to say, the operation of time filtering would appear to commute with the regression analysis).

Before concluding this section, we give some brief results from two further sets of regressions, one for day 9 hemispheric scores, and the other for

regional day 3 scores. The correlations between actual and regressed hemisphere scores are given in Table 1. Fig 11 shows the anomaly pattern associated with the factor structure constants from a regression of the set of 23 forecast EOF coefficients (only) against day 9 hemispheric skill scores. The magnitude of the correlation between regressed and observed skill scores is equal to 0.34, and this is close to the mean correlation coefficient for the twelve regional regressions in the third column of Table 1. The anomaly pattern appears to be composed of a combination of patterns already discussed above; ie the patterns in Fig 8 for regions 8, 12 and 1. In particular the PNA mode appears to be correlated with hemispheric skill scores, albeit less emphatically than with regional scores over the Pacific/North American region.

Finally, in the columns 7-9 of Table 1, we show some correlations derived from regressions for day 3 of the forecast. The regressions are for (M=46, N=23), and (M=N=23), both temporally filtered and unfiltered. Comparing columns 7 and 1, 8 and 3, 9 and 5, it can be seen that of the 4 regions, 1, 12, 8 and 9, described above, day 3 correlations are not consistently larger than day 9 correlations. This gives some support to the contention mentioned in the introduction that just because mean day 9 skill is lower than mean day 3 skill, it does not follow that it will be more difficult to predict day 9 skill than to predict day 3 skill (at least using the EOF predictor sets).

Anomaly maps showing the forecast factor structure constants for day 3 are shown in Fig 12. These should be compared with Fig 8 for the day 9 forecasts. It can be seen that whilst the pattern for region 1 is very similar to that for day 9, the patterns in the other regions are quite different. In particular, for regions 8 and 9 there is no evidence of the PNA mode. It can also be seen, and this is true for all the day-3 patterns for the other regions, that over the region of interest there is a negative height anomaly. Whist this is true for some of the patterns at day 9, it is by no means always true, as the pattern for region 8 in Fig 8 illustrates. The possible significance of this is discussed in section 5.

3.3 Skill of the short range forecast

The quality of the initial analysis is an important influence on day to day variations of forecast skill. Since we have no direct knowledge of analysis

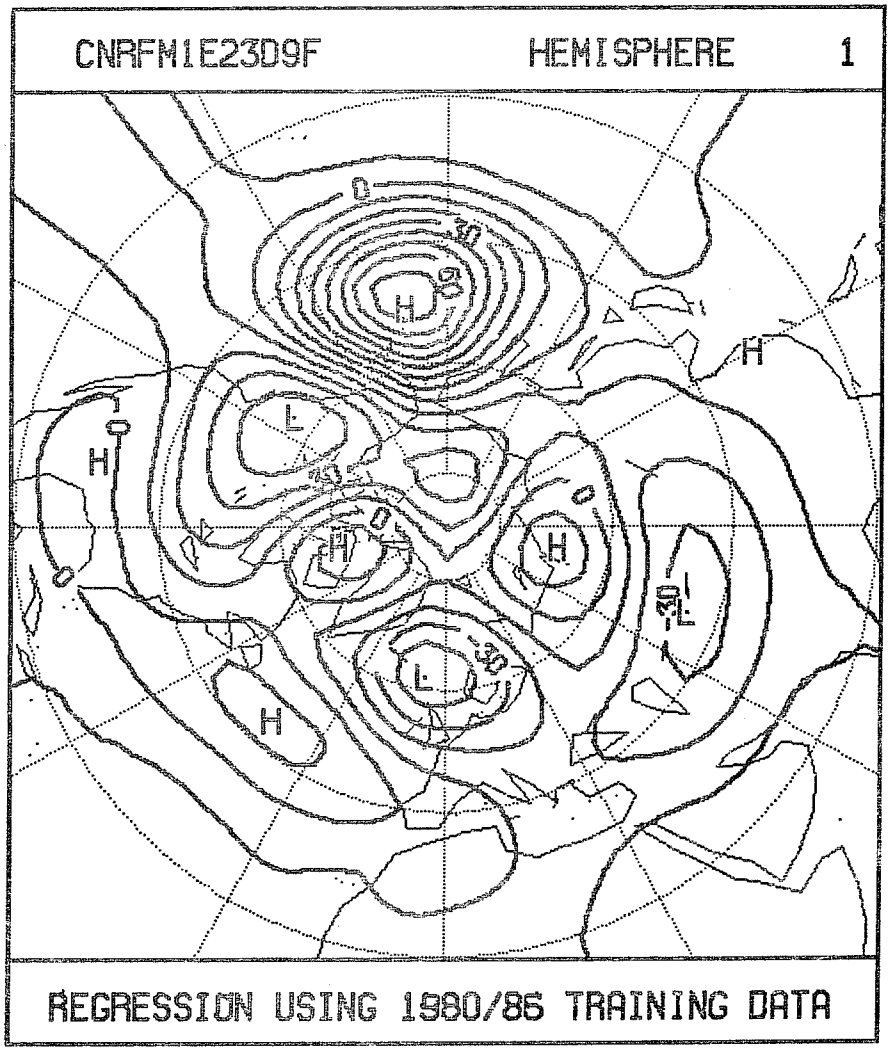

Fig. 11 As Fig. 8 but for a regression with day 9 hemispheric RMS error as predictand.

