AN EXPERIMENTAL SCHEME FOR THE PREDICTION OF FORECAST SKILL AT ECMWF

Franco Molteni and Tim Palmer
European Centre for Medium Range Weather Forecasts
Shinfield Park, Reading, Berkshire

1. INTRODUCTION

The skill of forecasts produced by any numerical weather prediction (NWP) model varies considerably on many different time scales, as is clearly shown by fluctuations in all the indices (for example, RMS error or anomaly correlation coefficient) used to quantify the similarity between predicted and observed fields. This variability has periodic components related to the seasonal cycle, random day-to-day components due to the variable quality of initial analysis, and low-frequency components that, some recent studies suggest, may be related to transitions between different circulation types or regimes (see Palmer, 1988; Tibaldi and Molteni, 1988).

The variations in forecast skill are substantial as forecast time approaches the limit of deterministic predictability, though with a present state-of-the-art model, they are already significant after 5 days. (For example, for the anomaly correlations of 5-day 500 hPa height forecasts over Europe issued by ECMWF in two recent winter periods, 20% of the forecasts had a correlation greater than .91, another 20% a correlation lower than .65). Such fluctuations are obviously associated with considerable variability in the practical usefulness of the forecast itself. Consequently, an a priori indication of the likely level of skill of a numerical forecast would be of enormous assistance to the forecaster in making the best use of the predicted fields, at least in the most predictable situations.

Dynamical predictability clearly depends on the instability properties of the atmospheric flow, and the idea that the departures between integrations started from similar initial conditions can give an estimate of the limits of atmospheric predictability has been the basis of a number of studies in the last decade (see the review by Shukla, 1985). However, NWP models have only recently reached a level of sophistication that allows them to simulate a realistic proportion of atmospheric variability; necessary for forecast dispersion to be a reliable indicator of forecast error. Indications that the

spread of an ensemble of forecasts started from slightly different analyses could be effective as a predictor for forecast error, not only on average but also on a case-to-case basis, came from the idealized study of Hoffman and Kalnay (1983) (so far, however this idea has found only partial confirmation in the studies of Molteni et al. (1987), Kalnay and Dalcher (1987), Murphy (1988) and Brankovic et al. (1988), where actual numerical forecasts are considered).

Whilst forecast spread can give an indication of error amplification due to dynamical instability, it cannot account for sources of error variability related to the quality of initial analyses, or regime-dependent systematic deficiencies in the model. On this basis, Palmer and Tibaldi (1988) studied a number of possible predictors for the forecast skill of the ECMWF model, including very-short-range forecast errors and indices of flow patterns, and concluded that at least the low-frequency components of the skill variations could be predicted to some extent. The promising results of their diagnostic study encouraged implementation of a quasi-operational experiment during winter 1987/88, in which indices of the skill of the ECMWF forecast are predicted statistically every day, from predictors computed from the forecast fields and the initial analysis.

The methodology and the results of the experiment are described in this paper, which is organized as follows. Section 2 describes the predictands, that is, the way in which we chose to quantify the skill of the forecasts, and the geographical areas for which these indices were computed. Section 3 discusses the nature of the predictors used in the statistical scheme; the mathematical algorithm is presented in Section 4. The statistical method was tuned and tested using values of predictands and predictors computed from forecasts and analyses for the winters 1985/86 and 1986/87, in which the T106 spectral model, currently used at ECMWF, was operational. The performance of the scheme on this training data set is described in Section 5, while Section 6 presents the actual results of the experiment for winter 1987/88. An example of how the nature of atmospheric dynamics can influence the prediction of forecast skill is discussed in Section 7, and finally our conclusions and plans for future developments are presented in Section 8.

2. PREDICTANDS

Root mean square (RMS) error and anomaly correlation coefficient (ACC) for the 500 hPa height field are probably the most common indices used to verify numerical weather predictions. Given the experimental nature of this study, we assumed that a prediction of these two parameters would give a reasonable indication of the overall skill of a forecast.

Kalnay and Dalcher (1987, KD hereafter) and Palmer and Tibaldi (1988, PT hereafter) found that the latter parameter is easier to predict than the former, at least in the short range, using forecast consistency as a predictor. PT showed that this is due to the fact that, in the short range, ACC is positively correlated with the amplitude of the anomaly itself, which in turn is mainly determined by the amplitude and phase of the large-scale planetary waves. These waves are usually very well predicted in the short range, due to their persistence over short time scales. On the other hand, short-range errors are mainly determined by failures in predicting the intensity of baroclinic waves, which account for a much lower proportion of the anomaly spatial variance. Therefore, the day-to-day variability of the ACC is largely determined by the variations in the amplitude of the large-scale anomaly whilst RMS error is influenced more by the variability of the skill in predicting baroclinic activity. It is not surprising that, when the spread of two or more forecasts started from slightly different initial conditions, is measured in terms of ACC, this measure is correlated with the ACC between each of the forecasts and the verifying analysis, as long as the initial pattern of large-scale planetary waves persists. This effect will be clearly illustrated by the results of the present study. However, this does not mean that ACC is not a practically useful measure of forecast skill, as has been confirmed by discussions with a number of operational forecasters. The scheme described below therefore includes ACC as a predictand.

The studies of PT and KD also show that the predictability of forecast skill depends on geographical region; not only does the overall accuracy of the prediction of forecast skill vary from region to region, but so does the best predictor of the skill variations. Consequently, better results were obtained for (at least some) areas of continental size (for example, Europe or North America) than for the Northern Hemisphere as a whole. This appears to be due

to the relative importance of low-freqency variability, baroclinic activity and the model systematic error in determining the variability over different areas.

From a practical point of view, predictions of forecast skill are more useful on a synoptic than on a hemispheric scale; on the other hand, the smaller the verification area, the shorter the forecast time for which that area can be considered a dynamically closed system. There is probably an optimum size for the areas over which predictions of forecast skill can be performed using predictors computed only on that area. In order to get some indication of this, we computed our predictands for 7 areas of very different extension: one Hemispheric area (the Northern Hemisphere), two continental areas (North America and Europe), and four sub-continental areas covering parts of the European continent. The skill scores were computed from grid-point fields over a regular latitude-longitude grid with a mesh of 3.75 in both directions; the coordinates of the latitude and longitude lines that delimit the seven areas are listed in Table 1; the limits of areas 2 to 7 are also shown in Fig. 1.

The need to provide a forecast of forecast skill is a consequence of the intrinsic probabilistic nature of even an apparently deterministic prediction of the state of an unstable, non-linear system such as the atmosphere. For this reason, it was considered that a probabilistic prediction of forecast skill should be provided; that is, an estimate of the probability distribution of any parameter that quantifies the possible distance between the forecast and the actual state of the atmosphere. The simplest way of representing a probability distribution is a histogram giving the probabilities of occurrence of a number of classes into which the range of variability of the predictand has been divided. For a number of reasons, it is convenient to choose the classes so that they are, a priori, equally likely. In practice, they are chosen to occur with the same frequency in the available data. We chose to divide the range of our predictands into five classes. Using the values of RMS error and ACC of forecasts started in the periods 1 Dec. 1985/28 Feb. 1986 and 1 Dec. 1986/28 Feb. 1987, we determined the quintiles of their distribution, that is the four values that divide the range of variability into five classes occurring with the same frequency. We also chose to limit our skill prediction to forecast days 3, 5, 7 and 9. Consequently, our

estimate of the skill of one forecast consists of the probabilities of occurrence of 5 classes of RMS error and 5 classes of ACC, for the 500 hPa height field over 7 areas and at 4 forecast times.

For the sake of coherence, we ordered the classes of both RMS error and ACC in such a way that class 1 corresponds to the most skilful forecasts according to both the skill scores, class 5 to the least skilful. This means that class 1 includes the lowest values of RMS error and the greatest values of ACC, and so on. Table 2 shows the quintiles of ACC at different forecast times for three areas of different size: area 1 (Northern Hemisphere), area 3 (Europe), area 5 (south-western Europe). One can clearly see that the variability of the skill increases with forecast time, and it is greater for smaller verification regions. One should also note that our classes simply reflect an ordering of the forecasts on the basis of their skill scores, and do not necessarily correspond to given levels of practical usefulness for the forecaster. For example, forecasts belonging to class 4 (corresponding to below-average skill) at day 3 can still be very useful for many purposes, whereas all the hemispheric forecasts with ACC belonging to class 2 (above- average skill) at day 9 are below the value of 0.6, which is commonly considered as a threshold for the practical usefulness of a medium-range forecast.

A prediction of forecast skill is clearly of most use when it is issued for the latest available forecast. However, PT showed that more effective predictors can be found for the skill of the forecast started one day before. First of all, the actual day 1 error is available; in addition, the spread between two consecutive forecasts is more correlated with the error of the less recent one. Waiting one day is obviously not viable, practically speaking, for short-range forecasts, but possibly some users might prefer to use a medium-range forecast started one day earlier if a good estimate of reliability could be given. In order to test this possibility, we decided to perform a prediction of forecast skill not only for the latest forecast, but also for the previous forecast (i.e, started the day before), with slightly different predictors. In the following, we shall refer to the statistical scheme for the skill prediction of "today's" forecast as the prognostic scheme, to the scheme for "yesterday's" forecast as the diagnostic/prognostic (D/P) scheme. The predictands for the latter scheme are the same as those for the former.

