Forecasting Forecast Skill

EUGENIA KALNAY† AND AMNON DALCHER*

NASA/Goddard Space Flight Center, Greenbelt, MD 20771

(Manuscript received 20 March 1986, in final form 26 July 1986)

ABSTRACT

We have shown that it is possible to predict the skill of numerical weather forecasts—a quantity which is variable from day to day and region to region. This has been accomplished using as predictor the dispersion (measured by the average correlation) between members of an ensemble of forecasts started from five different analyses. The analyses had been previously derived for satellite data impact studies and included, in the Northern Hemisphere, moderate perturbations associated with the use of different observing systems.

When the Northern Hemisphere was used as a verification region, the prediction of skill was rather poor. This is due to the fact that such large area usually contains regions with excellent forecasts as well as regions with poor forecasts, and does not allow for discrimination between them. However, when we used regional verifications, the ensemble forecast dispersion provided a very good prediction of the quality of the individual forecasts.

Although the period covered in this study is only one month long, it includes cases with wide variation of skill in each of the four regions considered. The method could be tested in an operational context using ensembles of lagged forecasts and longer time periods in order to test its applicability to different areas and weather regimes.

1. Introduction

The skill of both short and long range numerical weather forecasting varies from case to case, from region to region, and depends on other factors such as season, the state of the atmosphere and the boundary anomalies. For this reason, it has been suggested by Tennekes et al. (1986) that "no forecast is complete without a forecast of forecast skill."

There have been several effort towards a probabilistic approach to numerical weather prediction in order to provide an a priori estimation of forecast skill. Epstein (1969) developed a *stochastic-dynamic* prediction scheme including forecast equations for probability distribution of the atmospheric variables. Because of the size of the problem, this method is unfeasible except for the simplest models. Leith (1974) proposed instead the use of *ensemble forecasting,* such as Monte Carlo forecasting (MCF). He suggested that the dispersion of a small number of forecasts ($N \sim 8$) from randomly perturbed intitial conditions is a measure of the forecast skill of the mean forecast. Hoffman and Kalnay (1983) formulated the lagged average forecasting (LAF) method, which has the advantages of MCF but which in a medium range operational context can be attained at virtually no cost. It makes use of not only the latest operational forecast, but also of forecasts with the same verification time started one or more days earlier.

In the LAF forecast, members of the ensemble are weighted statistically. Hoffman and Kalnay (1983) showed that in a low-order model simulation, the LAF forecast was somewhat more accurate than a simple dynamical forecast or MCF forecast. Furthermore, as expected, the spread of the forecasts provided a good estimate of the ensemble mean breakdown time.[1] In a recent application of LAF to operational ECMWF forecasts, Dalcher et al. (1985) found that although LAF showed marked improvements upon the operational dynamical forecasts, the success in predicting individual forecast skill was only minimal. Dalcher et al. attributed this lack of success to the use of global verifications, which could have possibly masked regional variations in skill.

The purpose of this paper is to report the results obtained in a study using ensemble forecasting to predict *regional* forecast skill. We show that the use of regional verification leads to very good predictions of the quality of individual forecasts. However, when the whole Northern Hemisphere is used for verification the quality of the forecast skill prediction is degraded.

2. Data and results

In order to test the feasibility of predicting *regional* forecast skill using ensemble forecasting, we made use of a set of five different analyses and forecasts previ-

† Laboratory for Atmospheres.
* Sigma Data Services Corporation.

[1] However, this result is not conclusive since the dispersion of the forecasts is also included in the ensemble error.

ously utilized in data impact studies (Halem et al., 1982; Kalnay et al., 1986). The analyses, denoted "FGGE," "NOSAT," "NOTEMP," "NOWIND" and "NOCTW" were derived using the Goddard Laboratory for Atmospheres (GLA) four-dimensional analysis/forecast system (Baker, 1983; Kalnay et al., 1983). In the FGGE analysis, all the conventional and satellite data available during the period 5 January to 2 February 1979 was utilized. The NOSAT analysis was performed in a similar way, except that no satellite data (TIROS-N temperature retrievals and cloud tracked winds) were used. In the NOTEMP, NOWIND, and NOCTW, certain data sets were deleted (temperature, wind, and cloud-tracked winds, respectively).

Because of the greater density of conventional data in the Northern Hemisphere, it was observed in the data impact study that the forecasts generally tend to resemble each other, whereas in the Southern Hemisphere the impacts of the data on the forecasts are much larger. For this reason Northern Hemisphere forecasts derived from the five analyses can be considered as members of an ensemble of forecasts whose initial conditions have been moderately perturbed.