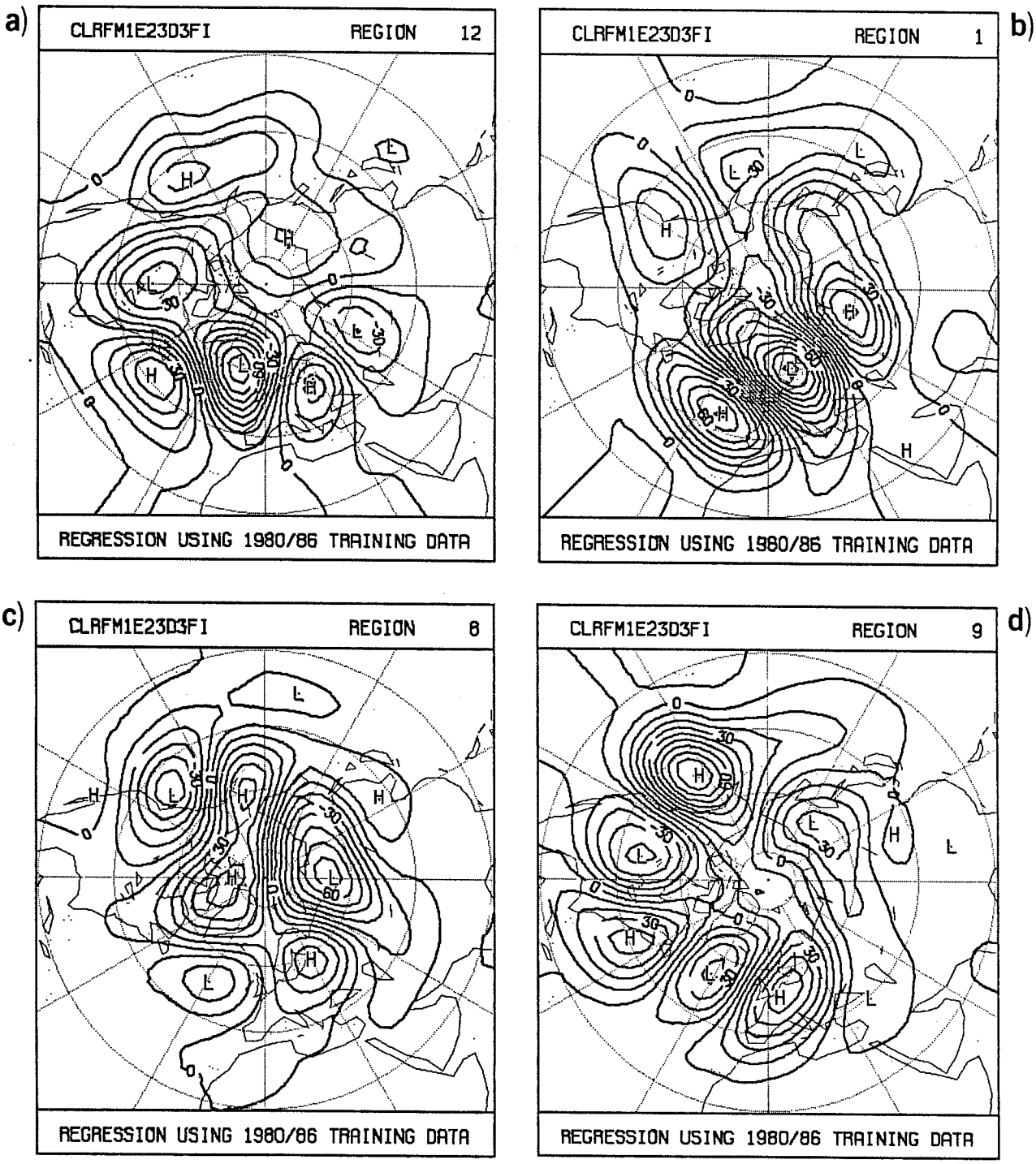

Fig. 12 As Fig. 8 but for day 3 RMS error and day 3 forecast EOF coefficients.

errors in our dataset, we must look for proxies. As discussed above, the day 1 forecast RMS error of the current forecast might serve as just such a proxy, though clearly this is not available at initialisation time. Again, as above, we refer to correlations between the skill of 'today's' forecast at day n>1 with the skill of the same forecast at day 1, as diagnostic. A 'very poor man's' analysis error for today's forecast is provided by yesterday's day 1 forecast error, in general agreeing with today's day 1 forecast error only in the low-frequency component. Correlations between today's day n forecast and yesterday's day 1 forecast are, in line with earlier terminology, prognostic. As with correlations between spread and skill, we show in Fig 13 the prognostic and diagnostic correlations between hemispheric RMS day n skill and hemispheric RMS day 1 skill. Results are qualitatively similar to those for spread/skill correlations shown in Fig 6. In particular, it can be seen that the diagnostic correlations are larger than the prognostic correlations, and that, in the short range at least, the 5-day mean filter has a more substantial effect on the prognostic correlations than on the diagnostic correlations.

The diagnostic correlation as a function of limited area is shown in Fig 14 for day 6 of the forecast. Two things appear clearly: the three limited areas for which such correlation is highest are limited areas 1, 2 and 9, and the correlation itself is significant in these areas only after the 5-day mean filter has been applied. Fig 4 showed the location of such limited areas, together with the climatology of the band-pass eddy activity during the Northern Hemispheric winter. We therefore deduce that short-range forecast skill is an indicator of medium-range forecast skill only in those areas of the Northern Hemisphere where baroclinic eddy activity is relatively low.

In Table 2 we show possible 'upstream influences' for skill/skill correlations. Specifically, where day 3 skill is evaluated in regions 1 and 8, we show temporally unfiltered correlations with day 1 skill, evaluated in the same region and in the three immediate upstream regions. The day 3 skill lags day 1 skill by up to 3 days. The 3 day lag gives the prognostic correlations, 2 day lag gives the diagnostic correlations, discussed above.