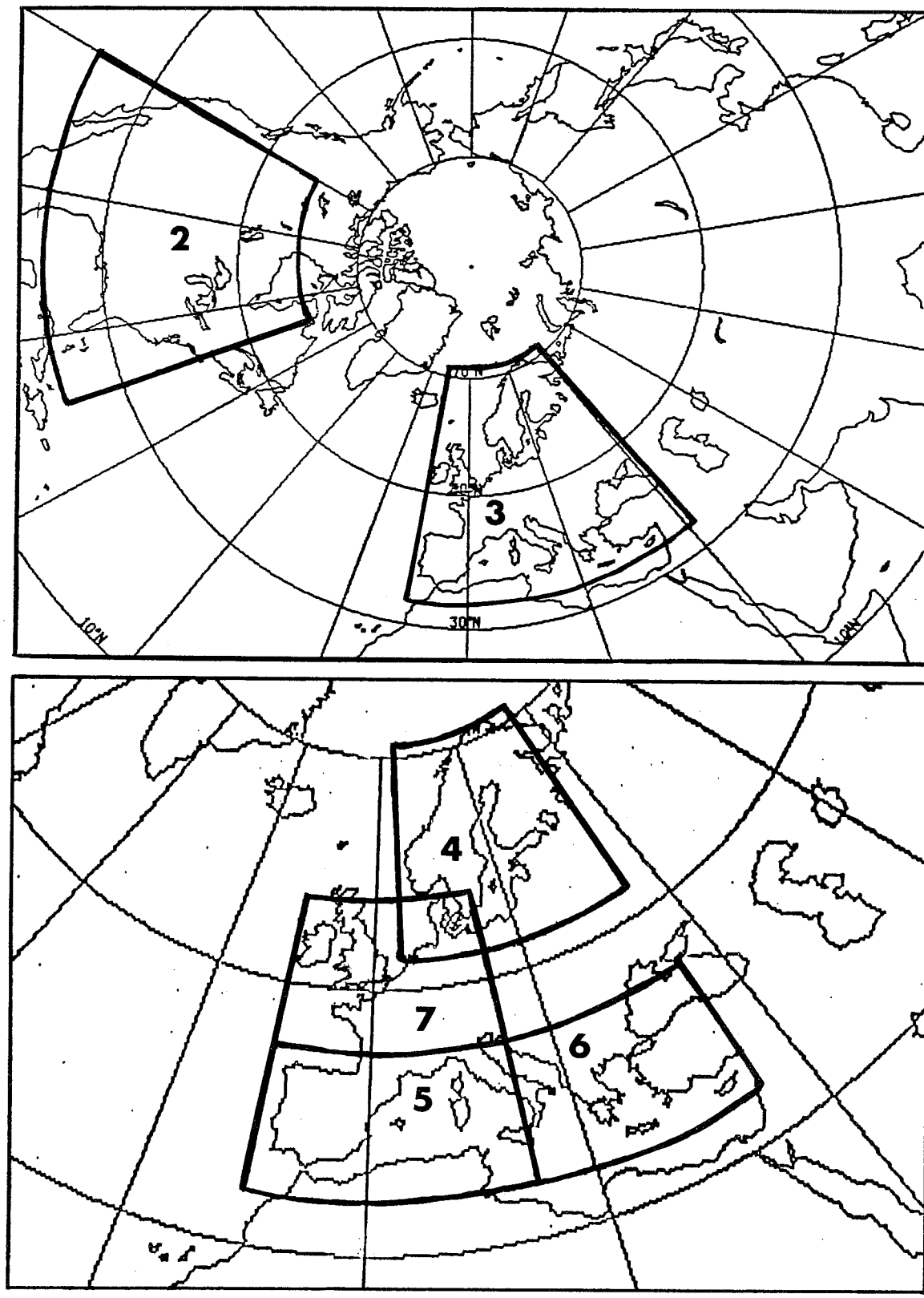

Fig. 1 Location of the verification areas 2 and 3 (top panel) and 4 to 7 (bottom panel).

VERIFICATION AREAS:		LAT(N)	LAT(S)	LON(W)	LON(E)
Area 1:	Northern Hemisphere	78.75	18.75	-	-
Area 2:	North America	60.00	22.50	-120.00	-71.25
Area 3:	Europe	71.25	33.75	-11.25	41.25
Area 4:	Northern Europe	71.25	52.50	3.75	37.50
Area 5:	South-Western Europe	45.00	33.75	-11.25	15.00
Area 6:	South-Eastern Europe	45.00	33.75	15.00	37.50
Area 7:	Central Europe	56.25	45.00	-11.25	15.00

Table 1: Boundaries of the verification areas

FC Day	Area	Quintiles of ACC			
3	1	.93	.92	.90	.89
	3	.97	.95	.93	.91
	5	.99	.97	.94	.88
5	1	.80	.77	.73	.68
	3	.91	.85	.79	.65
	5	.95	.90	.79	.57
7	1	.66	.58	.51	.41
	3	.80	.72	.55	.27
	5	.89	.76	.57	-.02
9	1	.51	.42	.33	.19
	3	.65	.45	.24	-.09
	5	.79	.45	.13	-.36

Table 2: Quintiles of the distribution of ACC in winters 1985/86 and 1986/87, for areas 1 (Northern Hemisphere), 3 (Europe) and 5 (South-Western Europe).

3. PREDICTORS

The predictors used in the statistical scheme are essentially those studied and discussed by PT, and are computed from observed and predicted fields of 500 hPa height. Each predictor is, conceptually at least, distinct from the others; in practice, however, they do overlap in terms of explained variance of forecast skill. They are outlined in rough order of importance below. With the exception of predictor (b), all of them are defined as RMS values of the difference between two fields, and for each of the 7 verification regions the computations are performed over the corresponding areas.

a) Spread between consecutive forecasts. The first predictor is an (inverse) measure of the consistency between forecasts initialized from consecutive 24 hour analyses. It is computed as the RMS difference between forecast fields verifying on the same day. So, for example, in order to forecast either RMS error or ACC over Europe at day 5 in the prognostic scheme, the RMS difference in 500 hPa height over Europe between the current 5-day forecast and the preceding 6-day forecast is computed. In the D/P scheme, conversely, the difference between yesterday's 5-day forecast and today's 4-day forecast is used to predict the skill of the former forecast.

b) Circulation index. The second predictor is an objectively defined measure of the hemispheric-scale forecast flow pattern. From a diagnostic point of view, the dependence of the forecast skill on the circulation regime has been documented in the works of Grønaas (1982), Tibaldi and Molteni (1988), Palmer (1988). Here the difficulty is in translating this knowledge into a prognostic tool, and for this purpose PT showed that the use of forecast fields, rather than initial analyses, gives better results. In order to define this index, forecast fields of 500 hPa height for six winters (from 1981/82 to 1986/87) were projected onto 23 empirical orthogonal functions (see Molteni, 1987). The 23 EOF coefficients were used as independent variables in a linear regression equation for the RMS error over each verification area. In order to take into account the changes in the operational forecast model from a gridpoint N48 model in the first two winters, to a spectral T63 model in the two central winters, and finally to a T106 spectral model in the last two winters, the data have been standardized using

appropriate means and standard deviations computed for the 3 different periods. The circulation index is defined as the output of the regression equation, that is, as a linear combination of the 23 EOF coefficients according to the regression weights. The index can be seen as the time coefficient of a rotated EOF, the rotation being defined in such a way that the time coefficient of one EOF has the maximum possible correlation with the RMS error, while the coefficients of all the other rotated EOF's are orthogonal to error time series. This rotated EOF is proportional to the covariance pattern between the standardized RMS error and the forecast field; such covariance patterns are shown in Fig. 2 for areas 1, 2 and 3 (Northern Hemisphere, North America and Europe, respectively) and for forecast days 3 and 9. One can clearly see that the patterns associated with above-average errors over the two continents are very different, especially at day 9, when they have almost opposite phases. The pattern for North America, as well as the hemispheric pattern, shows a negative PNA signature, in agreement with the findings of Palmer (1988). However, the small amplitude of the covariances, in comparison with the natural variability of the height field, reveals that a lot of error variance remains unexplained by this simple linear approach. We shall discuss this point further in Section 7.

c) Skill of very-short-range forecast. Our third predictor is the RMS error of the day 1 forecast over the region of interest. In the prognostic scheme, the error of the forecast preceding the current one is used; in the D/P scheme, the error of the current forecast is used. Due to the downstream propagation of initial errors, this predictor is a priori unlikely to give good results for small areas.

d) Forecast tendency. The fourth predictor is a measure of the variability occurring in the forecast from the initial to the verification time. It is defined as the RMS difference between the forecast height field at the verification time and the height field in the initial conditions. The results of PT suggested that this predictor could be effective in the medium range; lower variability tends to be associated with higher skill.

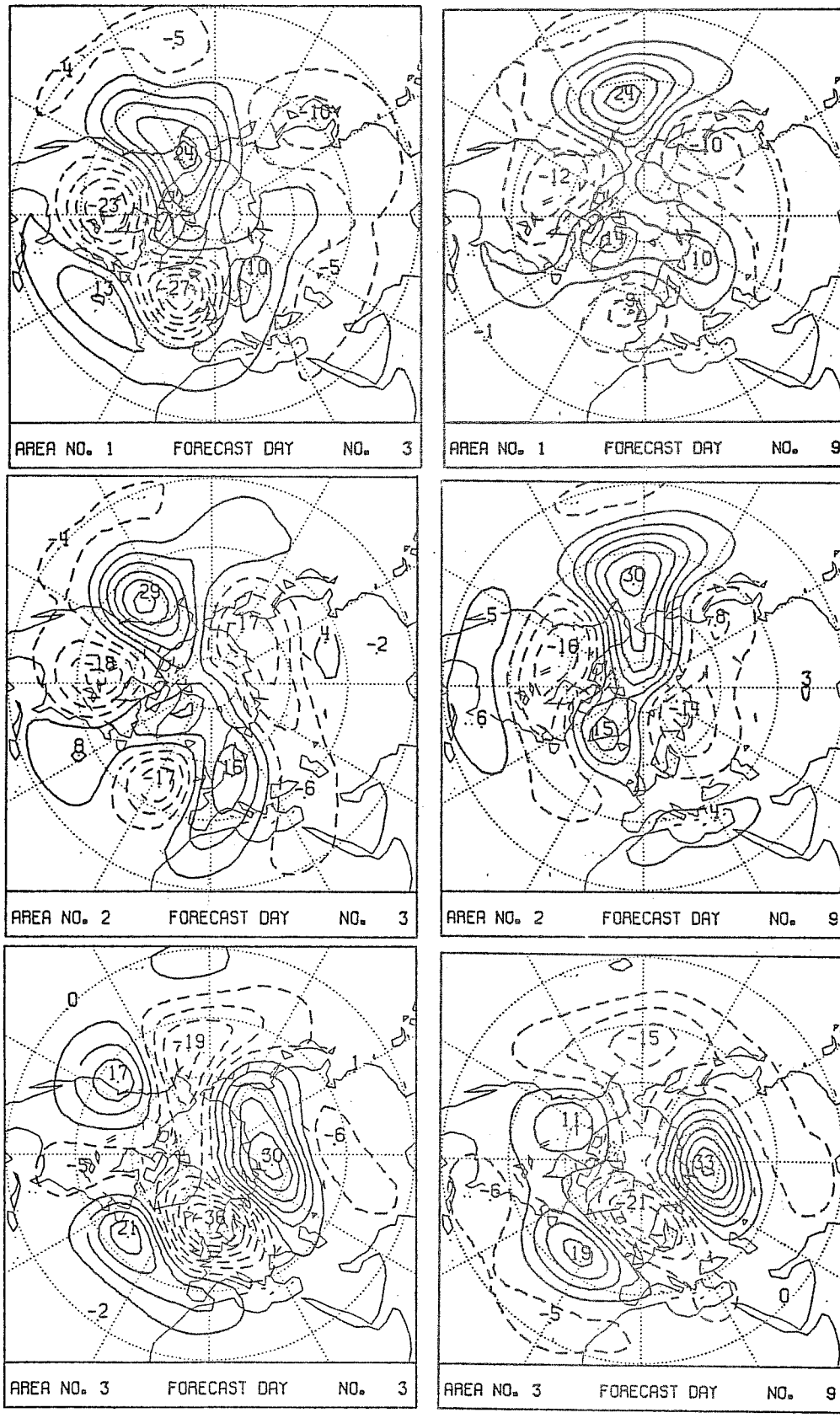

Fig. 2 Covariance patterns between the predicted 500 hPa height and the standardized RMS error over areas 1, 2 and 3, at forecast days 3 and 9.

e) Amplitude of the forecast anomaly. Finally, the RMS difference between the forecast and the climatological field of 500 hPa height is used as an additional predictor for the ACC at forecast day 3 and 5. We decided to use this approach, rather than computing the predictors as RMS values for the forecast of RMS error and as anomaly correlations for the prediction of the ACC, in order to make the effect of the variability in the anomaly amplitude more evident.

Time series of the five predictors were computed for all the forecasts in the two winters 1985/86 and 1986/87, which constitute our training data set. Data for only these two winters were also used for the circulation index, even though the regression weights were computed from a 6-year sample in order to improve their statistical significance; in fact, the circulation index kept its predictive skill when compared with actual errors in the last two winters. As already done for the predictands, the quintiles of the distributions of the predictors were found, and the continuous values were converted into a class index ranging from 1 to 5. The classes were ordered so that class 1 of each predictor should correspond to class 1 of each predictand, and so on, if a perfect correlation existed between the two variables.