In the following attempt to predict forecast skill, we compare the forecasts from the analyses NOSAT, NOTEMP, NOWIND, NOCTW (which for simplicity will be denoted B, C, D, E, respectively) with the forecast from the FGGE analysis (denoted analysis A). Figures 1a, b, c, d present correlation of 500 mb and sea level pressure anomalies, with respect to climatology, corresponding to 5-day forecasts started from 0000 UTC 21 January 1979, verified *regionally* over North America and Europe. The bottom (dashed) curve represents the anomaly correlation of the forecast whose skill we wish to predict, verified against the ECMWF analysis. The top (solid) curve represents the arithmetic mean R of the anomaly correlation ρ of the forecasts from analyses B, C, D, E verified against the forecast from the analysis A. Note that R can be computed at forecast time.

As expected, the forecasts from the different analyses are generally closer to each other than to the ECMWF analysis, indicated by the fact that the solid line corresponds to a higher correlation than the dashed line. Nevertheless, there is a degree of parallelism between the two curves, suggesting that, indeed, the dispersion between the members of an ensemble of forecasts can be used to predict the breakdown of the forecast skill. For example, over Europe, where the forecast is excellent with an anomaly correlation ρ above 0.8 for most of the 5 days, the average anomaly correlation of forecast versus forecast R remains above 0.9. Over North America, the forecast remains accurate at 500 mb, but it breaks down at sea level pressure, with the correlation becoming smaller than 0.6 after only 2.75 days. The verification of forecast versus forecast reflects this deterioration, becoming less than 0.85 at about 3.2 days.

In what follows we use the time at which the average forecast/forecast correlation crosses the value $R = 0.85$ as predictor of the time of forecast skill breakdown. The predictand is the observed time of forecast breakdown, i.e., the time at which the forecast correlation with analysis becomes less than $\rho = 0.60$. We have used results from 14 forecasts started every other day beginning on 0000 UTC 7 January 1979 and ending on 0000 UTC 2 February 1979. The verifications were performed over four regions: North America, Europe, North Atlantic and North Pacific, and for both sea level pressure (SLP) and 500 mb heights. Figure 2 pre-

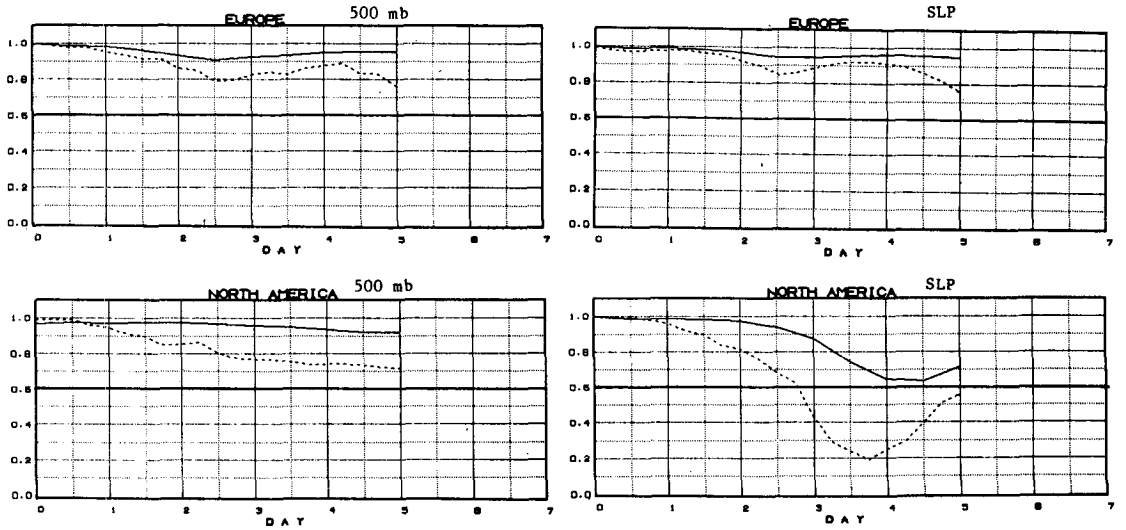

FIG. 1. Anomaly correlations corresponding to 5-day forecasts from 21 January 1979. Solid: Average forecast-forecast correlation; dashed; forecast-analysis correlation.

