Skill and Return of Skill in Dynamic Extended-Range Forecasts

JEFFREY L. ANDERSON

Geophysical Fluid Dynamics Laboratory, Princeton, New Jersey

HUUG M. VAN DEN DOOL

CAC/National Meteorological Center, Washington, D.C.

(Manuscript received 15 April 1993, in final form 27 August 1993)

ABSTRACT

The skill of a set of extended-range dynamical forecasts made with a modern numerical forecast model is examined. A forecast is said to be skillful if it produces a high quality forecast by correctly modeling some aspects of the dynamics of the real atmosphere; high quality forecasts may also occur by chance. The dangers of making a conclusion about model skill by verifying a single long-range forecast are pointed out by examples of apparently high "skill" verifications between extended-range forecasts and observed fields from entirely different years.

To avoid these problems, the entire distribution of forecast quality for a large set of forecasts as a function of lead time is examined. A set of control forecasts that clearly have no skill is presented. The quality distribution for the extended-range forecasts is compared to the distributions of quality for the no-skill control forecast set.

The extended-range forecast quality distributions are found to be essentially indistinguishable from those for the no-skill control at leads somewhat greater than 12 days. A search for individual forecasts with a "return of skill" at extended ranges is also made. Although it is possible to find individual forecasts that have a return of quality, a comparison to the no-skill controls demonstrates that these return of skill forecasts occur only as often as is expected by chance.

1. Introduction

Since the advent of successful numerical weather predictions there has been a constant quest to extend the range for which useful numerical forecasts can be produced. Attempts to produce forecasts for lead times up to two weeks were pioneered by Miyakoda et al. (1972). Although success in forecasts for lead times past ten days has generally been limited, there have been some examples of highly successful forecasts of monthly means (Miyakoda et al. 1983). More recently, many of the world's operational prediction centers have become interested in the prospect of long-range numerical forecasts. At the National Meteorological Center (NMC) this has led to a number of large experiments on dynamical extended-range forecasting (DERF) (Tracton et al. 1989).

The task of evaluating the skill of a forecast can be as involved and complex as producing the forecast itself. Even for simple point forecasts of a single discrete variable, the problem of verifying forecasts is not entirely straightforward (Murphy and Winkler 1987). Here, a much more complex problem is of interest: evaluating the largest lead times for which a dynamical forecast of the global circulation has any predictive skill. Clearly, it is not possible to completely solve this problem; however, the method presented here should be able to produce a considerable amount of new information.

It is helpful to distinguish between the concepts of forecast skill and forecast similarity. A forecast is said to have skill if it is quite similar to some verifying observation because the forecast method (a numerical model in the examples presented here) has successfully captured some portion of the dynamics of the real atmosphere. On the other hand, forecasts that are similar to a verifying observed field can also be generated by chance. For instance, randomly selecting a field from a set of historical observations is a forecast method that clearly has no skill by the preceding definition. Nevertheless, there will be some such random forecasts that are somewhat similar to their verifying observations purely by happenstance. Determining whether a forecast is skillful is especially difficult for lead times near the limits of skillful prediction. In this case, a skillful model may produce forecasts that are, on the average, only slightly more similar to the verifying observations than are forecasts produced by some control with no skill.

Corresponding author address: Dr. Jeffrey L. Anderson, Geophysical Fluid Dynamics Laboratory, Princeton University, P.O. Box 308, Princeton, NJ 08542.

Saha and Van den Dool (1988) have already pointed out some of the subtleties involved in evaluating predictive skill. They demonstrated that a numerical weather prediction (NWP) model of late 1980s vintage lost the ability to predict the 1-day time tendency of the circulation field at leads of less than 10 days. They used a comparison to control forecasts with no skill at predicting the time tendency (persistence of the forecast valid one day previously) to evaluate when forecasts of the time tendency were no longer skillful.