In summary, the dependent data set used to tune our statistical prediction scheme consisted of 9 time series of 180 class indices (one value for each day in the two winters) for each verification area: two series for the predictands and seven for the predictors, since predictors (a) and (c) are different for the prognostic and the D/P scheme, while the other three predictors are identical. We are aware of the limitations of this data set, but continual development of the operational model prevented us from having a statistically uniform sample for a longer period.

4. THE ALGORITHM FOR SKILL PREDICTIONS

Given the time series of one predictand and one predictor (converted into class indices) from the training sample, a simple way of analysing the correlation between the two variables is the computation of a contingency table (in our case, a 5 x 5 matrix), in which each element P(i,j) gives the proportion of cases in which the i-th class of the predictand and the j-th class of the predictor occurred contemporarily. The classes have been defined in such a way that the total frequency of each of them is 1/5. Hence, if each

element of the contingency table is multimplied by 5, we obtain the so called conditional probability $P(i|j)$, that is, the probability of occurrence of the i-th class of the predictand when the j-th class of the predictor has occurred. If an exact dependence existed between the two variables, in the form of a monotonically increasing function (not necessarily linear), then $P(i|j)$ would be 1 for $i = j$, 0 otherwise. If, conversely, no correlation at all existed between predictand and predictor, $P(i|j)$ would be equal to 1/5 for each i and j.

A skill score which quantifies the correlation between the time series can be defined as

$$S = [(\sum_{i=1}^{5} P(i|i)) - 1] / 4 \qquad (1)$$

This score is 1 for a perfect correlation, 0 if the two variables are statistically independent.

The matrix $P(i|j)$ provides an immediate way of performing a probabilistic prediction of forecast skill from an independent data set if only one predictor is taken into account. Once the value of the predictor is computed and converted into a class index j, the vector $[P(i|j), i=1,5]$ deduced from the training set can be assumed as the best estimate for the probabilities of occurrence of the 5 classes of the predictand.

When more than one predictor is available, one must combine the conditional probabilities deduced from various predictors. A number of statistical techniques are available for this purpose, which extend the concept of multiple linear regression to probabilistic or categorical variables. The output of these techniques is obviously optimized in order to get the best fit over the training sample; given the very limited size of our training set, there was a strong risk of overfitting the data, i.e. obtaining results with poor statistical significance. Therefore, we decided to test a very simple method, on the basis that results obtained on the training set would have comparable levels of skill for the independent data. The method consisted of

taking a simple weighted average of the conditional probabilities given by different predictors, with the weights being proportional to the skill score S of each predictor.

An example of our procedure is given in Table 3. The conditional probability matrices for one predictand and two hypothetical predictors are given. The skill score is 0.1 for the first predictor, 0.2 for the second one; unfortunately, such low values are typical for our predictors when the full range of daily variability of RMS error and ACC is considered. Now, let us suppose that from the height field of an independent forecast one obtains 2 as the class index of predictor 1, 1 as the index for predictor 2. Our probabilistic forecast for the predictand will be a weighted average of the second column of the first matrix and the first column of the second matrix, with weights 1/3 and 2/3 respectively. In this example, the output indicates that a forecast skill above or much above the average should be expected.

This procedure has been applied independently for the prediction of RMS error and ACC over different areas and at different forecast times. Despite its simplicity, it gives better skill forecasts than those obtained selecting only the most correlated predictor for any predictand, area and forecast time in the training set. These results will be discussed in the next section.

Operationally, the probabilities have been graphically displayed in the form of histograms. Fig. 3 gives an example of the actual output for ACC over the Northern Hemisphere, produced by the prognostic scheme for the forecast started on 8 December 1987. For this day, a clear indication of a very good forecast was provided for forecast day 3 and 5; a good forecast is also predicted for day 9, but with much less confidence, while the results for day 7 give a very confused indication, which appears in contrast with the results for other forecast times. In a few days, inconsistencies were also noticed between predictions of RMS error and ACC, beyond what could be expected because of the inexact correlation between the two measures of skill. The fact that independent forecasts were issued for the two predictands and the four forecast times is the cause of these shortcomings. However, as we shall see in Sect. 7, another apparent source of uncertainty for the forecaster, namely the occurrence of bimodal probability distributions in the medium range, in many cases reflected a real feature of the large-scale dynamics.

CONDITIONAL PROBABILITY MATRICES

S = (1.4-1)/4 = 0.1		Class of predictor 1				
		1	2	3	4	5
	1	.25	.25	.20	.20	.10
Class of	2	.25	.30	.20	.15	.10
	3	.20	.20	.25	.15	.20
Predictand	4	.20	.15	.15	.25	.25
	5	.10	.10	.20	.25	.35

S = (1.8-1)/4 = 0.2		Class of predictor 2				
		1	2	3	4	5
	1	.40	.25	.20	.10	.05
Class of	2	.25	.30	.20	.15	.10
	3	.20	.20	.30	.15	.15
Predictand	4	.10	.15	.15	.35	.25
	5	.05	.10	.15	.25	.45

If predictor 1 is in class 2 and predictor 2 is in class 1, the estimated probabilities are:

$$(1/3)\begin{bmatrix} .25 \\ .30 \\ .20 \\ .15 \\ .10 \end{bmatrix} + (2/3)\begin{bmatrix} .40 \\ .25 \\ .20 \\ .10 \\ .05 \end{bmatrix} = \begin{bmatrix} .35 \\ .27 \\ .20 \\ .11 \\ .07 \end{bmatrix}$$

Table 3: Example of a hypothetical forecast skill prediction (See text for explanation)

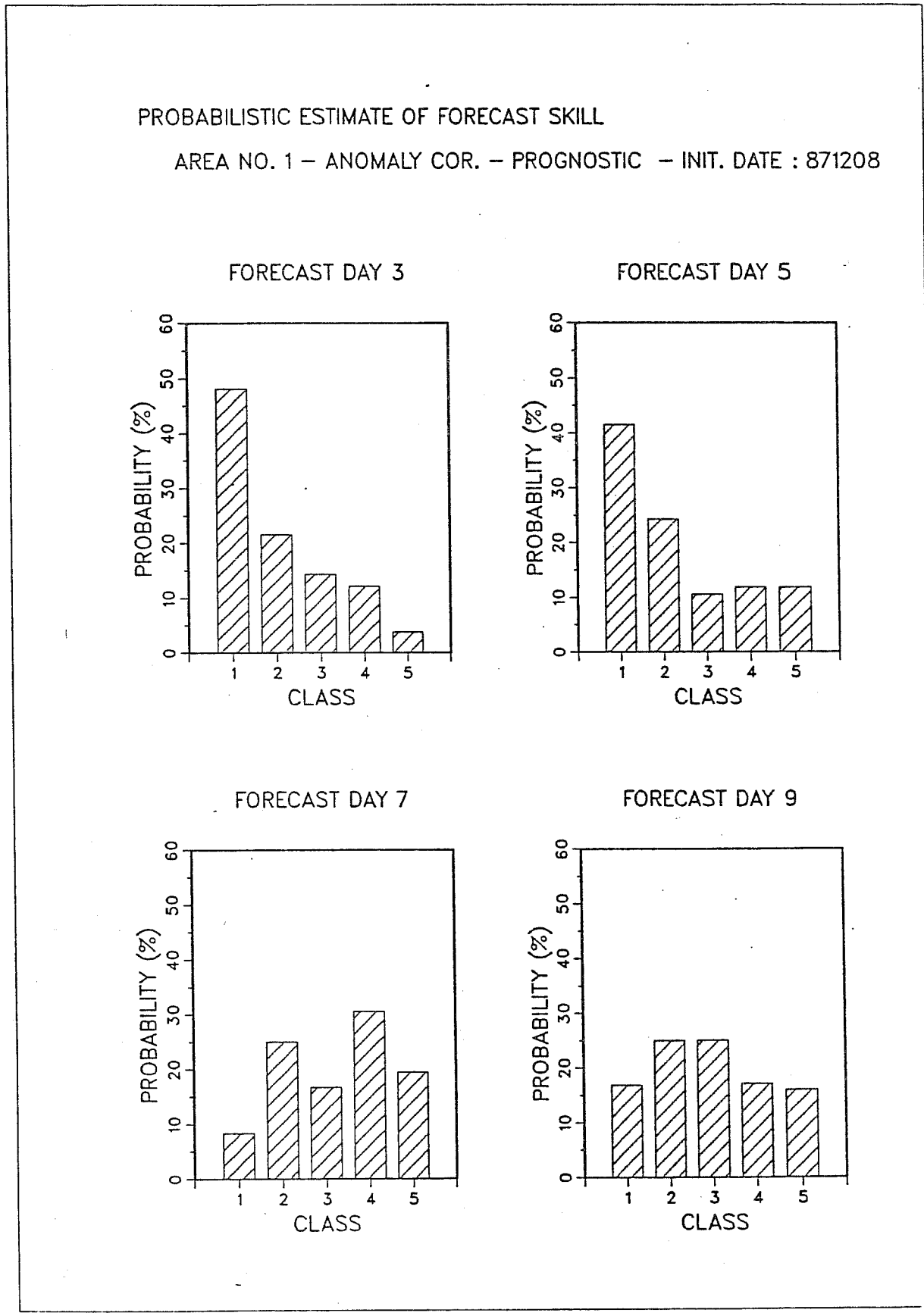

Fig. 3 Example of graphical output from a quasi-operational prediction of forecast skill during winter 1987/88.

5. RESULTS ON THE TRAINING SAMPLE

The results of the application of our statistical prediction scheme to the training data are summarized in Table 4. This table gives the best predictor, the skill score S of this predictor, and the skill score of the full scheme for each verification area and forecast time. A skill score for the full scheme was computed by selecting for each day the skill class with the highest predicted probability; the proportion of correct forecasts (that is, cases with coincidence between observed and predicted class of skill) over the two winters, C, was computed and converted into a normalized skill score defined as (C - 0.2) / 0.8, which again is zero for a random forecast and 1 for a perfect forecast.

In the prediction scheme for RMS error, the circulation index was the best predictor in more than half of the cases (15 out of 28), particularly at day 3 and 9; in general, spread appeared as the second predictor in order of skill, and was the most skilful one in 8 cases. However, the skill score of even the best predictor was usually very low, and only in a few cases exceeded 0.1. As a result the predicted probability distributions were often almost flat, reflecting the strong uncertainty in the prediction.

The situation improved clearly in the diagnostic/prognostic scheme for RMS error, where the spread from the latest forecast was almost always the best predictor. Its skill score was greater than 0.1 in all cases (with only one exception), and between 0.2 and 0.3 in 8 cases. Even though the highest scores occurred in the short range, when the D/P scheme is not useful operationally, an improvement over the prognostic scheme could be still detected at day 7 and 9.

The addition of the anomaly amplitude as a predictor for ACC at day 3 and 5 shows its clear effect in the results of the prognostic scheme. The anomaly amplitude is the best predictor over all areas at day 3 and 5, with a skill score generally between 0.1 and 0.2. At forecast day 7 and 9, when this predictor is not used, the situation is similar to that for the prognostic prediction of RMS error; the circulation index becomes the best predictor in most cases, but with very low skill.