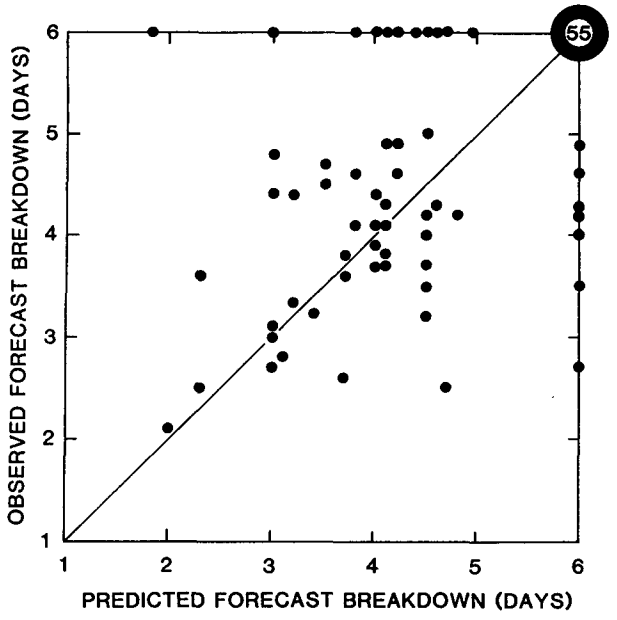

FIG. 2. Scatter diagram of predicted vs observed time of forecast breakdown. Breakdowns after 5 days are arbitrarily plotted as 6 days. The 55 cases of breakdown both predicted and observed to take place after 5 days are represented by a large dot.

sents a scatter diagram of the predicted time of forecast breakdown versus the observed time. Since the forecasts were only 5 days long, forecasts that remained accurate for more than 5 days are arbitrarily represented by a breakdown at 6 days. The large dot represents 55 cases of breakdown both predicted and observed to take place after 5 days. In only 11 cases the forecast remained skillful for more than 5 days but was incorrectly predicted to break down before 5 days. Of the 46 cases of early breakdown, 39 were correctly predicted to have skill lasting less than 5 days, and only 7 were predicted to last more than 5 days. These data are represented schematically in Table 1.

Table 1 suggests that we can skillfully predict whether an individual forecast will remain good for 5 days, i.e., whether its correlation against analysis will stay above 0.6. In order to judge the performance of the quality prediction we need to compute

$$p = (n_{11} + n_{12})/(n_{11} + n_{12} + n_{21} + n_{22})$$

$$p_1 = n_{11}/(n_{11} + n_{21}), \text{ and } p_2 = n_{12}/(n_{12} + n_{22})$$

were n_{ij} are the elements of Table 1. Thus, p is the proportion of good forecasts, i.e., those that did not break down in 5 days among all forecasts; p_1 is the proportion of good forecasts among the forecasts predicted to be good; p_2 is the proportion of good forecasts among the forecasts predicted to be poor. The a priori skill prediction is successful if $p_1 > p > p_2$. Direct calculation shows that each of the inequalities $p_1 > p$, $p > p_2$, $p_1 > p_2$ is equivalent to the other two and also

TABLE 1. Contingency table indicating the proportion of forecasts which remained skillful ($\rho \geq 0.60$) for more than 5 days vs those predicted to remain skillful beyond 5 days. The table is based upon 14 forecasts verified over 4 regions, both at 500 mb and sea level pressure (SLP).

	Predicted breakdown	
Observed breakdown	≥5 days	<5 days
≥5 days	55	11
<5 days	7	39

to $n_{11}n_{22} > n_{12}n_{21}$. Therefore the method is successful if the determinant of the contingency table is positive. In the case of Table 1 we have $p = 0.59$, $p_1 = 0.89$ and $p_2 = 0.22$ so that, indeed, $p_1 > p > p_2$ and the method is successful. We can estimate the statistical significance of this success by using a standard test for the difference between two proportions (Snedecor and Cochran, 1967, Chapter 8). The test is based on the statistic

$$Z = \frac{p_1 - p_2}{[p(1-p)(1/N_1 + 1/N_2)]^{1/2}} \quad (1)$$

which, under the null hypothesis of no prediction skill ($p_1 = p_2 = p$), has a normal distribution with mean zero and standard deviation 1. Here $N_1 = 55 + 7$, and $N_2 = 11 + 39$ are the number of forecasts predicted to be good and poor, respectively. A continuity correction due to the fact that the entries in the matrix of Table 1 are integers, while Z has a continuous distribution, is made by subtracting $0.5(1/N_1 + 1/N_2)$ from the numerator in (1) (if $p_1 > p_2$). The value of Z derived from Table 1 is 6.94, which corresponds to a one-sided significance probability of 10^{-11}. Even if the number of degrees of freedom is reduced by taking into account the fact that we have not used an independent sample, and that there is some dependence between the 500 mb and SLP verifications, it is clear the results are still highly significant.