In this paper, a method is presented to help distinguish between skillful forecasts and unskillful forecasts that may appear skillful due to random effects. This is done by comparing a scalar measure [the anomaly correlation (AC) defined in section 3] of the similarity between observed and forecast fields for a set of DERF forecasts to ACs for a similar set of control "forecasts" that is known a priori to have no useful skill. To be skillful, the DERF forecasts must be clearly superior to the no-skill control. This idea is similar to the traditional use of persistence or climate mean forecasts as no-skill controls to which dynamical forecasts can be compared. Hoffman and Kalnay (1983) suggested a more sophisticated use of a no-skill control but did not have a chance to test their proposed methodology.

A simple example of the proposed method is a comparison to analogs (Van den Dool 1989). There is only limited similarity between any observation and its best historical analog. Therefore, any forecast that is more similar to its verifying observation than the best natural analogs is potentially skillful, no matter what its lead time.

In the following, the DERF forecasts are first compared to a control set as a function of lead time in an attempt to find the longest lead times for which skill exists. The 5% and 95% limits and mean of the DERF forecast AC distribution are compared to the control as a function of lead. In addition, a statistical test is applied to compare the entire AC distributions. In this fashion, an upper bound can be placed on the extreme lead limits for which statistically significant forecast skill exists.

Since only a limited number of extended-range forecasts have been made to date with modern sophisticated NWP models, very little is yet known about the behavior of forecasts past 10 days lead time. One of the recurring themes in discussions of such long lead forecasts is the possibility of forecasts exhibiting the elusive "return of skill." A forecast with return of skill would exhibit a minimum of AC at a lead time of n days and then an increase of skill thereafter. In earlier DERF experiments, Tracton et al. (1989) found suggestions of return of skill. Their Fig. 30 shows a return of skill in ACs for 10-day average forecasts. The theme of return of skill is often discussed in meetings, sometimes in jest, but has rarely found a place in the published literature.

Nevertheless, there are some possible explanations for a return of skill. For instance, if a model successfully reproduces some wavelike phenomenon but with an error in phase speed, the model wave will be periodically in phase with the observed wave resulting in periodic higher AC. Palmer (1993) has recently pointed out another possible cause of return of skill. In his case, the apparent return is simply a result of random chance and would be discounted by the method presented here.

Examples of individual forecasts with an apparent return of skill can easily be found in the DERF90 experiment. It is important to examine if these return of high AC cases are statistically significant or just a result of random chance. Again, the same method is applied by comparing possible return of skill cases in DERF forecasts to randomly occurring return of higher AC cases from the no-skill control. A significantly greater number of return of high AC cases in DERF than in the control would be evidence of return of skill.

Section 2 presents the datasets used to examine skill and return of skill, while section 3 develops the measures used for verification of individual forecasts. Section 4 explains the no-skill control forecasts comparison that is applied in section 5. Section 6 presents a more robust statistical technique for comparing forecasts with the controls. Section 7 examines return of skill and is followed by conclusions in section 8.

2. Data

The datasets used to examine extended-range forecast skill are described in this section. Unless otherwise stated, all results presented are for 500-mb height fields obtained from archived spectral T21 representations. In addition, 300- and 700-mb height fields and streamfunction fields at all three levels were examined in an identical fashion. The results were not qualitatively affected by the choice of level or field.

a. DERF90

The DERF90 dataset, produced at NMC by Saha, Kalnay, Kanamitsu, and Van den Dool, consists of a series of extended-range forecasts produced by NMC's Medium-Range Forecast (MRF) Model. Comprehensive information about the MRF can be obtained from Kalnay et al. (1990) and Kanamitsu et al. (1990). In the DERF90 experiment, forecasts out to 90-day leads were made for 128 consecutive days starting on 3 May 1990. Progressively shorter lead forecasts were appended to complete the block of forecasts depicted in Fig. 1. Only the "Lorenz block" of forecasts verifying between 1 August and 6 December 1990 are used here.