RMS prognostic	RMS diag./prog.	ACC prognostic	ACC diag./prog.

BEST PREDICTOR S : spread C : circ. index E : day 1 error
T : fc. tendency A : anomaly ampl.

fc. day	area 1 2 3 4 5 6 7	area 1 2 3 4 5 6 7	area 1 2 3 4 5 6 7	area 1 2 3 4 5 6 7
3	E C C C C C C	S S S S S S S	A A A A A A A	A A A S A A A
5	T S S S S C C	S S S S S S S	A A A A A A A	S A S S S A S
7	C E S S T C S	S E S S S S S	C E E C C C E	S E S C S S S
9	C C S C C T C	C S S S S S C	C C C C C S E	C S S S S S S

SKILL SCORE of BEST PRED. + : .0 – .1 * : .1 – .2 X : > .2

fc. day	area 1 2 3 4 5 6 7	area 1 2 3 4 5 6 7	area 1 2 3 4 5 6 7	area 1 2 3 4 5 6 7
3	* + + + + + *	X X X * X * *	* * * * X X *	* * * X X X *
5	+ + + * + + *	* * X X * + X	+ * * * * * *	+ * * * * * *
7	+ * + * + + +	X * * * * * *	+ + + * + + +	* + * * + + +
9	* + * + * + *	* * * * * * *	* + + + + + +	* + + * + + *

SKILL SCORE of FULL SCHEME + : .0 – .1 * : .1 – .2 X : .2 – .3 X > .3

fc. day	area 1 2 3 4 5 6 7	area 1 2 3 4 5 6 7	area 1 2 3 4 5 6 7	area 1 2 3 4 5 6 7
3	X X X * X X X	X X X X X X X	X X X X X X X	X X X X X X X
5	X * * * * * X	X X X X X * X	* X X X * X *	X X X X X X X
7	* X X X * X *	X X X X X * X	* X * * X * *	X * X * X X *
9	X X X X X X X	X X X X X X X	X X * X X * *	* X X X X X X

Table 4 Best predictor (top panel), skill score of the best predictor (centre panel) and skill score of the full prediction scheme (bottom panel) for all the predictands, verification areas and forecast times, as deduced from the training data.

In the D/P scheme for ACC, the 'diagnostic' spread generally becomes the best predictor from day 5 onwards, but the improvement in skill score is not so strong as for the RMS error. At day 3 and 5, anomaly amplitude and diagnostic spread have comparable skill scores.

When the skill score of the full scheme (for both RMS error and ACC, and for both the prognostic and the D/P version) is compared with that of the most skilful predictor, a clear improvement can be seen for all areas and forecast times, despite the simplicity and non-optimal nature of our algorithm. In most cases, the skill score of the full scheme was between 0.2 and 0.3; a number of cases with skill score over 0.3 were also found, even though most of them were obtained by the D/P scheme in the short range. Such scores, even if not exceptional, are close to what could be expected on the basis of the diagnostic study of PT, if one takes into account that we have tried to predict the full daily range of variability of the forecast skill, and not only its low-frequency component. An example of how much time-filtering the observed and predicted variables can improve results will be shown in the next section.

6. RESULTS OF THE QUASI-OPERATIONAL PREDICTIONS FOR WINTER 1987/88

The results of the daily predictions of forecast skill during the period 1 December 1987/28 February 1988 are summarized in Table 5. The top panel is the exact counterpart of the bottom panel in Table 4: it shows the skill score of the prognostic and the D/P scheme for RMS error and ACC, defined as $(C - 0.2)/0.8$, where C is the proportion of correct forecasts (again, a forecast is assumed to be correct if the observed class of skill was the one with the highest probability in the prediction). The bottom panel presents the results of verifications performed in terms of correlation between the observed and the 'expected' class index. The expected class index is defined as a continuous parameter in the following way:

$$EI(k) = \sum_{i=1}^{5} P(i,k)\, i \qquad (2)$$

where i is the class index, k the index of the day in winter 1987/88 (k=1,..,90) and P(i,k) the predicted probability for class i on day k. If

RMS prognostic	RMS diag./prog.	ACC prognostic	ACC diag./prog.

SKILL SCORE + : .0 – .1 * : .1 – .2 X : > .2

fc. day	area 1 2 3 4 5 6 7	area 1 2 3 4 5 6 7	area 1 2 3 4 5 6 7	area 1 2 3 4 5 6 7
3	+ + + + + + +	* * * * X * *	* * * * X X *	X * + X X X X
5	* + + + +	+ * + + X * +	X * * * * * *	X + * * + * *
7	+ + + + +	+ + + + * *	+ +	+ + + + +
9	+ +	+ + +	+ + + *	+ + +

CLASS INDEX CORREL. + : .0 – .2 * : .2 – .4 X : .4 – .6 X ~ > .6

fc. day	area 1 2 3 4 5 6 7	area 1 2 3 4 5 6 7	area 1 2 3 4 5 6 7	area 1 2 3 4 5 6 7
3	+ + + + + + +	X X X X X X X	X X X X X X X	X X X X X X X
5	+ * * + + + +	X * * * * * X	X * * X X X X	X * * X X X X
7	+ + + +	+ * * + X * *	+ +	+ + * + * +
9	+ + + + +	+ * * * * *	+ + * +	+ + + + +

Table 5 Skill score (top panel) and class-index correlation (bottom panel) for all the predictand, verification areas and forecast times, as computed from the quasi-operational predictions of forecast skill during winter 1987/88.

OI(k) is the observed class index, a class-index correlation (CIC) can be defined as

$$CIC = \sum_{k=1}^{90} (EI(k)-3)(OI(k)-3) \; / \; [\; (\sum_{k=1}^{90} (EI(k)-3)^2)(\sum_{k=1}^{90} (OI(k)-3)^2)]^{\frac{1}{2}} \quad (3)$$

Since 3 is the class index corresponding to average skill, a positive value of the CIC is given if both the observed and the expected class index indicate an above-average or a below-average skill. This verification method takes into account the fact that if a distinction between 5 classes seems to be beyond the practical skill of our prediction scheme, a cruder estimate could still be useful to operational forecasters.

When comparing the (normalized) proportions of correct forecasts obtained on the training data with those in winter 1987/88, the results appear disappointing: the clear improvement over individual predictors shown in Table 4 has been strongly reduced in most cases. There are very few correct forecasts at day 7 and 9; the skill score of the prognostic scheme for RMS error is generally below .1, and the only predictions with practical usefulness seem to come from the prognostic scheme for ACC at day 3 and 5.

The situation becomes more optimistic when the verification is performed in terms of class-index correlation. Over the two large continental areas (North America and Europe), a CIC between .2 and .4 is obtained with the prognostic scheme for RMS error at day 5; although these values are low in an absolute sense, they are not very far from the theoretical upper limits for areas of similar size estimated by Kruizinga and Kok (1988, this volume). The results of the prognostic scheme are much better for ACC at day 3 and 5, with CIC generally greater than .4. The fact that this improvement over the RMS error predictions disappears at day 7 and 9 clearly shows that it is due to the inclusion of the anomaly amplitude as a predictor at days 3 and 5. For both RMS error and ACC, the results are much better with the D/P scheme than with the prognostic scheme. In general, the improvement is much more evident at day 3 and 5, when its usefulness is limited, but for RMS error, CIC between .2 and .4 can be found over 5 areas even at day 9.

Since no specific information about the initial or very-short-range error of a given forecast is used in our prognostic scheme, this scheme essentially gives a measure of the instability of the circulation regime occurring during the period from the initial to the verification date, together with indications about some systematic deficiencies of the model, typical of that regime. Therefore, a better correlation can be expected between observed and predicted skill scores if a time filter is used to remove the day-to-day variability induced by the quality of the initial analysis, as shown by PT. In our case, verifications for running 3-day means of predicted and observed RMS error and ACC were performed in the following way. For the observed values, 3-day averages of the actual scores were computed, and the mean value was subsequently converted to a class index. For the predictions, a 3-day average of the 'expected' index given by Eq. (3) was considered. The use of 3-day means, although less effective than the 5-day filter used in PT, did not change too much the range of variability of the scores, so that the same classes defined for the daily values could still be used.

Scatter plots comparing the class-index correlations for daily values and 3-day means of RMS error and ACC are shown in Fig. 4; predictions from the prognostic scheme at forecast day 3 and 5 are used in this verification. One can see that the improvement in skill is greater for RMS error than for ACC, and is greater for hemispheric or large continental areas than for the small sub-continental regions. The reason is that the correlation between ACC and the anomaly amplitude exists on all time scales, so that the daily predictions of ACC have some skill even in forecasting the day-to-day fluctuations of this index. Moreover, the smaller the area, the stronger the high-frequency variability of the anomaly amplitude. In fact, when the 3-day filter is applied, the predictions of RMS error over areas 2 and 3 at day 5 become as good as the corresponding predictions of ACC; conversely, on average there is no improvement in the ACC predictions for the sub-continental regions 4 to 7.

Figures 5 to 9 show time series of daily values for observed and predicted class indices at day 5. Figs. 5 and 6 present the results for RMS error over North America and Europe respectively, while Figs. 7, 8 and 9 refer to the predictions of ACC over the Northern Hemisphere, Europe and south-western Europe. Here the predicted class is the one with the highest probability; an open circle marks the days in which this probability is greater than .3, a