Similar contingency tables are presented separately for each verification region in Table 2. The large value of the diagonals show that the predictor provides excellent discrimination between the skillful forecasts lasting either less or more than 5 days in each of the regions. On the other hand, when we verify over the whole Northern Hemisphere, that discrimination is lost, as seen in Table 3.

In this case, $p = 0.64$, $p_1 = 0.61$, $p_2 = 0.70$, and $Z = -0.06$, with a one sided probability of 0.52, confirm-

TABLE 2. As in Table 1 but for the given individual verification regions.

North America		Europe		North Atlantic		North Pacific	
8	4	14	1	19	1	14	5
2	14	2	11	2	6	1	8

TABLE 3. As in Table 1 but for the whole extratropical Northern Hemisphere.

11	7
7	3

ing the lack of a priori prediction skill, and even showing a negative value of Z associated with the violation of the condition for skill $p_1 > p > p_2$. Table 4 summarizes the values of the conditional probabilities p_1, p, p_2, the Z statistics and the significance probability for Tables 1, 2, and 3.

We see that using regional verifications we are very successful in the qualitative prediction of forecast skill, whereas the method does not work with large verification areas, which are affected by different synoptic systems and include regions of both good and poor forecast skill. It should be noted that the skill of the predictions is not sensitive to the precise value of R chosen as a threshold. Similar skills were obtained varying R by up to 0.05.

It is interesting to point out that we were less successful when we attempted to use other measures of forecast skill such as rms errors and S1 scores. The advantages of the use of correlations may be due to the fact that correlations are bounded. If we view each forecast as a many-dimensional vector, the correlation between two forecasts is the cosine of the angle between the corresponding vectors. Since for angles that are not too large, $\cos\alpha \approx 1 - \alpha^2/2$, we see that the average correlation R, that we have chosen as predictor, is indeed a measure of scaled spread among the forecasts.

3. Shorter range prediction

We now apply a similar method to determine a priori the time at which a forecast ceases to be "very good", defined as the time the regional anomaly correlation verified against the ECMWF analysis reaches the value $\rho = 0.80$. An inspection of the 112 forecasts resulted in the choice of the time at which the average forecast/forecast correlation reaches $R = 0.92$ to be used as predictor.

Figure 3 is a scatter diagram of prediction versus observed time at which the forecasts cease to be very good. Those cases in which the crossing occurred after 5 days were arbitrarily assigned to 5.5 days. Again, we observe excellent a priori predictive skill. We can construct a table (Table 5) similar to Table 1, in which we count the proportion of forecasts predicted and observed to remain very good beyond 3.5 days.

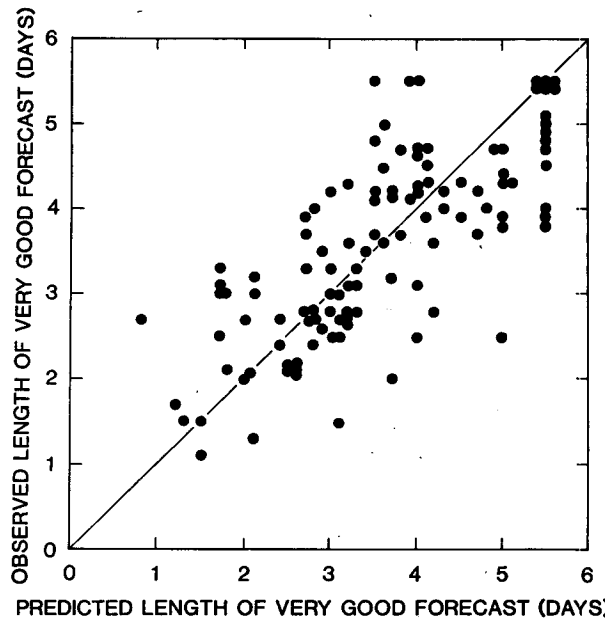

FIG. 3. Scatter diagram of predicted vs observed length of very good forecast, corresponding to an anomaly correlation of more than 80%. Forecasts remaining very good beyond 5 days are arbitrarily plotted at 5.5 days.

The statistic value derived from Table 2 is $Z = 7.73$, which has a one-sided significance probability of 1.2×10^{-15}. Similar tables (Table 6) for the individual regions also show excellent skill.

In this case, even the verification over the whole extratropical Northern Hemisphere indicates some skill as shown in Table 7.

Table 8 summarizes the values of the conditional probabilities, the Z statistics and its significance probability for the contingency Tables 5, 6, and 7.