For day 3 skill evaluated in region 1 the highest correlations, for all lags, occurs when day 1 skill is also evaluated in region 1. However, one can see

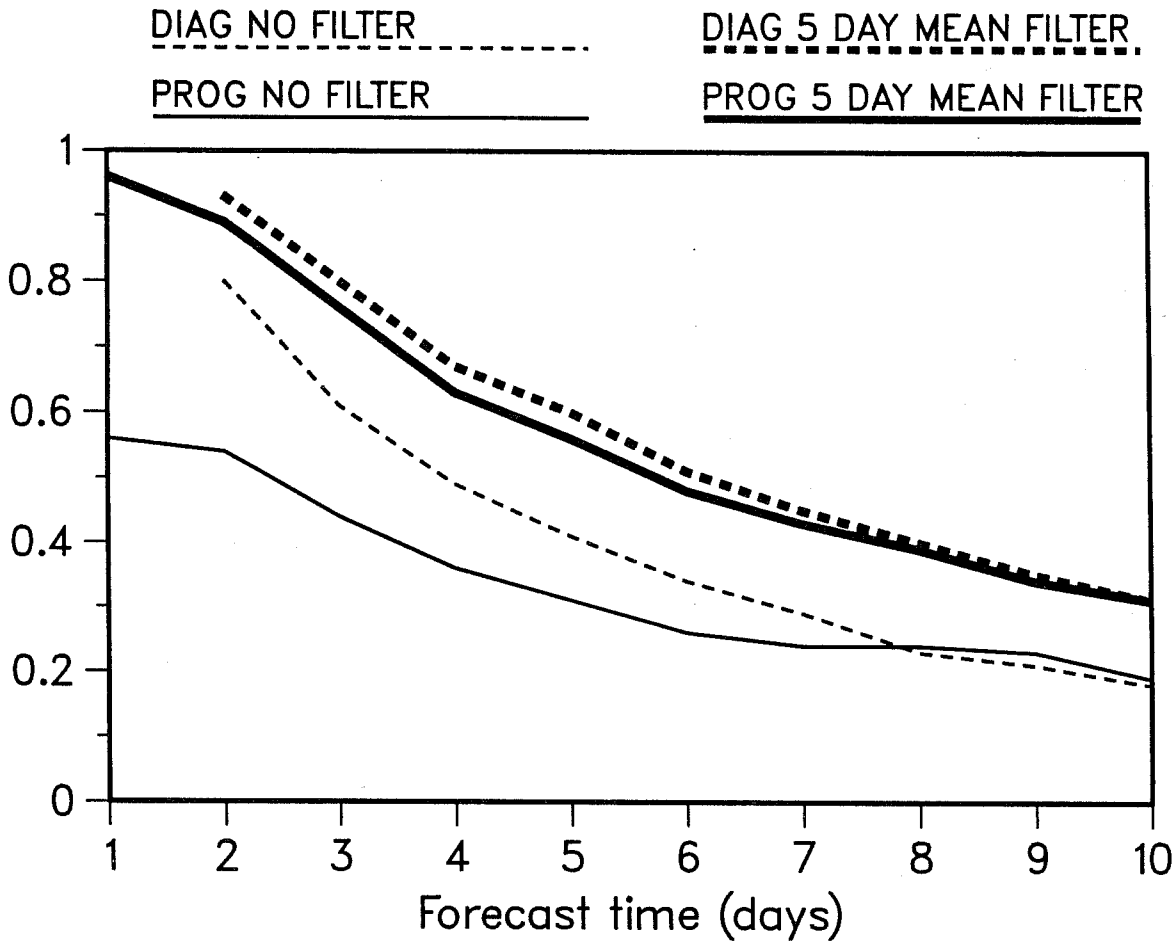

Fig. 13 Correlation between RMS skill and the RMS error of a day 1 forecast. Description of curves as in Fig. 6.

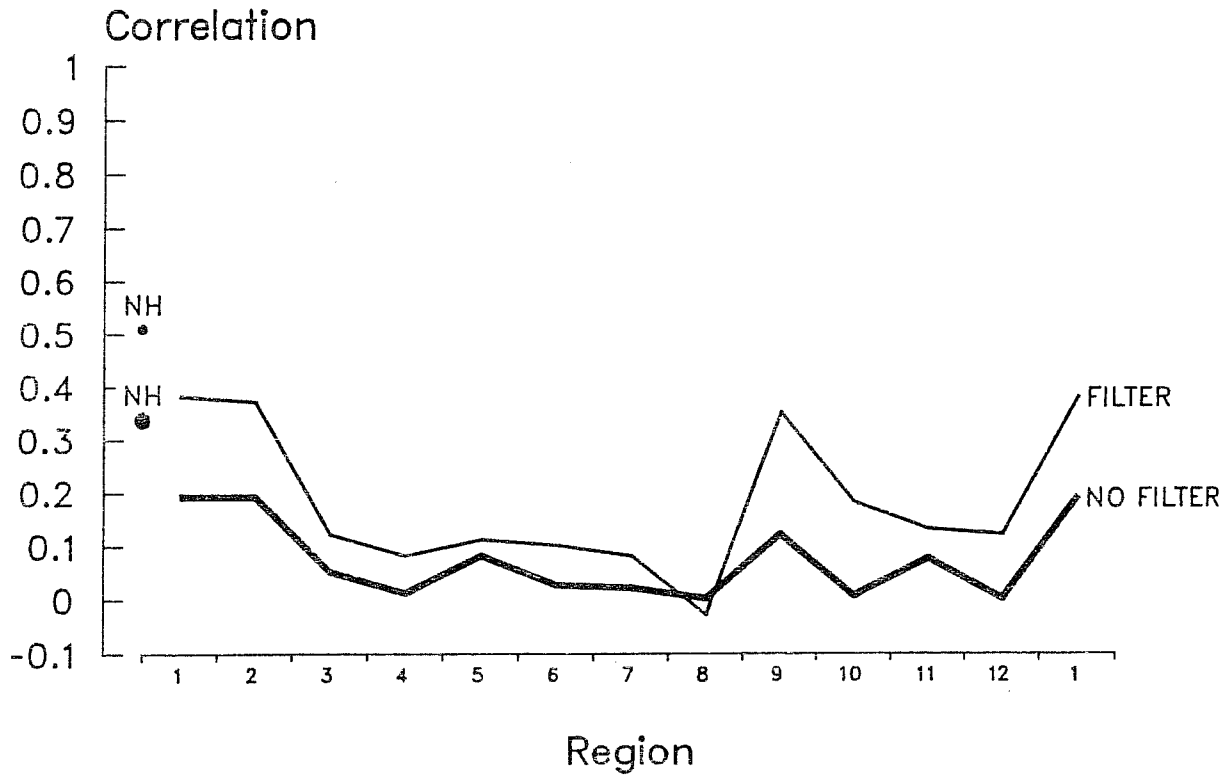

Fig. 14 Correlation between day 1 and day 6 forecast skill (RMS error) with and without 5-day mean filter as a function of longitude (limited area. The black dots on the left represent hemispheric values.