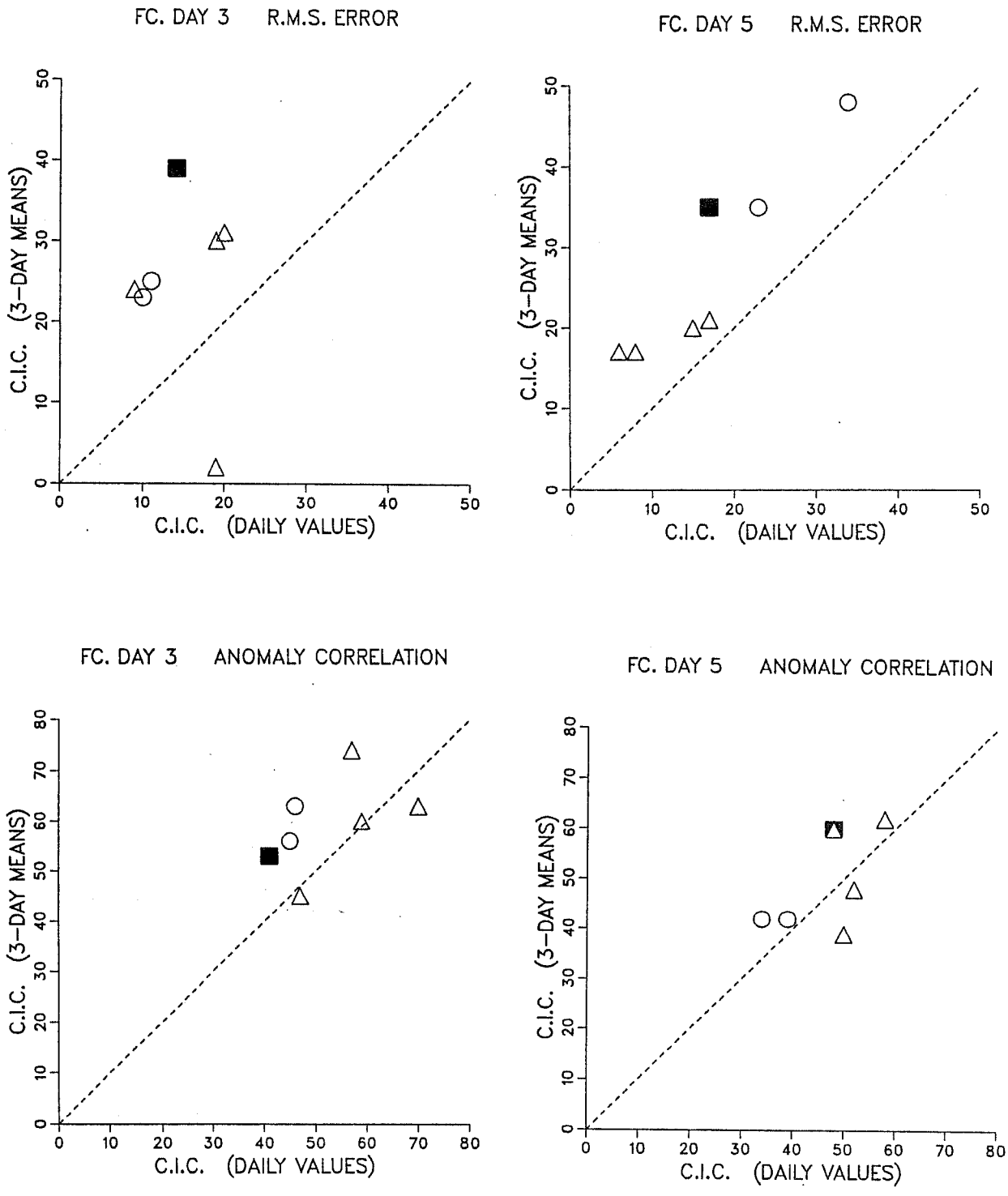

Fig. 4 Comparison between class-index correlations computed from daily values and 3-day means of observed and predicted RMS errors and ACC's. Black square: Northern Hemisphere (area 1). Circles: North America and Europe (areas 2 and 3). Triangles: European subcontinental areas (4 to 7).

VERIFICATION OF SKILL PREDICTIONS

FORECAST DAY 5 – AREA NO. 2 – R.M.S. ERROR

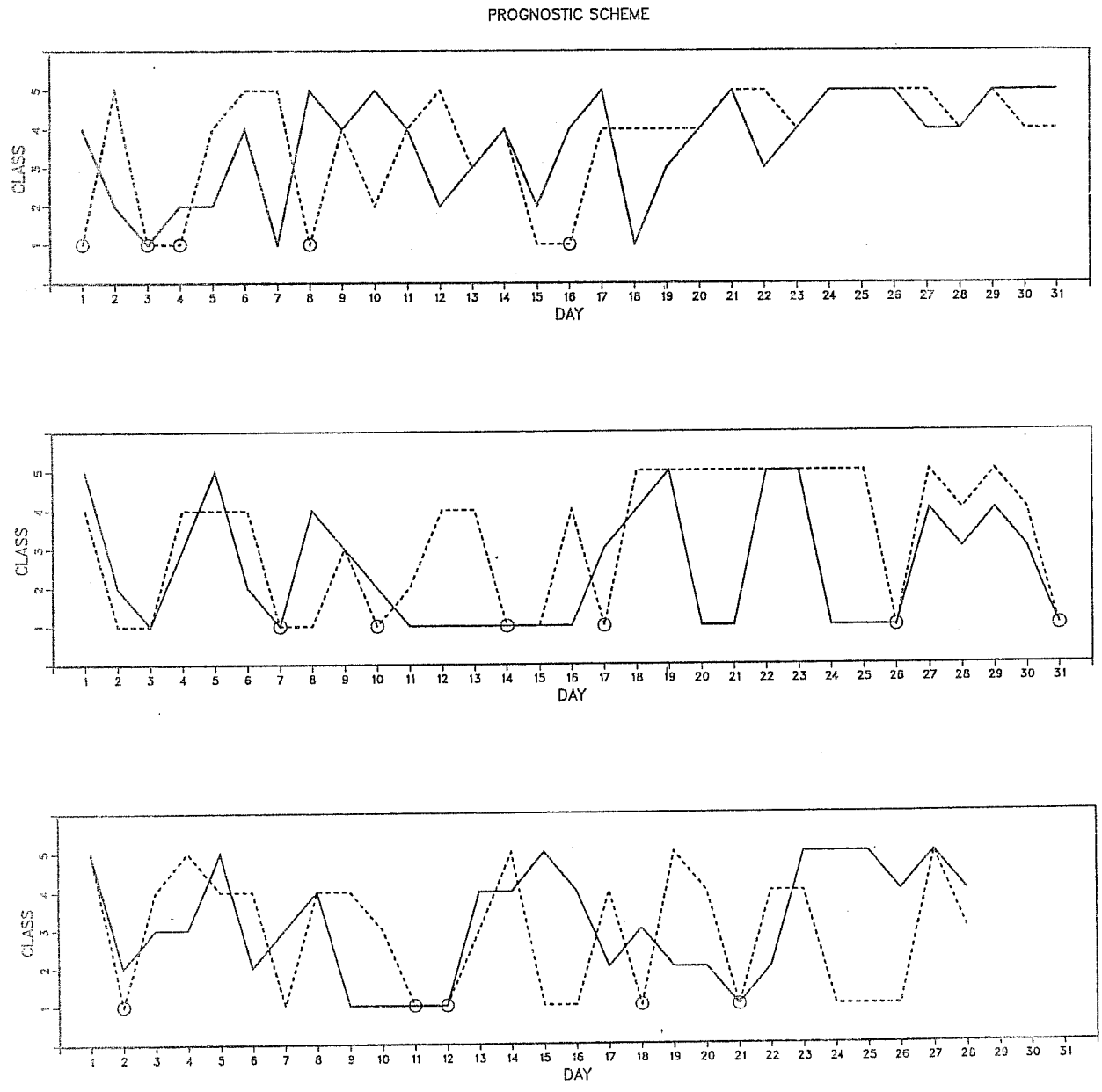

Fig. 5 Time series of observed (solid line) and predicted (dashed line) classes of RMS error at forecast day 5 over North America in December 1987 (top panel), January 1988 (centre panel) and February 1988 (bottom panel). The predicted class is defined as the class with the highest probability: a circle on the dashed line indicates that, in a particular day, this probability was higher than .3 (open circle) or higher than .4 (full circle).

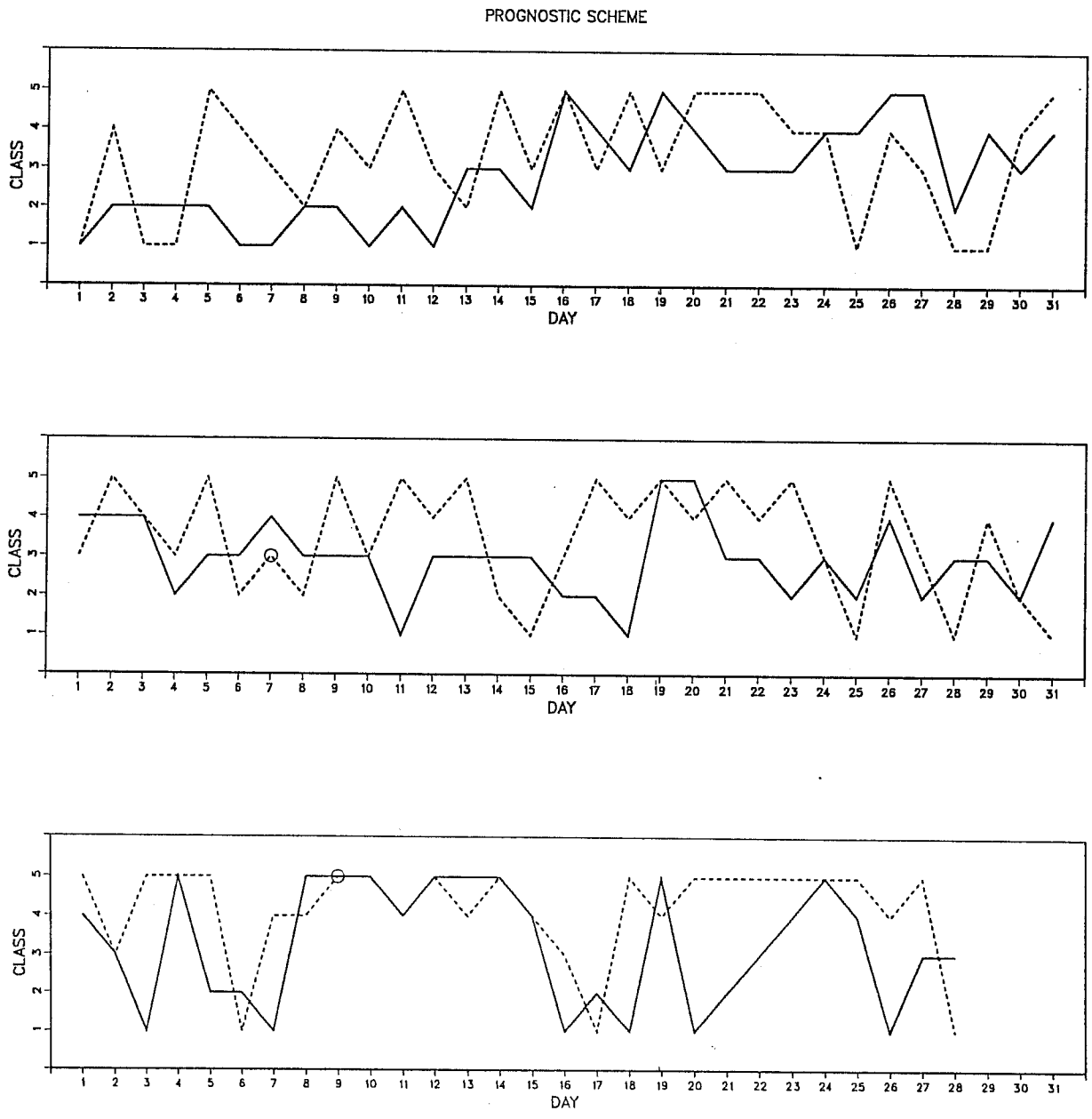

Fig. 6 As in Fig. 5, but for Europe.

VERIFICATION OF SKILL PREDICTIONS

FORECAST DAY 5 – AREA NO. 1 – ANOMALY COR.