Similar skills were obtained when the value of R was perturbed by up to 0.02. At this shorter range R is closer to its maximum value of one and therefore there is less room for variations.

TABLE 4. Partial probabilities, Z statistics and significance probability for the proportion of forecasts that remain good ($\rho \geq 0.60$) beyond 5 days.

Region	p_1	p	p_2	Z	prob (Z)
Four regions	0.89	0.59	0.22	6.94	2.10^{-11}
North America	0.80	0.43	0.22	2.96	5.10^{-3}
Europe	0.88	0.54	0.08	3.77	8.10^{-5}
North Atlantic	0.90	0.71	0.14	3.86	6.10^{-5}
North Pacific	0.93	0.68	0.38	3.69	1.10^{-3}
Northern Hemisphere	0.61	0.64	0.70	−0.06	0.52

TABLE 5. Contingency table indicating the proportion of forecasts that remained very good ($\rho \geq 0.80$) for more than 3.5 days among those predicted to remain very good beyond 3.5 days. The data included 14 forecasts verified over 4 regions, both at 500 mb and SLP.

	Predicted very good skill	
Observed very good skill	≥3.5 days	<3.5 days
≥3.5 days	54	8
<3.5 days	6	44

TABLE 6. As in Table 5 but for the given individual regions.

North America		Europe		North Atlantic		North Pacific	
7	2	13	2	18	2	16	2
1	18	4	4	1	7	0	10

4. Relationship between predicted and observed correlations

Based on the results presented so far, we constructed a correspondence between forecast/analysis and forecast/forecast correlation using the following procedure: For each value of the average forecast/forecast correlation R, we determined the median time of crossing T, i.e., the time at which half of the forecasts reached the value R at least once. The median forecast/analysis correlation ρ was then determined such that at time T, half of the forecasts had not yet reached ρ.

Figure 4 presents the derived relationship $\rho(R)$. Since the forecasts were only 5-days long, we only have data for $\rho \geq 0.66$ and $R \geq 0.86$. However, because of the almost linear relationship between the two correlations apparent in Fig. 4, we adopted the following calibration formula:

$$\rho = 2.205R - 1.215 \qquad (2)$$

which we also used to extrapolate below $R = 0.86$.

We now present for each of the 14 forecasts over Europe the observed anomaly correlation ρ and the predicted correlation $\rho(R)$ derived from (2). Figure 5 contains the sea level pressure results, and Fig. 6 the results for 500 mb heights. As indicated by Tables 4 and 8, the results over Europe are representative of the four regions, being somewhat better than those over North America, and slightly worse than the results over the North Atlantic. It should be noted that the accuracy in skill prediction is similar at both levels, and that the sample includes cases with high as well as low predictability. These observations are also true in the other three regions. Although there are cases of skill prediction that are too optimistic or too pessimistic, Figs. 5 and 6 indicate that, overall, our method provides estimates of day-to-day skill that would be useful in an operational context.

5. Discussion

We have shown that is it possible to predict the skill of numerical weather forecasts using as predictor the dispersion between the members of an ensemble of forecasts. When the Northern Hemisphere was used as a verification region, the prediction of skill was rather poor. This is due to the fact that such large area usually includes regions with excellent forecasts as well as regions with poor forecasts, and does not allow for discrimination between them. However, when we used regional verifications, the ensemble forecast dispersion provided a very good prediction of the quality of the individual forecasts.

Our results are clearly preliminary, since they cover only a period of 1 month, and may not hold under all circumstances. Nevertheless, they are very encouraging and suggest that indeed, the dispersion of the members of a small ensemble of forecasts can be a good a priori predictor of forecast skill, even if the ensemble is as small as three or four. (In our ensemble, the forecast from the NOCTW analysis remained almost invariably very close to the FGGE forecast, and therefore contributed little discrimination.)

In this paper we pooled together data from different geographical regions and from two vertical levels. Because of this we may have introduced some dependence

TABLE 7. As in Table 4 but for the whole extratropical Northern Hemisphere.

11	3
5	9

TABLE 8. As in Table 4 but for the proportion of the forecasts that remain very good ($\rho \geq 0.8$) beyond 3.5 days.