Table 2

a)

Region 10	11	12	1	Lag (days)
.03	.00	.07	.37	0
.00	.07	.14	.45	1
.10	.23	.11	.42	2
.16	.09	.05	.31	3

b)

Region 5	6	7	8	Lag (days)
.06	.00	.10	.23	0
.07	.03	.14	.21	1
.10	.11	.15	.17	2
.10	.04	.02	.08	3

Table 2 Correlation coefficients between day 3 skill in a) region 1 with day 1 forecast skill in regions 10-1 inclusive. b) region 8 with day 1 forecast skill in regions 5-8 inclusive. Both simultaneous and lagged correlations (day 1 skill leading day 3 skill) are shown.

secondary maxima in upstream regions for positive lags. For a 1 day lag, the secondary maxima is in region 12; for a 2 day lag it is further upstream in region 11, and for a 3 day lag it occurs in region 10. This is certainly suggestive of the effects of both propagating and stationary short range forecast errors effecting the day 3 European forecast errors. However, it would appear that the growth of quasi-stationary errors is stronger.

A similar, though less clear-cut pattern can be seen for region 8. For a three day lag the 'stationary' and 'propagating' signal have approximately equal (albeit weak) amplitude.

It should be mentioned that with both these skill/skill correlations and the spread/skill correlations, we have considered the more generalised correlation matrix between day n skill and day m skill/spread, for various time lags. In no cases were results deemed to be significantly better than those described here, and in most cases the correlations were smaller. Some description of these correlation matrices are given in Palmer and Tibaldi (1986,87).

3.4 Forecast RMS transience

Motivation for the use of this predictor, defined as the RMS difference between 500mb forecast height and the 500mb height of the initial conditions, evaluated either regionally or hemispherically, came from purely diagnostic studies of interannual variability of the RMS error of persistence forecasts.

Fig 15 shows a scatter diagram of the winter mean 500mb height error of the model forecast error (ordinate) and persistence error (abscissa) for the seven winter periods 1980/81 to 1986/87, for region 1. For this region in particular, the model performance for 1986/87 certainly fell short of expectations, especially since a number of model and analysis changes had been made to the forecast system (see Palmer and Tibaldi, 1987). Fig 15 shows there is a marked correlation between RMS forecast error and RMS persistence error, with 1986/87 being one of the least persistent winters (and 1985/86 being the most persistent) of the seven year sample. Note also that the difference in model error between 1986/87 and 1980/81, years with similar persistence error but with very different models, is significantly smaller than the difference in model error between 1986/87 and 1985/86, years with similar models but very different persistence error. This is certainly

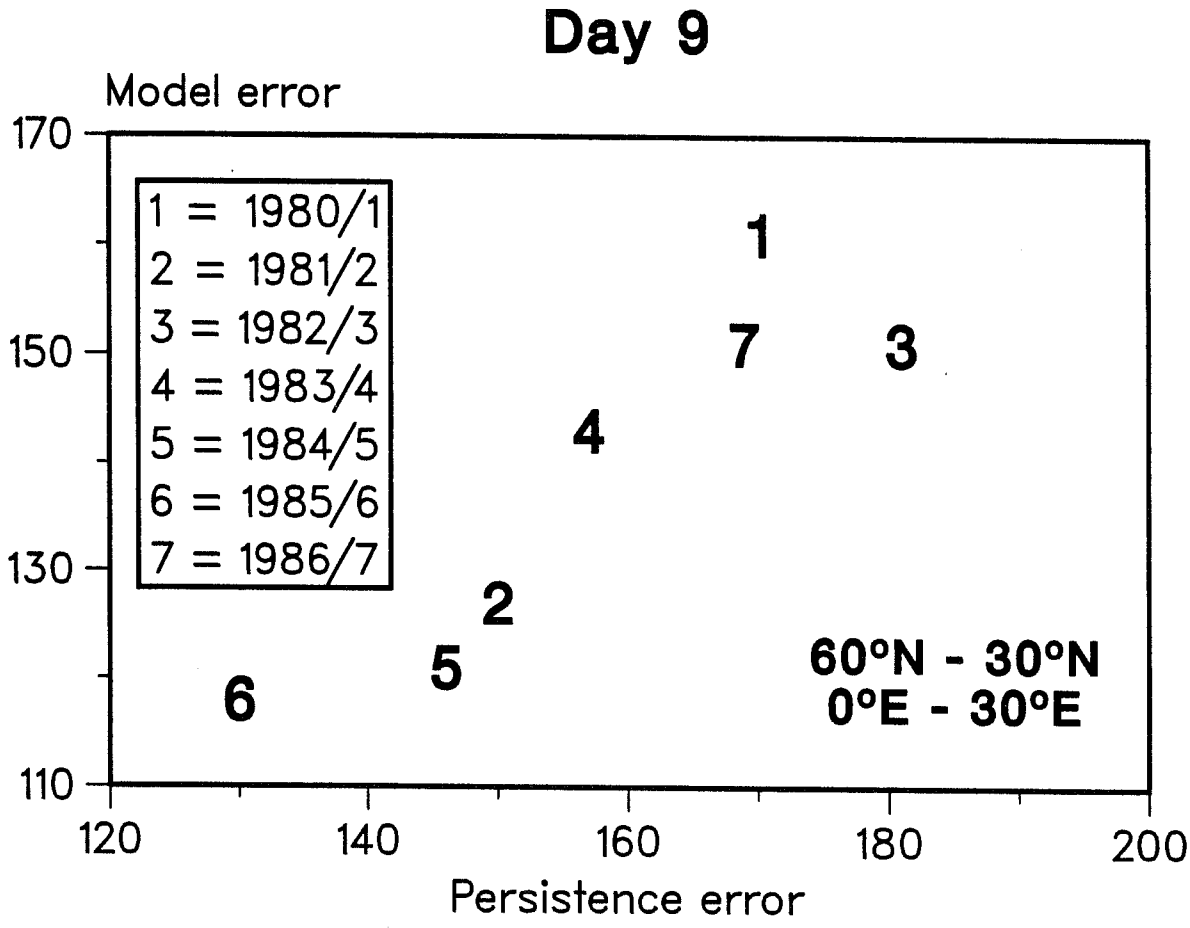

Fig. 15 Scatter plot of winter mean day 9 500 mb height RMS error against the day 9 500 mb height RMS error of a persistence forecast (in region 1) 1=1980/1, 2=1981/2 ... 7=1986/7.