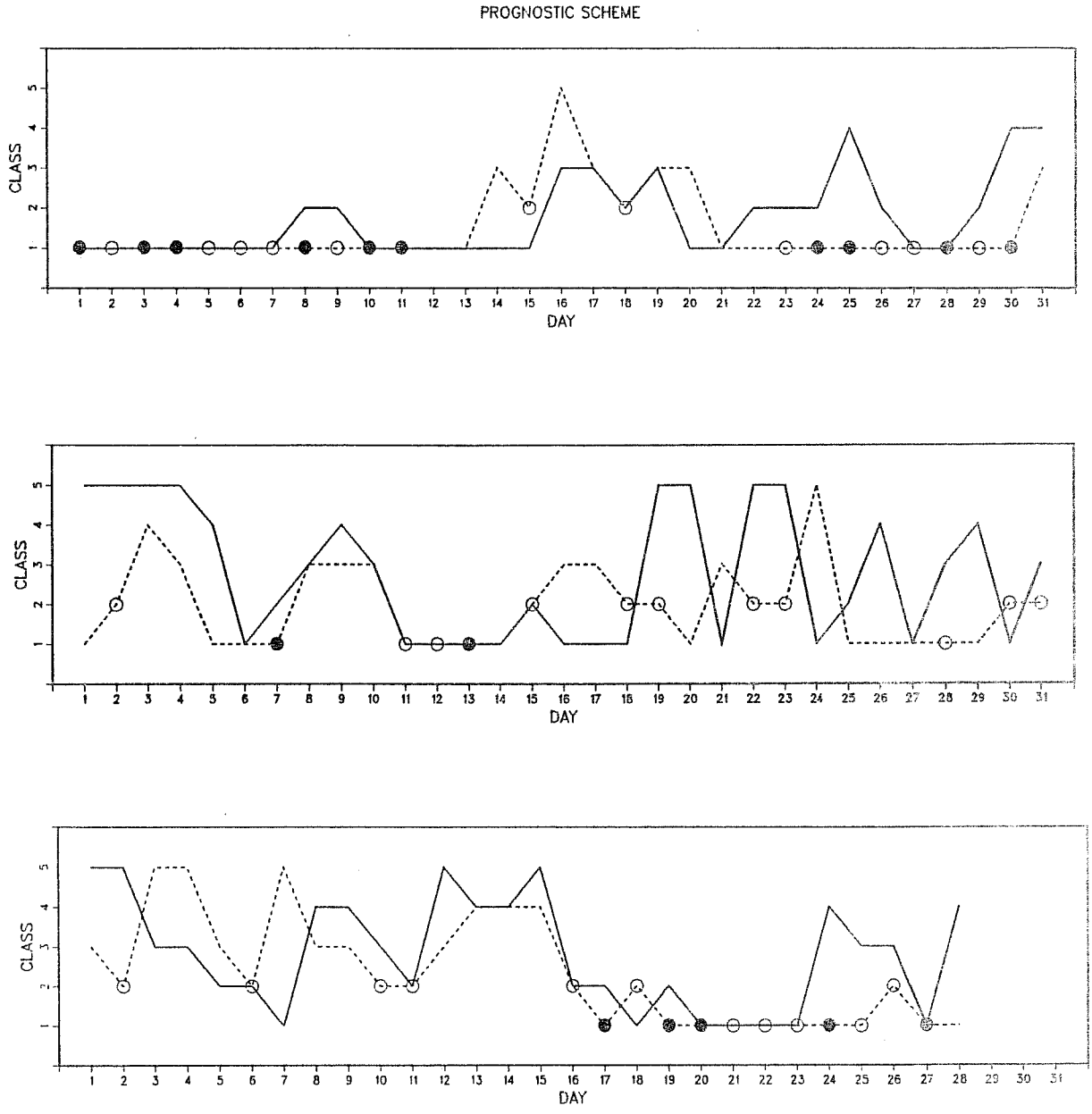

Fig. 7 As in Fig. 5, but for anomaly correlation over the Northern Hemisphere.

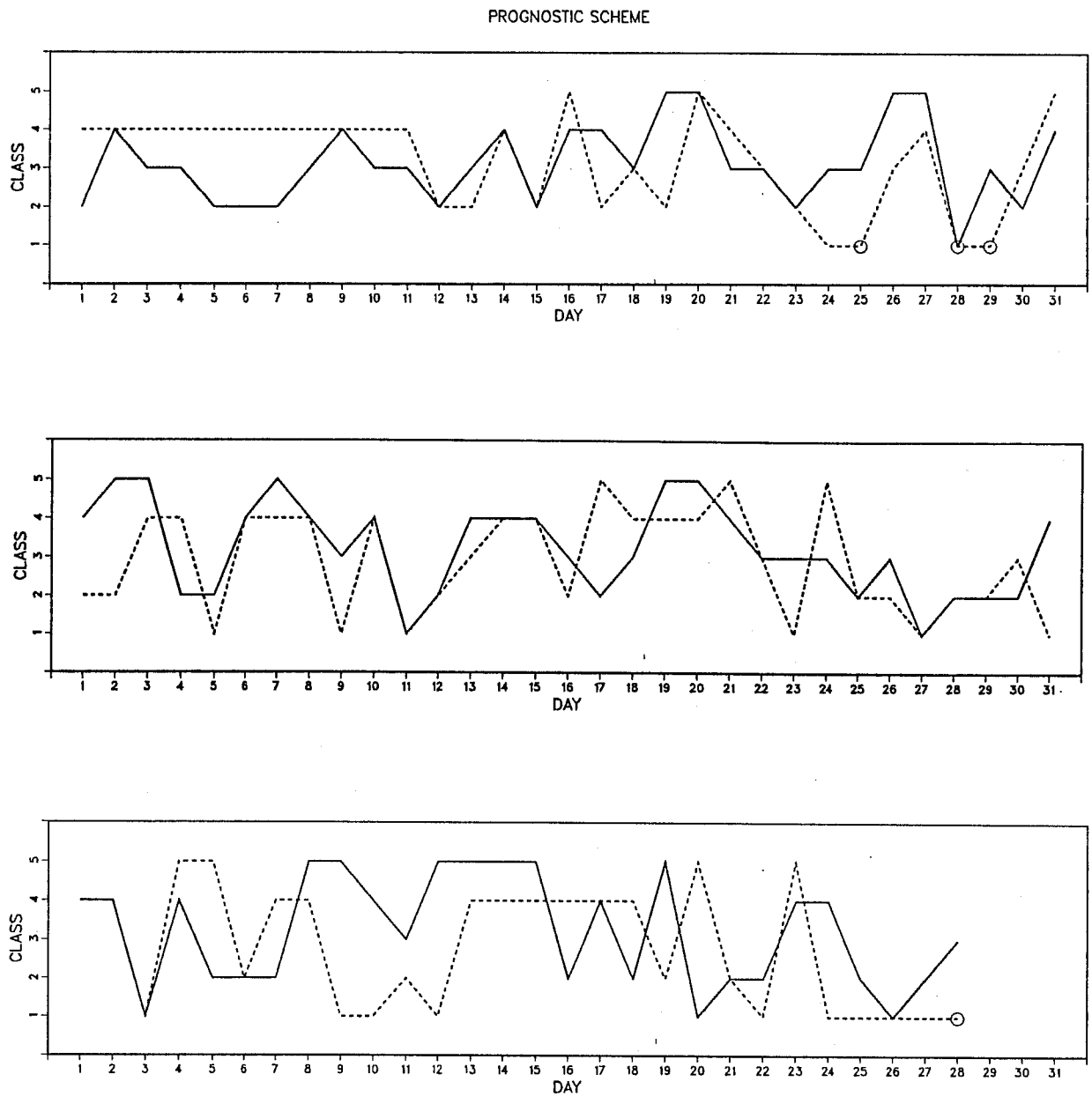

Fig. 8 As in Fig. 5, but for anomaly correlation over Europe.

VERIFICATION OF SKILL PREDICTIONS

FORECAST DAY 5 – AREA NO. 5 – ANOMALY COR.

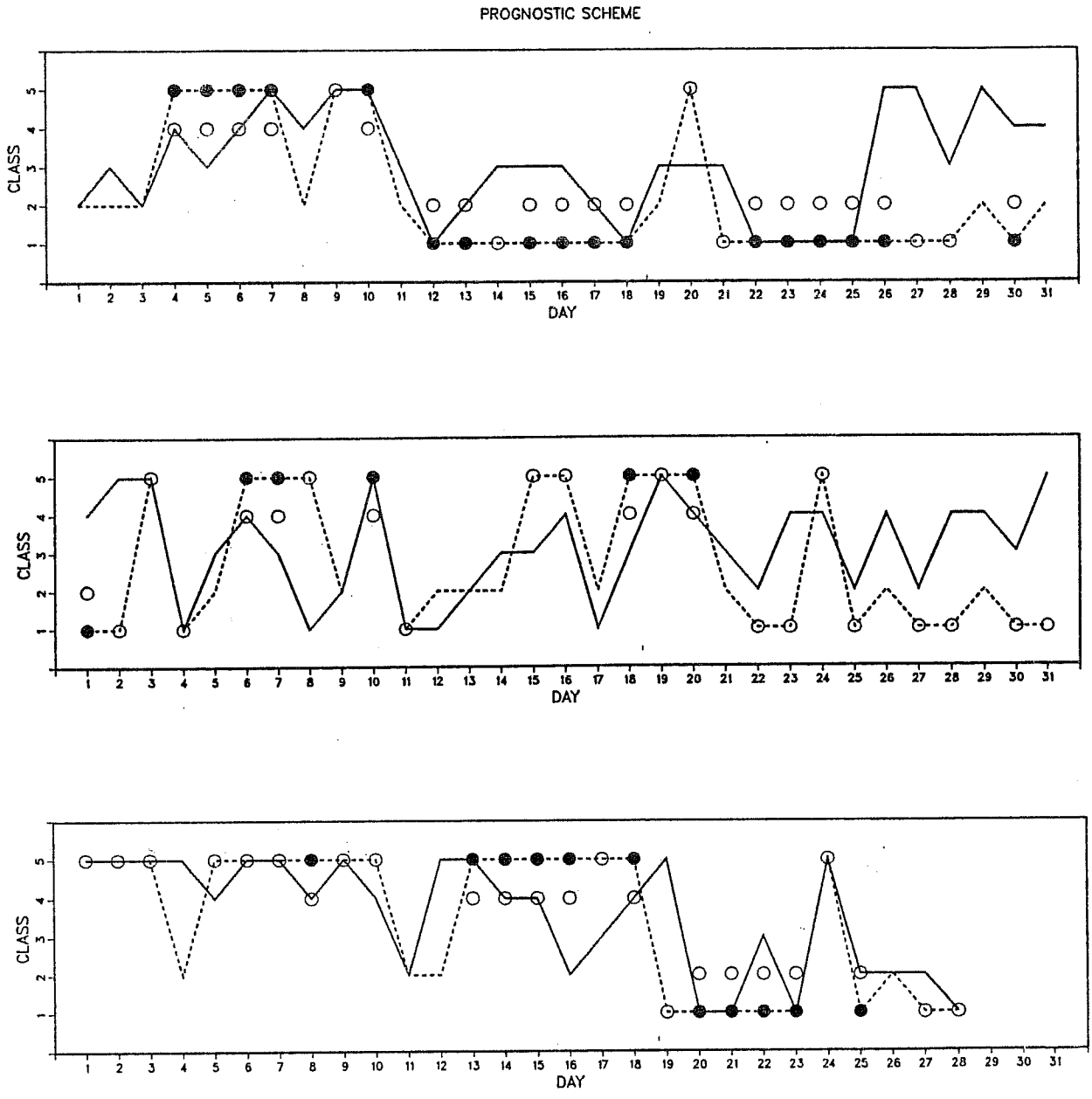

Fig. 9 As in Fig. 5, but for anomaly correlation over South-Western Europe.