The model used for DERF90 was a reduced resolution (T40) version of the MRF that was operational during the summer of 1990 (White and Caplan 1990). The boundary conditions for the DERF90 runs were

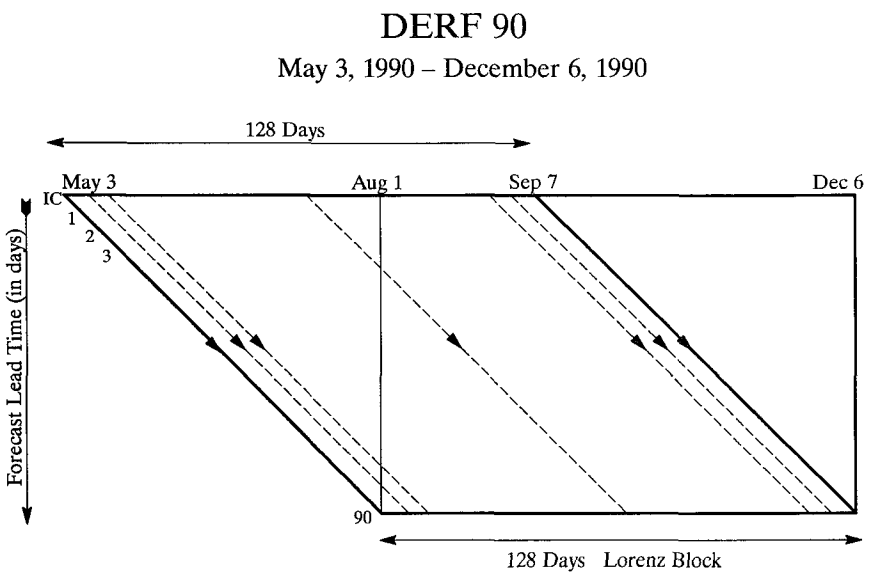

FIG. 1. Schematic of the DERF90 forecast experiment. Forecasts in the rectangular Lorenz block that verify on the 128 days from 1 August through 6 December are used.

designed to be as realistic as possible. Snow depth and soil moisture were interactive but were attracted to an evolving climatology with a 90-day e-folding time. Sea surface temperature (SST) was initially the observed field but was also damped toward climatology with a 90-day e-folding. The sea ice was represented by observed initial anomalies for 30 days and switched discontinuously to the climatological distribution for leads greater than 30 days. Some of the basic results of DERF90 can be found in Van den Dool (1993).

b. *Observed data and climatology*

The observed analyses used to verify the DERF90 forecasts were obtained from NMC's global data assimilation system (GDAS). The instantaneous daily 0000 UTC observations from the years 1987–1991 are used here.

An observed climatological mean field is required for calculation of the AC defined in the next section. The 10-yr Climate Diagnostics Data Base (CDDB) consisting of 10-yr mean monthly fields from September 1978 through August 1988 is used here. These climatologies are based on twice-daily NMC GDAS data archived in real time. The monthly means are linearly interpolated to produce monthly climate means valid on each individual date.

3. Verifying individual forecasts

The method for evaluating skill described in the next section uses a scalar measure of the similarity between a forecast field and an analysis field. Throughout this study, the anomaly correlation

$$\mathrm{AC} = \frac{\sum_i (F_i - C_i)(V_i - C_i)}{[\sum_i (F_i - C_i)^2 \sum_i (V_i - C_i)^2]^{1/2}}$$

is used to measure this similarity. Here C_i, F_i, and V_i are the climate, forecast, and verification height values at grid point i, and the area-weighted sum extends over all grid points in a given geographic region. This definition of AC was used by Miyakoda et al. (1972) and Saha and Van den Dool (1988) (who also summed in time) but is slightly different from the AC used operationally at NMC.

The AC has been applied to three different geographical regions. For all regions, the data points are on the 33° latitude × 64° longitude Gaussian grid corresponding to the T21 spectral representation of the data. Two regions are the Northern Hemisphere (NH) and Southern Hemisphere (SH), all points north of 20°N or south of 20°S, respectively. The North American (NA) region contains all points between latitudes 30° and 80°N and between longitudes 70° and 140°W.