indicative of the relative importance of interannual variability of the atmospheric general circulation in influencing regional scores in the medium range. (It also suggests that none of the modifications to the model or analysis scheme between the winters 1985/86 and 1986/87 were responsible for the relatively poor forecasts over Europe during 1986/87, a fact further confirmed by numerical experimentation).

In Palmer (1988) it is shown that a similar correlation holds for region 7 over the north east Pacific. However, as shown in Palmer and Tibaldi (1987), the correlation breaks down in other regions, and, in particular, breaks down on the hemispheric scale. However, in view of the success of this diagnostic relationship, at least over Europe, it is worth considering whether the degree of persistence of a given forecast, (as opposed to the degree of persistence of the atmosphere) could be used as a predictor of forecast skill. Because of the above diagnostic results, we have only considered this potential of this predictor on a regional basis. Table 3 summarises the results, showing the correlation between the day n RMS error, and the RMS difference between the day n forecast 500mb height, and 500mb height from the initial analysis, using daily unfiltered data. Columns 1 and 2 show results for n=3; columns 3 and 4 for n=6; and columns 5 and 6 for n=9. Only regions 12, 1, 8, and 9 are shown (no significantly higher correlations were found in the other regions). For unfiltered values the magnitude of the correlations generally increase with forecast time. This would seem to make sense, after all we would expect the model to forecast a change of weather regime if it occurred within a day or so of the initial conditions. In most cases, time filtering the data increases the magnitude of the correlations, though not in a particularly coherent way.

In columns 7 and 8 of Table 3 we show a minor variant of this tendency predictor. Instead of calculating the RMS difference between the day 9 forecast and initial state, we calculate the RMS difference between the day 9 forecast and the day 3 forecast. The rationale for this is that whilst the model may be relatively good at predicting transient behaviour within the first three days of the forecast, any predicted change after three days may be somewhat unreliable. Results of correlations using this predictor do show increased skill for region 1, however, the magnitude of the correlation is not consistently larger compared with columns 5 and 6 for regions 8, 9, and 12.

Table 3

Region	day 0/day 3		day 0/day 6		day 0/day 9		day 3/day 9	
	filter off	filter on	filter off	filter on	filter off	filter on	filter off	filter on
12	.17	.27	.18	.19	.25	.25	.23	.28
1	.24	.50	.24	.23	.29	.36	.32	.46
8	.13	.26	.16	.41	.21	.25	.18	.22
9	.14	.28	.11	.08	.27	.37	.22	.39

Table 3 Correlation coefficient between RMS error and the RMS transience predictor for regions 12, 1, 8 and 9.

In conclusion, we would appear to have somewhat mixed results with this predictor. The correlations are positive, though there is no clear pattern to the results. Nevertheless we felt that, overall, it was worth testing this predictor, at least for the day 9 forecasts, on the independent 1986/7 data.

4. PREDICTION OF FORECAST SKILL FOR THE WINTER OF 1986/87

In this section we present some results of attempts to 'predict' variability in the the skill of forecasts of the winter 1986/87 using the variety of predictors discussed above.

In Table 4 we show results for a variety of EOF /skill regressions for day 9 and day 3 scores in regions 12, 1, 8 and 9. Except column 4, the regressed and observed skill scores have been smoothed with the running five day mean filter before the correlation coefficient was calculated. It should be noted that the predictand for these regressions was ln(RMS). This was done to avoid regressed RMS error becoming negative (as it occasionally did when RMS itself was used as predictand) and has the added advantage of making the measure unbounded. The factor structure constants are essentially unchanged by this transformation of the predictand.

The first two columns show that, overall, for day 9, there is no advantage to be had by using EOF coefficients from both initial and forecast data; indeed for regions 12 and 8, the correlation coefficients are lower with M=46 than with M=23. For the other two regions, there is probably no significant difference. Broadly similar results hold for day 3 regressions. For example, comparing columns 5 and 6, the two largest corelations are reduced using M=46. On the other hand the two smallest correlations are increased with M=46, though it could be argued that they are sufficiently small in magnitude that this increase is not significant. Overall this appears to confirm our earlier speculation that the higher correlations using forecast and initial data in the dependent sample was largely illusory. Furthermore, it can be seen that at day 9 the forecasts with only nine EOFs are inferior to those with 23 EOF regressions in regions 12 and 8 (cf columns 2 and 3); again in regions 1 and 9 there is probably no significant difference. The difference between the 9 and 23 EOF regressions are also marked for day 3 forecasts (columns 6 and 7). For region 8 in particular, the correlation drops from 0.47 with 23 forecast EOFs to -0.22 with 9 forecast EOFs. As noted in section 3, 9 EOFs may give a poor

Table 4

Region	1 day 9 M=46 N=23 (5-day filter)	2 day 9 M=23 N=23 (forecast EOFs 5-day filter)	3 day 9 M=09 N=09 (forecast EOFs 5-day filter)	4 day 9 M=23 N=23 (forecast EOFs no filter)	5 day 3 M=46 N=23 (5-day filter)	6 day 3 M=23 N=23 (forecast EOFs 5-day filter)	7 day 3 M=09 N=09 (forecast EOFs 5-day filter)
12	-0.02	0.22	0.10	0.12	0.10	0.03	0.09
1	-0.21	-0.07	0.05	-0.16	0.16	0.05	-0.18
8	0.31	0.34	0.20	0.15	0.42	0.47	-0.22
9	0.31	0.30	0.32	0.30	0.25	0.27	0.26

Table 4 Correlation coefficients between regressed and actual skill scores for regions 12, 1, 8 and 9 for (independent) forecast data from 1986/87.

prediction of the skill scores because the structure factor patterns were generally less geographically localised than with 23 EOFs.