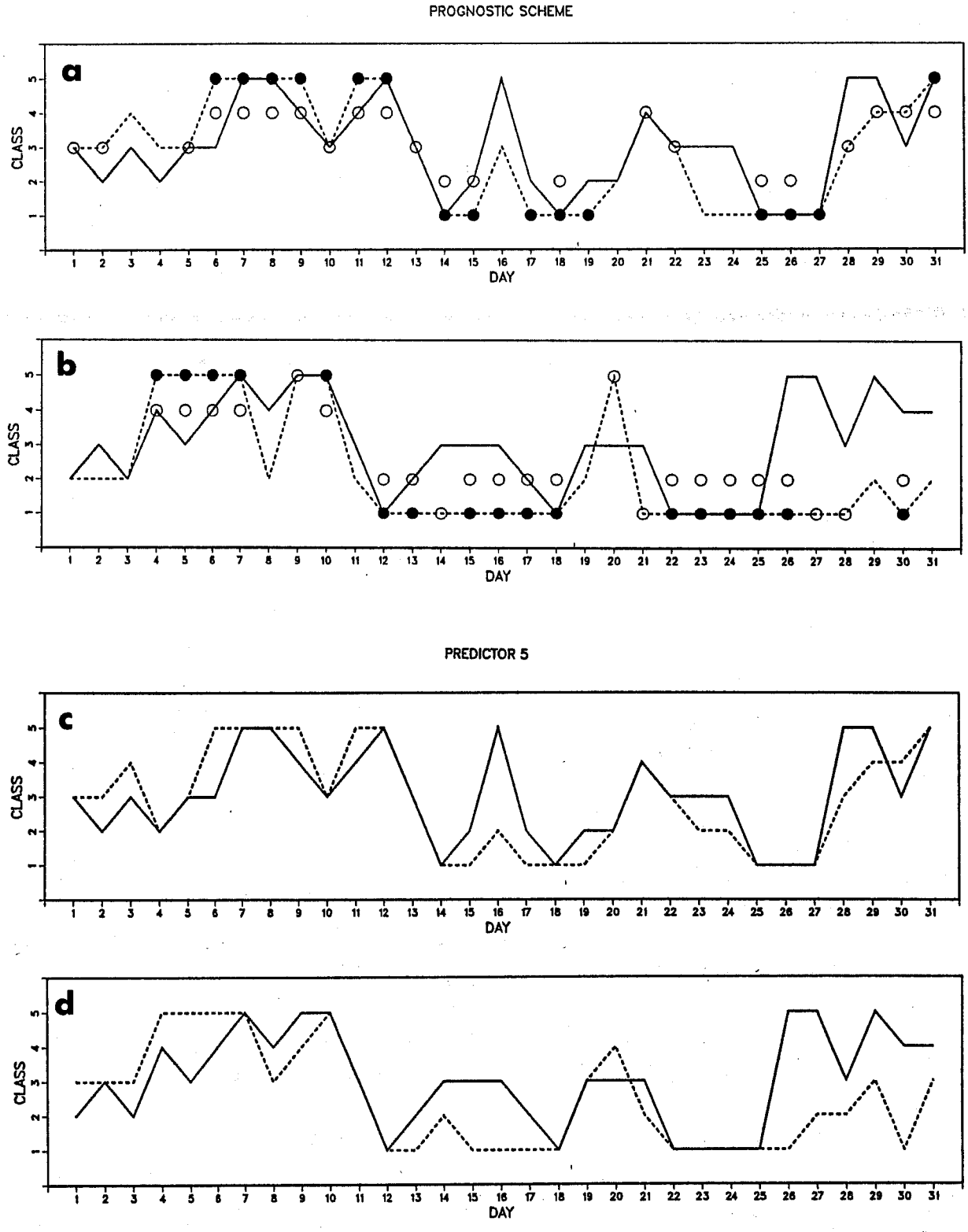

Fig. 10 a) Time series of observed (solid line) and predicted (dashed line) class of ACC of forecast day 3 over South-Western Europe in December 1987. The predicted class is the class with the highest probability as estimated by the prognostic scheme.
b) As in a), but for forecast day 5.
c) As in a), but here the predicted class is assumed as the class of the forecast anomaly amplitude.
d) As in c), but for the forecast day 5.

full circle the days in which the maximum probability exceeds .4. From the proportion of open and full circles one can clearly see that the predictions of RMS error suffer from a greater uncertainty than those for ACC; however, at least some periods of persistent high RMS error were reasonably forecast (see the second half of December 1987, for example). The predictions for ACC are clearly better, and in general the probability distributions have a stronger peak. Over the Northern Hemisphere, the low-frequency variations of ACC are well captured, while over south-western Europe the scheme seems to be able to forecast many high-frequency fluctuations.

The fact that the strong skill of ACC predictions over small areas is almost totally due to the correlation between ACC and anomaly amplitude is clearly demonstrated in Fig. 10. Here, the predictions of ACC over south-western Europe provided by the full prognostic scheme at day 3 and 5 in December 1987 are compared with those obtained using the anomaly amplitude as the only predictor: the results are practically identical. We want to point out that these predictions are not totally trivial, since they imply the existence of a good correspondence between the observed and the predicted amplitude of the anomaly; besides, they can still be useful practically. However, this type of practical skill has little in common with our knowledge about the instability processes that determine the dynamical predictability of the atmospheric motion.

7. THE EFFECTS OF ATMOSPHERIC DYNAMICS

As we discussed in Sect. 2, probabilistic prediction is the most appropriate way of providing a forecast of forecast skill from a scientific point of view. On the other hand, we have seen in the previous section that the most efficient way of extracting valuable indications from our skill predictions was to compute an 'expected class index', which summarizes the probabilistic information in a single number. So, is it really worth adopting the probabilistic approach? Can we say something more about atmospheric predictability with a histogram rather than with a single number, which would be easier to estimate by means of regression techniques and easier to interpret from the forecaster's point of view?

A reasonable compromise might be to provide, in addition to the expected value of the skill score, a standard error of this estimate, that is, the mean and

the standard deviation of the estimated distribution of the score. It is clear that this information is sufficient if the distribution of the score is gaussian, or at least if the probability density is a unimodal function with a known mathematical form. Fig. 11 shows probability density estimates of RMS error at forecast days 3, 6 and 9 over the Northern Hemisphere, deduced from the 180 forecasts in our training sample. The statistical algorithm that produced these curves is dependent on a smoothness parameter, which was set to a sufficiently high value to minimize the probability of getting a multimodal distribution by chance. One can see that at day 3 and day 6 the probability density function is unimodal, with a positive but not excessive skewness, whereas a multimodality appears at day 9.

Other indications can be deduced from the conditional probabilities of error classes for different values of the predictors. Fig. 12 shows probabilities for the 5 classes of RMS error at day 9, computed by averaging the conditional probabilities corresponding to classes 1 + 2 (high expected skill) and 4 + 5 (low expected skill) of a given predictor. The top panels show the histograms for the Northern Hemisphere corresponding to different values of the circulation index, the bottom panels show the histograms for Europe for different classes of spread from the previous forecast. All these histograms are bimodal. In the case of the circulation index, the explanation may be that low values of this index indicate that the forecast is in a more stable region of the phase space, but this does not necessarily mean that the actual atmosphere is in that region too; conversely, even in an unstable situation a good initial analysis can lead the forecast into the right direction. In the case of the spread, small differences between forecasts sometimes occur when both the forecasts fail in reproducing a transition in the circulation pattern; on the other hand, a large discrepancy from the previous forecast may be due to the fact that the earlier forecast, and not the later one, is wrong.

In reality, these explanations may justify the existence of a poor correlation between the predictors and the predictand, but are not sufficient to account for bimodality. Bimodality can be explained if we assume that the distribution of atmospheric states is not unimodal, and that the medium-range errors are mainly due to failures in forecasting transitions between regions

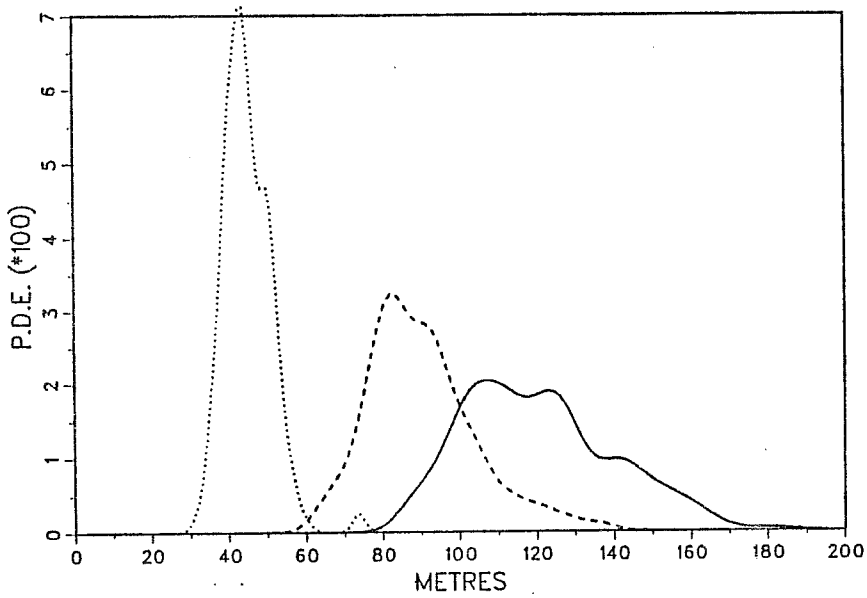

Fig. 11 Probability density estimates of RMS error at forecast day 3 (dotted line), day 6 (dashed line) and day 9 (solid line) over the Northern Hemisphere, as deduced from the 180 forecasts in our training set.

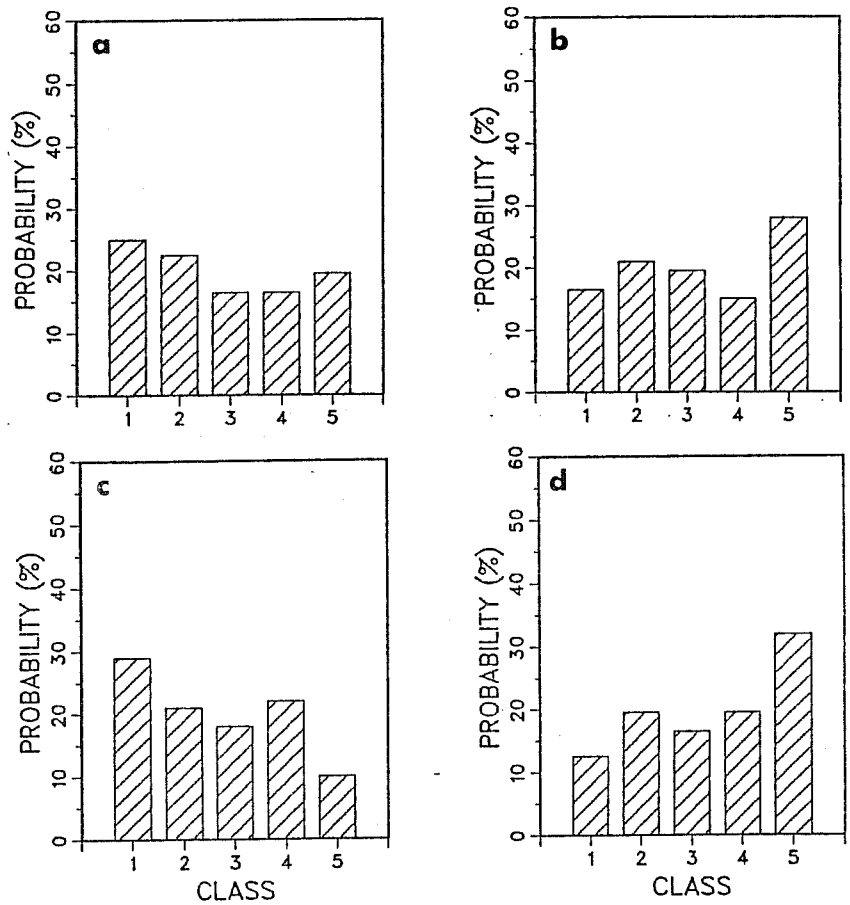

Fig. 12 Conditional probabilities for RMS error classes at forecast day 9.
a) Northern Hemisphere, circulation index in classes 1 or 2;
b) Northern Hemisphere, circulation index in classes 4 or 5;
c) Europe, spread in classes 1 or 2;
d) Europe, spread in classes 4 or 5.

of the phase space where the atmospheric states tend to cluster themselves (see the idealized model discussed in Moritz and Sutera, 1981).