Region	p_1	p	p_2	Z	prob (Z)
Four regions	0.90	0.55	0.15	7.73	5.10^{-15}
North America	0.88	0.32	0.10	3.52	2.10^{-4}
Europe	0.76	0.54	0.18	2.63	4.10^{-3}
North Atlantic	0.95	0.71	0.22	3.52	2.10^{-4}
North Pacific	1.00	0.64	0.17	4.16	2.10^{-5}
Northern Hemisphere	0.69	0.50	0.25	1.91	3.10^{-3}

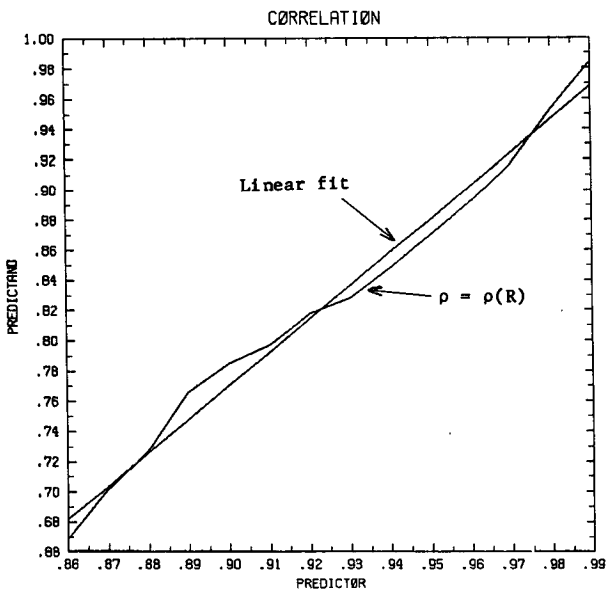

FIG. 4. Relationship between the average forecast/forecast correlation R and the forecast/analysis correlation ρ derived using the median method, and its linear fit.