In the 4th column of Table 4, results for day 9 are shown with no time filtering of regressed and actual scores. Whilst there is an overall reduction of skill to quite small values, the scores for region 9 are the same as when the filter was applied. Presumably this result is not statistically significant, since a drop in correlation was found when removing the temporal filter on the training data in region 9 (see columns 3 and 5 of Table 1).

In Fig 16 we compare results from three different predictors for day 3 in regions 1 and 8; in Fig 17 we show similar results for day 9. In each graph, the dashed line shows the actual forecast skill, the solid line shows the predicted skill. The correlation between the two curves is shown (unbracketed) at the top right hand corner of each graph. In brackets the corresponding correlations using the training data are shown. In all cases, both curves have been smoothed by the five-day running mean filter. For day 3, the three predictors are the forecast flow patterns with 23 EOFs, the day 3/day 4 forecast spread, and the day 1 RMS error of 'yesterday's' forecast. In fact the highest level of skill is achieved by the spread indicator in region 1 and the EOF indicator in region 8. The day 1 forecast error is a good indicator of skill in region 1 , less so in region 8, consistent with the discussion above. Note in Fig 16 that there is no obviously strong correlation between the predictor sets, indicating some degree of independence. On the other hand, there are periods when all three indicators agree to some extent (e.g. for region 1, the trough around day 50, or the peak around day 90). Clearly, one would expect to be more confident about a prediction if all predictors broadly agreed.

The predictors used to forecast day 9 skill, shown in Fig 17,are the forecast flow patterns with 23 EOF coefficients, the day 9/ day 10 spread, and the RMS transience between day 9 and the initial conditions. It can be seen that the spread indicator is again the most skilful in region 1, whereas RMS transience is most skilful, and EOF and spread indicators equally skilful, in region 8.

With the exception of the EOF predictors over Europe, and the transience predictor over the Pacific, the correlations between regressed and actual

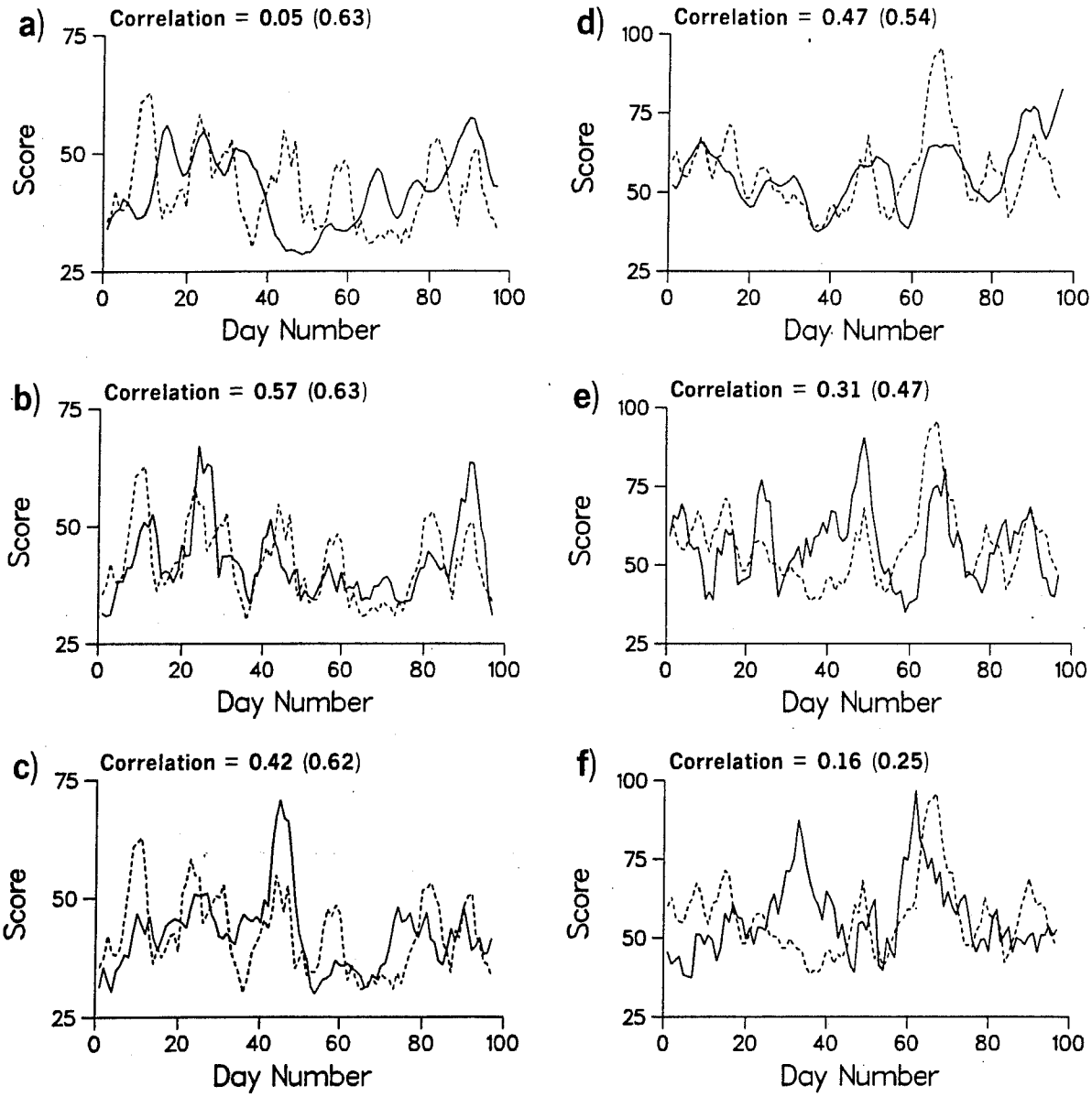

Fig. 16 Prediction of day 3 forecast RMS error for the winter 1986/7 (day 1 = Dec 1) using
a) skill in region 1, and EOF regression
b) skill in region 1, and day 3/day 4 spread
c) skill in region 1, and day 1 forecast error
d) skill in region 8, and EOF regression
e) skill in region 8, and day 3/day 4 spread
f) skill in region 8, and day 1 forecast error.
Solid line is regressed skill, dashed line actual skill. 5-day running filter applied to both curves. Unbracketed number is correlation between two curves. Bracketed number is the correlation expected from 1980/1-1985/6 data.