The discovery of bimodality in the amplitude of large-scale planetary waves (see Sutera, 1986; Hansen and Sutera, 1986) provided the first indication of a complex structure in the atmospheric phase space. Recently, Molteni et al. (1988) found 6 clusters in the space generated by 3 rotated empirical ortogonal functions (REOF's) of the large scale eddies, and showed that transitions between these clusters can account for the bimodality in the wave amplitude. However, they also found that this bimodality disappears in the region of the phase space corresponding to positive projections onto the first REOF, while it is is strongly enhanced in the opposite region. This first REOF is shown in Fig. 13a; it can be described as a hemispheric-scale version of the well known Pacific North-American (PNA) pattern. Proofs that medium-range forecasts are more difficult in negative PNA states were discussed by Palmer (1988), and also cases with high values of our circulation index for the Northern Hemisphere (corresponding to large errors) have a negative PNA signature (see Fig. 2).

If the multimodality in the error is a reflection of the multimodality of the distribution of atmospheric states, we should expect to see it more clearly when the observed atmospheric anomaly has a negative projection on the pattern shown in Fig. 13a. We computed probability density estimates for RMS error over the Northern Hemisphere at day 9 stratifing the forecasts in our training data set according to the projection on this pattern. Fig. 13b shows the result for the positive cases, Fig. 13c for negative cases. Only the second density function is multimodal, as expected from the observed structure of the phase space.

We can conclude that a probabilistic approach to the forecast of forecast skill is necessary at least in the medium range, if we want to take account the actual nature of atmospheric dynamics, and that bimodality in the predicted probabilities can reflect a real aspect of the problem of atmospheric predictability.

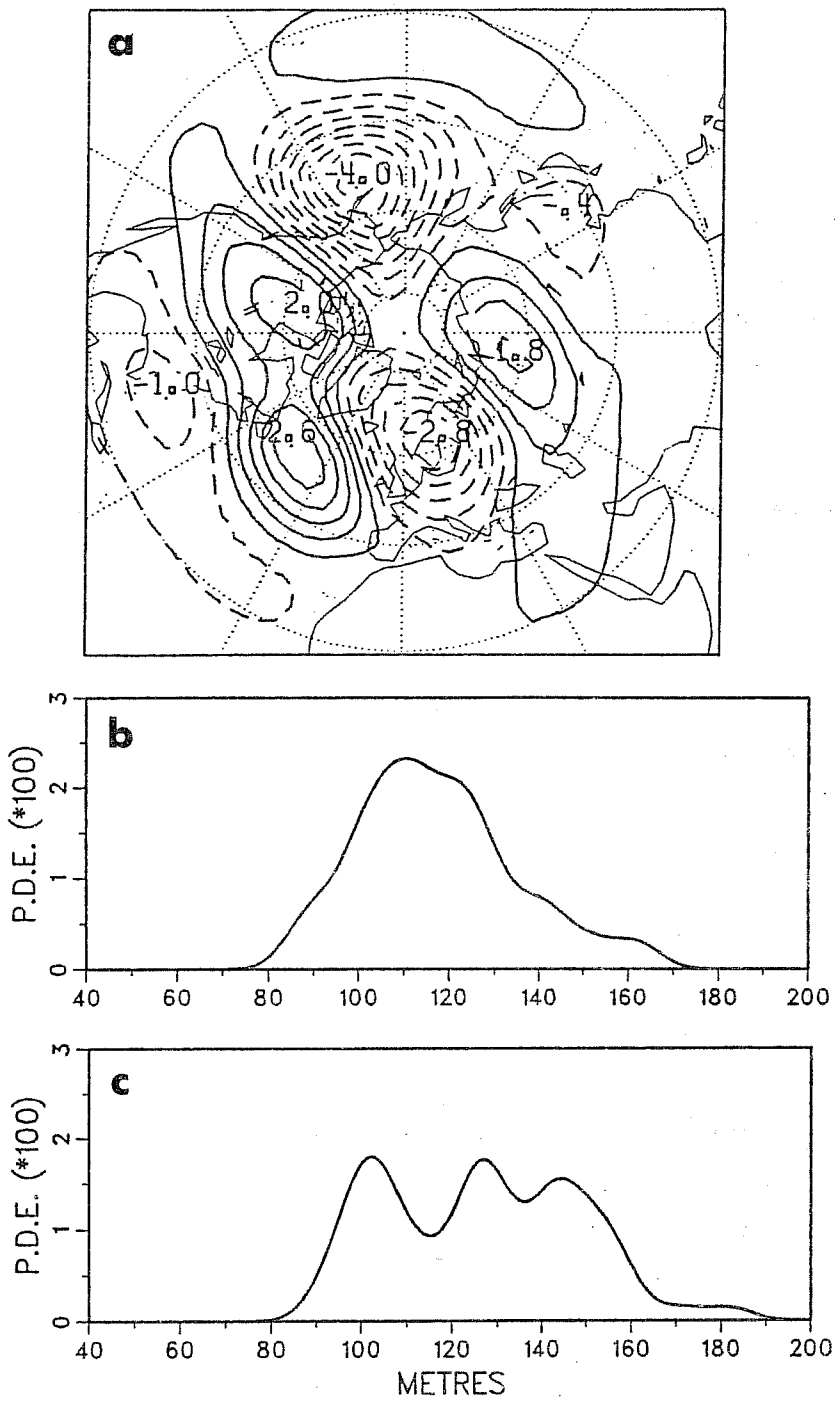

Fig. 13 a) First rotated EOF used for the definition of circulation regimes in Molteni et al. (1988);
b) Probability density estimate for RMS error at forecast day 9 over the Northern Hemisphere, when the observed anomaly has a positive projection on the pattern shown in (a);
c) As in b), but for negative projections of the observed anomaly on the pattern shown in (a).

8. SUMMARY AND CONCLUSIONS

In this paper, we have discussed the methodology and the results of an experimental, quasi-operational prediction of forecast skill performed at ECMWF during winter 1987/88. The predictions consisted in probabilities of occurrence for 5 classes of skill; the classes were defined as categories of RMS error and anomaly correlation coefficient (ACC) for the 500 hPa height field, which occurred with equal frequency during the two previous winters and could therefore be considered as, a priori, equally likely. The skill predictions were issued for 7 verification areas and 4 forecast times (namely, days 3, 5, 7 and 9).

The experimental predictions up to day 5 have provided encouraging results, but a lot of additional work is needed before they can be used operationally. The predictions of RMS error over two large continental areas (North America and Europe) captured at least some low-frequency variations of skill. Better results were obtained for anomaly correlation. For the Northern Hemisphere, the correlation between the observed class of ACC and the 'expected' value computed from the probability distributions was about .5; in some small regions covering part of the European continent, the ACC predictions were able to reproduce even some high-frequency variations. However, this additional skill of the ACC predictions was entirely due to the strong correlation between the ACC itself and the amplitude of the anomaly, and has little to do with our knowledge about the error growth due to atmospheric instability.

The unimodal distribution of the error in the first half of the forecast interval, and the results of our verifications, suggest that the useful information given by a probabilistic forecast of forecast skill up to day 5 can be effectively summarized by an expected value of a given skill score and a standard error of this estimate. This is no more true in the medium-range, when the non-linear nature of atmospheric dynamics produces bimodal distributions of the error, at least for some circulation regimes. Our experimental predictions have shown little skill at days 7 and 9; it is our opinion that more knowledge about the ability of the model in forecasting transitions to anomalous circulation regimes should be incorporated in a statistical prediction scheme, in order to provide useful information about the expected skill in the medium range. In the future, this knowledge will

also be essential for a correct interpretation of the results of more dynamically based methods, such as those based on ensemble forecasting.

Results from this experimental scheme are sufficiently encouraging to warrant an extension for the coming winter 1988/89. Some of the possible shortcomings of the present scheme (e.g. independence of results for RMS error and ACC, and different verification times) will be addressed by using the predictors to estimate parameters in simple error growth equations. Further details will appear elsewhere.

Acknowledgements

We are grateful to S. Tibaldi for constructive discussion at various stages during this work, in particular with regard to the analyses discussed in Section 7. Many useful comments and suggestions regarding the evaluation of the experimental skill predictions came from H. Böttger, B. Strauss, T. Tuna and researchers from the Meteorological Services of various member states, during an expert meeting held at ECMWF in February 1988.

References

Brankovic, C., F. Molteni, T.N. Palmer, S. Tibaldi and U. Cubasch, 1988: Extended range ensemble forecasting at ECMWF. This volume.

Grønaas, S., 1982: Systematic errors and forecast quality of ECMWF forecasts in different large-scale flow patterns. ECMWF Seminar/Workshop on Interpretation of Numerical Weather Prediction Products, 13-24 September 1982. ECMWF, Reading, UK., 161-206.

Hansen, A.R. and A. Sutera, 1986: On the probability density distribution of large-scale atmospheric wave amplitude. J.Atmos.Sci., 32, 3250-3265.

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

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

Kruizinga, S., and C.J. Kok, 1988: Evaluation of the ECMWF experimental skill prediction scheme and a statistical analysis of forecast errors. This volume.

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

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, Reading, UK.

Molteni, F., S. Tibaldi and T.N. Palmer, 1988: Regimes in the wintertime extratropical circulation. I: Observational evidence. Submitted to Q.J.R.Meteor.Soc.

Moritz, R.E., and A. Sutera, 1981: The predictability problem: effects of stochastic perturbations in multiequilibrium systems. Adv. in Geophys., 23, 345-383.

Murphy, J.M., 1988: The impact of ensemble forecasts on predictability. Q.J.R.Meteor.Soc., 114, 463-493.

Palmer, T.N., 1988: Medium and extended range predictability, and stability of the PNA mode. Q.J.R.Meteor.Soc., 114, 691-713.

Palmer, T.N. and S. Tibaldi, 1988: On the prediction of forecast skill. Mon.Wea.Rev., in press.

Shukla, J., 1985: Predictability. Adv. in Geophys., 28B, 87-122.

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

Tibaldi, S., and F. Molteni, 1988: On the operational predictability of blocking. Proceedings of the ECMWF Seminar on 'The nature and prediction of extra-tropical weather systems', 7-11 September 1987, Vol. II, 329-371. ECMWF, Reading, U.K..