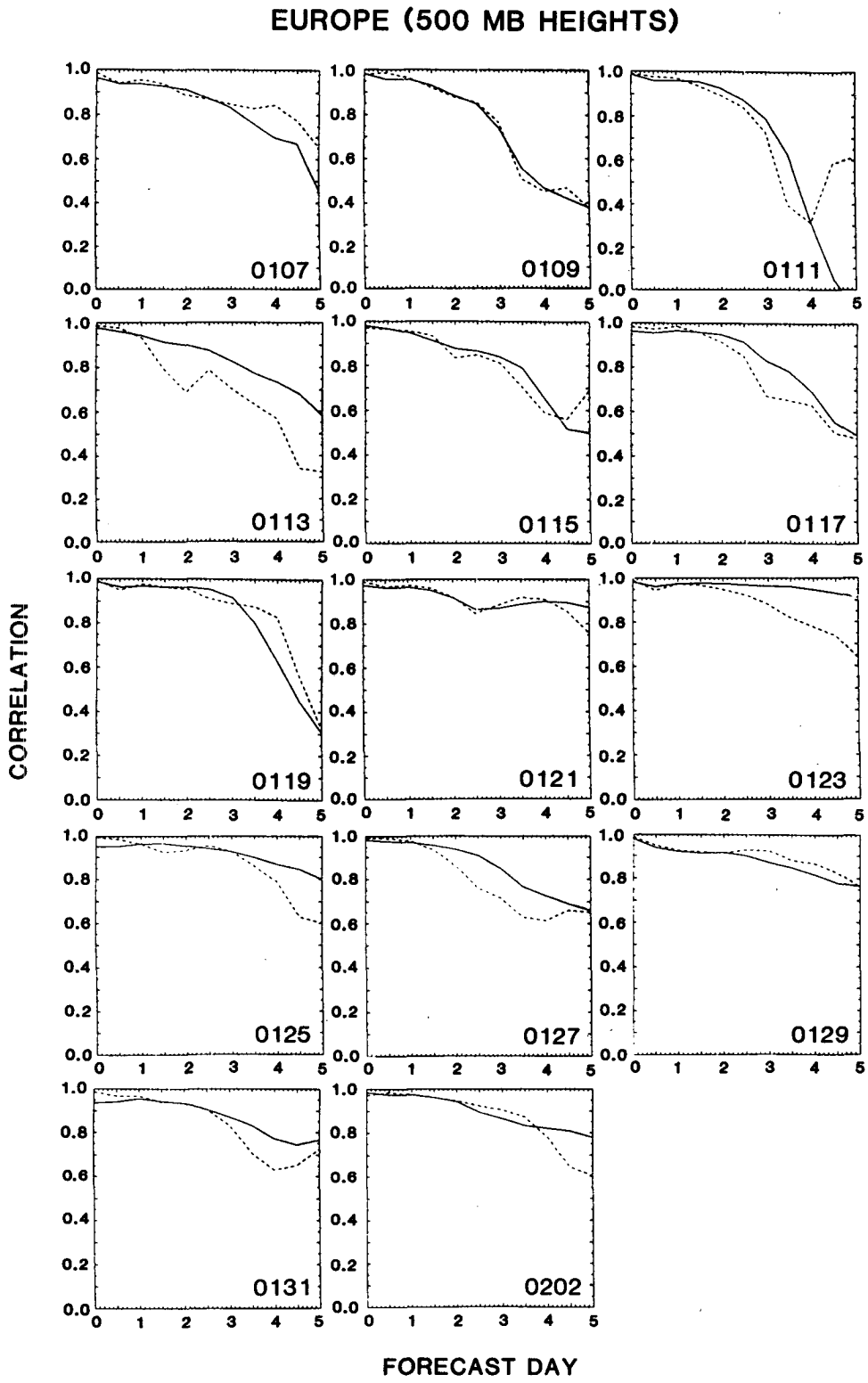

FIG. 5. Comparison of the sea level pressure observed (dashed) and predicted (solid) anomaly correlation for each of the 14 5-day forecasts verified over Europe. The abscissa is forecast time, the ordinate is correlation, and the month and date of the initial conditions are indicated by the four digits inside each panel.

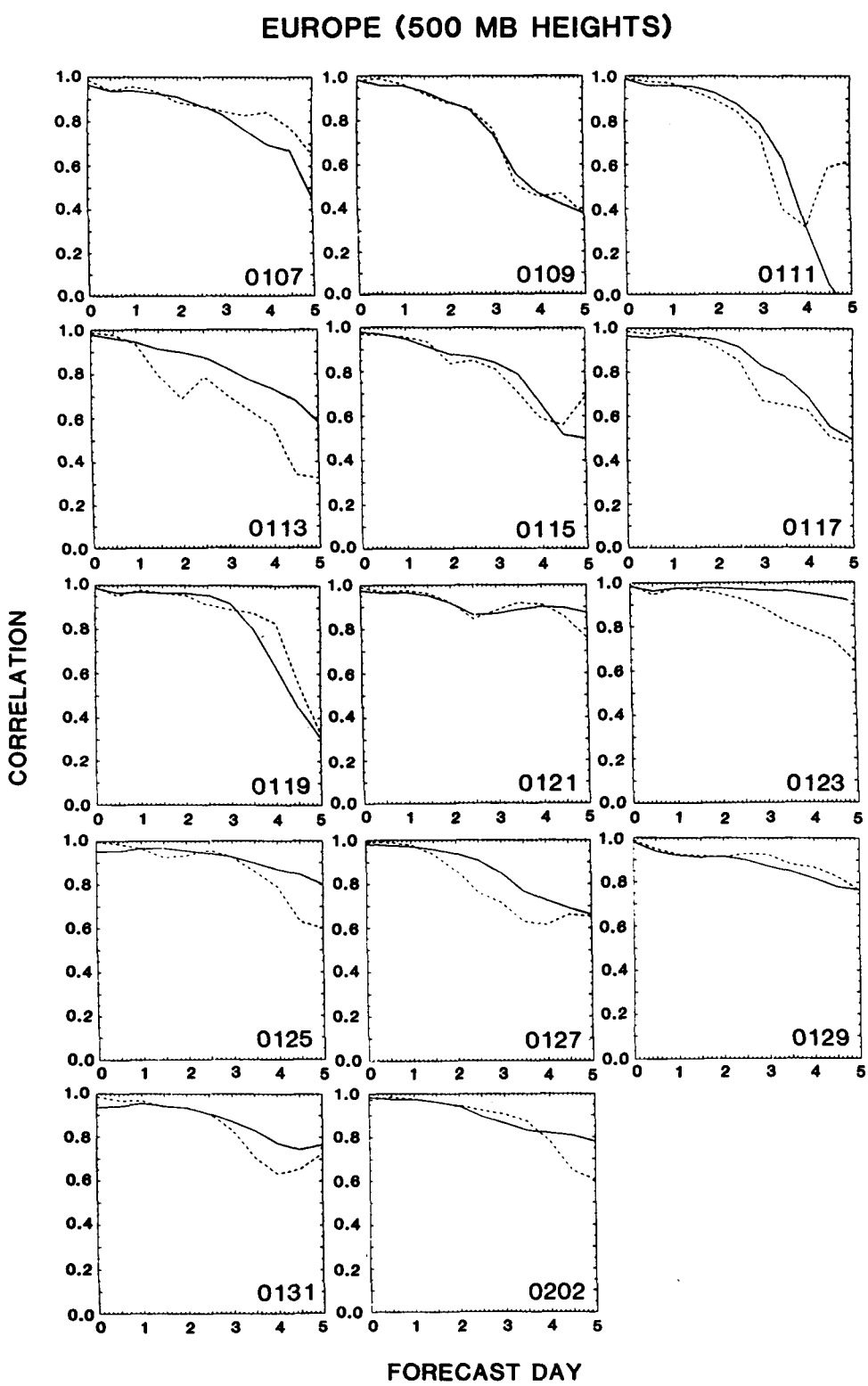

FIG. 6. As in Fig. 5 but for 500 mb heights.