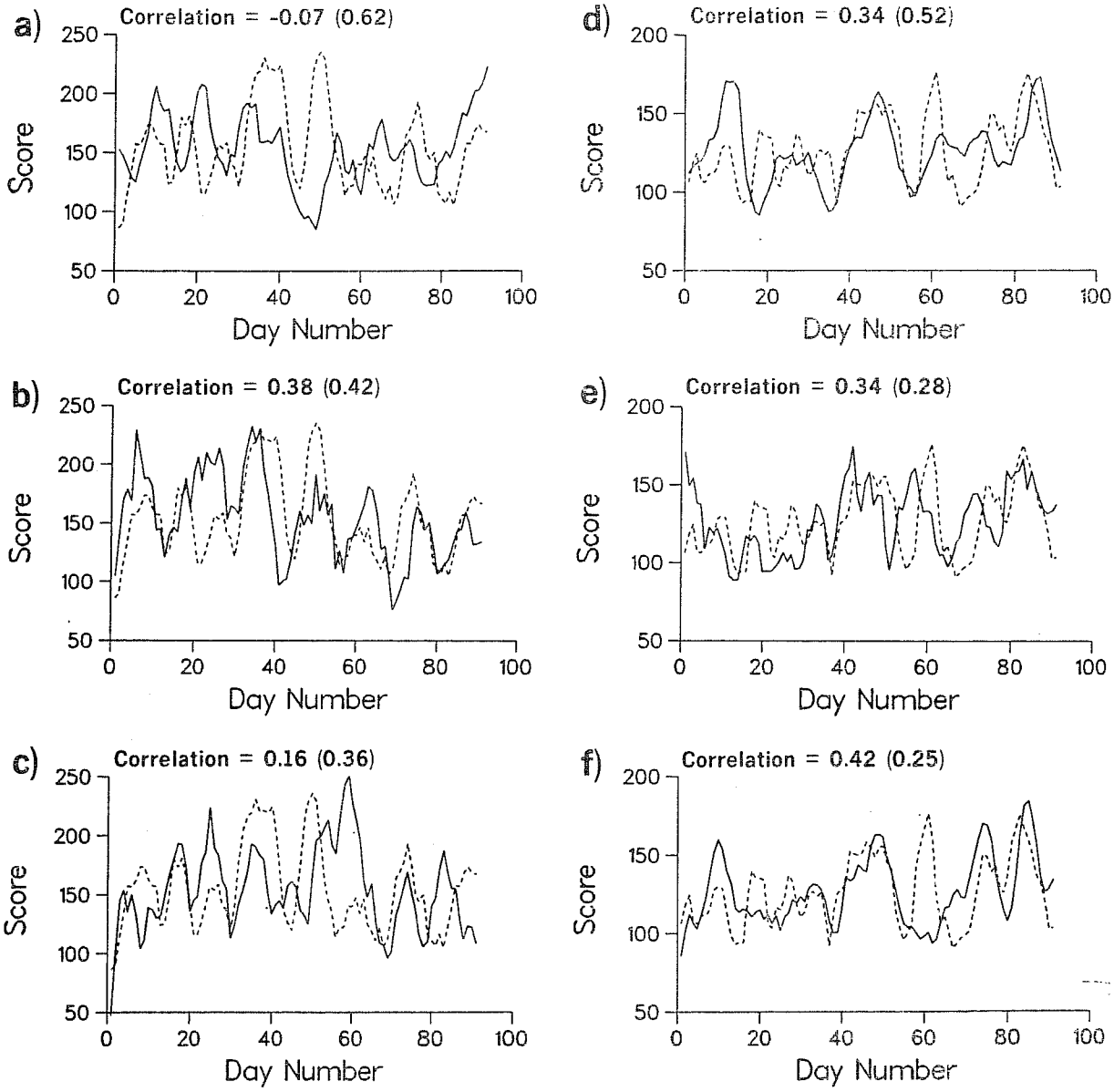

Fig. 17 As Fig. 16 but for day 9 (and day 9/day 10 spread) and c) and f) showing RMS transience predictor.

scores for 1986/7 are about the same or a little less than those derived from the training period. With these exceptions, this could be described as reasonably satisfactory. The relatively poor performance of the EOF predictor over Europe may indicate that the training data is insufficiently homogeneous to give reliable results. As will be discussed in the next section, there are indications that model systematic error strongly influences the factor structure constants for the European regions, and clearly the model systematic error has certainly changed over the training years as different models come into operational use. Of course the penalty for using only more recent years in the training data is the danger of incomplete statistical sampling.

The performance of the transience predictor (much better than expected in region 8 on day 9) is at least consistent with our observation, using the transience predictor on training data, that the correlation between predicted and actual scores for different verification times and regions does not exhibit clearly identifiable patterns. It would appear that there is some useful signal in this diagnostic; however, further work is required to find a more reliable indicator based on forecast persistence.

It would appear that all techniques show some promise as potential predictors, and the most obvious conclusion is that some combination of them may prove optimal. This will be discussed in a further paper, currently in progress.

5. DISCUSSION

We have investigated the potential of four different types of predictor to forecast both regional and hemispheric forecast skill of the ECMWF forecast system. The four predictors, defined objectively, gave measures of the consistency between adjacent forecasts, patterns of forecast flow anomalies associated with either skilful or unskilful forecasts, the short range skill of earlier forecasts, and the degree of persistence of an integration from initialisation time to verification time. It was found that when high frequency day to day fluctuations in forecast skill were filtered out, some encouraging results were obtained.

Possible reasons for the success of these predictors were put forward. Some of these were, from a dynamical point of view, somewhat trivial, relating, for example to the extent to which the forecast skill is dominated by model

systematic error. However, there were also strong indications that variability in the intrinsic stability of the atmosphere played an important role, particularly over the Pacific/North American region. On the hemispheric scale, mechanisms relating to forecast skill variability are both complex and interactive.

However, it was clear that the predictors also failed to capture a significant fraction of the variability of forecast skill. It is easy to imagine some refinement of the EOF regression technique that may give enhanced skill. For example, the training data could be classified according to some indicator of large scale flow (see for example Sutera, 1986), and the regression run separately on each sample. This would give two sets of factor structure constants appropriate to each class of the large scale flow indicator. This technique could be generalised further.

However, such generalisation would not address the question of variability in the quality of the initial analysis. We showed that the day 1 forecast error of the current forecast could provide information about forecast error later in the integration. In this paper, this was done simply by correlating RMS errors. However, a more satisfactory and complete approach would be to perform a regression of EOF coefficients of the short range forecast error against forecast error later in the forecast. This work is in progress and will be reported at a later date. If, operationally, it was considered viable to wait 6 hours, and use this regression analysis on the 6 hour forecast error, we may have a more satisfactory measure of analysis variability than is given in this paper.

In order to implement a scheme to predict forecast skill operationally, it would seem desirable to combine the results of the predictors in some optimal way. In Palmer and Tibaldi (1987), it is suggested that a probabilistic categorical approach be used. The overall output would be the probability that a given forecast lies in one of five a priori equally likely categories. For further details, see Palmer and Tibaldi (1987).

The approach suggested in this paper is only a first attempt at what is clearly a complex problem. Nevertheless we hope to have taken a step towards the more complete forecasting system that Tennekes et al. (1987) so strongly advocate.

Further diagnosis of these results and discussion of mechanisms or low-frequency skill variability are given in Palmer and Tibaldi (1989).

References

Arpe, K., A. Hollingsworth, M.S. Tracton, A.C. Lorenc, S. Uppala and P. Kållberg, 1985: The response of numerical weather prediction systems to FGGE level IIb data. Part II: Forecast verifications and implications for predictability. Quart.J.R.Met.Soc., 111, 67-101.

Branstator, G., 1986: The variability in skill of 72-hour global-scale NMC forecasts. Mon.Wea.Rev., 114, 2628-2639.

Epstein, E.S., 1969: Stochastic dynamic prediction. Tellus, 21, 729-759.

Fisher, R.A., 1921: On the "probable error" of a coefficient of correlation deduced from a small sample. Metron., 1, 1-32.

Frederiksen, J.S., 1986: Instability theory and nonlinear evolution of blocks and mature anomalies. Adv. in Geophys., 29, 277-304.

Grønaas, S., 1985: A pilot study on the prediction of medium range forecast quality. ECMWF Tech. Memo. No. 119. ECMWF, Shinfield Park, Reading, Berkshire, UK.

Hoffman and E. Kalnay, 1983: Lagged averaged forecasting: An alternative to Monte Carlo forecasting. Tellus, 35A, 100-118.

Hollingsworth, A., U. Cubasch, S. Tibaldi, C. Brankovic, T.N. Palmer and L. Campbell, 1988: Mid-latitude atmospheric prediction on time scales of 10-30 days. Atmospheric and Oceanic variability. Ed. H. Cattle. Royal Meteorological Society, Bracknell. p 117-152.

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

Lau, N.G., G.H. White and R.L. Jenne, 1981: Circulation statistics for the Extratropical Northern Hemisphere Based on NMC analyses. NCAR Tech.Note NCAR/TN-171+STR.

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

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

Miyakoda, K., J. Sirutis and J. Ploshay 1987: One month forecast experiments - without anomaly boundary forcings. Mon.Wea.Rev., 114, 2363-2401.

Molteni, F., 1987: Empirical Orthogonal Function Analysis of the Zonal and eddy components of 500 mb height fields in the northern extratropics. ECMWF Tech.Rep. No.61.

Molteni, F., U. Cubasch and S. Tibaldi, 1987: 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, Berkshire, UK.

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

Palmer, T.N., 1988: Medium and extended range predictability, and stability of the PNA mode. Quart.J.Roy.Met.Soc. To appear.

Palmer, T.N. and S. Tibaldi, 1986: Forecast skill and predictability. ECMWF Tech.Memo No.127. ECMWF, Shinfield Park, Reading, Berkshire, UK.

Palmer, T.N. and S. Tibaldi, 1987: Predictability studies in the medium and extended range. ECMWF Tech. Memo. No. 139. ECMWF, Shinfield Park, Reading, Berkshire, UK.

Palmer, T.N. and S. Tibaldi, 1989: On the prediction of forecast skill. To appear in Mon.Wea.Rev.

Schubert, S.D., 1985: A statistical dynamical study of empirically determined modes of atmospheric variability. J.Atmos.Sci., 42, 3-17.

Simmons, A.J., 1987: Barotropic instability, and anomalies of the extratropical northern winter circulation. Pontificae Academiae Scientiarum Scripta Varia 'Drought, El Nino and Teleconnections'. Published by Pontificia Academia Scientarium. Vatican City.

Simmons, A.J., J.M. Wallace and G. Branstator, 1983: Barotropic wave propagation and instability and atmospheric teleconnection patterns. J.Atmos.Sci., 40, 1363-1392.

Sutera, A., 1986: Probability density distribution of large-scale atmospheric flow. Adv. in Geophys. 29, 227-249.

Tennekes, H., A.P.M. Baede and J.D. Opsteegh, 1987: ECMWF workshop on predictability in the medium and extended range. ECMWF, Shinfield Park, Reading, Berkshire, UK.

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

Wallace, J.M., S. Tibaldi and A.J. Simmons, 1983: Reduction of systematic forecast errors in the ECMWF model through the introduction of an envelope orography. Quart.J.R.Met.Soc., 109, 683-718.