in the sample and therefore overestimated the level of significance. To study the robustness of our conclusions we tested separately each vertical level in each of the four verification regions. Even with only 14 forecasts in each set, the results were all significant at the 95% level except for the sea level pressure over North America which was significant at the 90% level.

The fact that we used the same criterion $R(\rho)$ and that the results are not sensitive to the precise value chosen suggests that the method will have applicability beyond this dataset.

The physical reasons of the variations in forecast skill should be studied in depth. Two processes that can result in reduced regional predictability are instabilities (such as cyclogenesis), and changes of weather regimes (such as the development or decay of a blocking high). In both cases one may expect that the increased difficulty of predicting a sudden transition will be also associated with a larger dispersion among the members of an ensemble of forecasts. This, of course, is the basic idea behind the use of ensemble forecasting for skill prediction. More understanding of the processes involved should come from a studies of the temporal and spatial variability of atmospheric instabilities (Tennekes et al., 1986) and of the characteristics and typical exit times of weather regimes (Mo and Ghil, 1985).

The present procedure to predict forecast skill has been applied to an ensemble of moderately perturbed analyses, which is akin to the use of Monte Carlo Forecasting (Leith, 1974), although the differences between analyses reflect data uncertainties rather than random errors. The results of Hoffman and Kalnay (1983) using a low-order model suggest that the use of lagged average forecast ensembles could be equally effective, although it may be necessary to perform more frequent medium range forecasts than once a day. The method should be tested in an operational context, using longer time periods in order to test the extent of its applicability to different areas, seasons and weather regimes.

Acknowledgments. We acknowledge helpful comment from R. N. Hoffman, R. Livezey, H. Tennekes and an anonymous reviewer. We are very grateful to Manina Almeida and Robert Rosenberg for processing the anomaly correlations. The figures were prepared by Laura Rumburg and the text typed by Benita Richardson, Mary Ann Wells and Donna Candido.

REFERENCES

Baker, W. E., 1983: Objective analysis and assimilation of observational data from FGGE. *Mon. Wea. Rev.*, **111,** 328–342.

Dalcher, A., E. Kalnay, R. Livezey and R. N. Hoffman, 1985: Medium Range Lagged Average Forecasts. *Preprints from the Ninth AMS Conference on Probability and Statistics in Atmospheric Sciences.* Virginia Beach, pp. 130–136.

Epstein, E. S., 1969: Stochastic-dynamic prediction. *Tellus,* **21,** 739–759.

Halem, M., E. Kalnay-Rivas, W. E. Baker and R. Atlas, 1982: An assessment of the FGGE Satellite Observing System during SOP-1. *Bull. Amer. Meteor. Soc.,* **63,** 407–426.

Hoffman, R. N., and E. Kalnay, 1983: Lagged Average Forecasting: An alternative to Monte Carlo Forecasting. *Tellus,* **35A,** 100–118.

Kalnay, E., Balgovind, W. Chao, D. Edelmann, J. Pfaendtner, L. Takacs and K. Takano, 1983: Documentation of the GLAS Fourth Order GCM. NASA Tech. Memo. 86064, NTIS N8424048.

——, J. C. Jusem and J. Pfaendtner, 1986: The relative importance of mass and wind data in the FGGE observing system. *Reprints from the AMS National Conference on the Scientific Results of FGGE.* Miami, FL.

Leith, 1974: Theoretical skill of Monte Carlo Forecasts. *Mon. Wea. Rev.,* **102,** 409–418.

Mo, K. C., and M. Ghil (1986): The statistics and dynamics of persistent anomalies. *Preprints of the Ninth AMS Conference Probability and Statistics in Atmospheric Sciences.* Virginia Beach, 416–421.

Snedecor, G. W., and W. G. Cochran, 1967: *Statistical Methods.* 6th edition, 1978 printing, Iowa University Press.

Tennekes, H., A. P. M. Baede and I. D. Opsteegh, 1986: Forecasting forecast skill. *Proceedings of the ECMWF Workship on Predictability in the Medium and Extended Range,* Reading, England.