The root-mean-square error

$$\mathrm{rms} = \left[\frac{\sum_i (F_i - V_i)^2}{n}\right]^{1/2},$$

where n is the number of grid points, has also been applied as a complement to AC results (see discussion in section 8).

4. A method for assessing forecast skill

Figure 2 shows an extended lead DERF90 forecast verifying on 6 December and an analysis for 6 Decem-

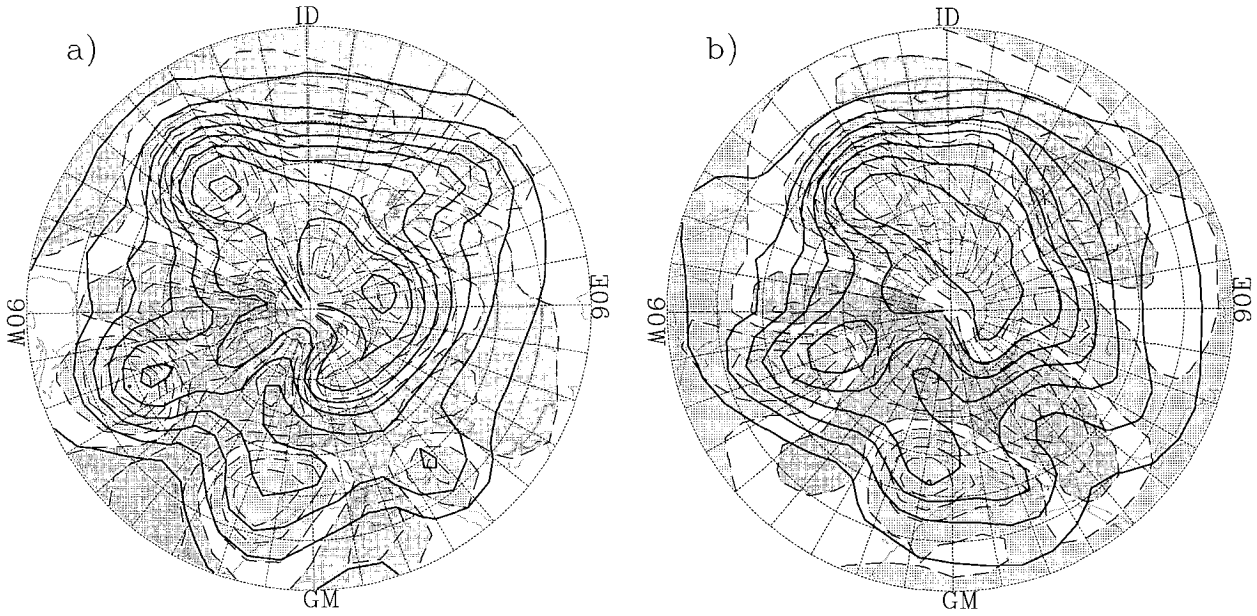

FIG. 2. The 500-mb height fields for (a) 6 December 1987 and (b) the 79-day lead DERF90 forecast valid on 6 December 1990. Positive (negative) departures from the CDDB climatology are shaded dark (light) starting at ±30-m anomalies. The contour interval is 60 m starting at ±30 m.

ber. The forecast and verification have a Northern Hemisphere AC of 0.62, above the threshold of 0.60 that is frequently used to demarcate skillful forecasts (Hollingsworth et al. 1980). Unfortunately, the verification field is for 6 December 1987, while the DERF forecast is, of course, for 1990. This clearly demonstrates the hazards involved in evaluating single forecasts from a large set; some high AC forecasts will occur by chance even in sets of forecasts with no reasonable possibility of skill.

Figure 3 shows a DERF90 forecast for 14 August 1990 and the observations for 14 August 1989. The AC between the forecast and the 1989 analysis over the North American region is 0.87 in this case. For fore-

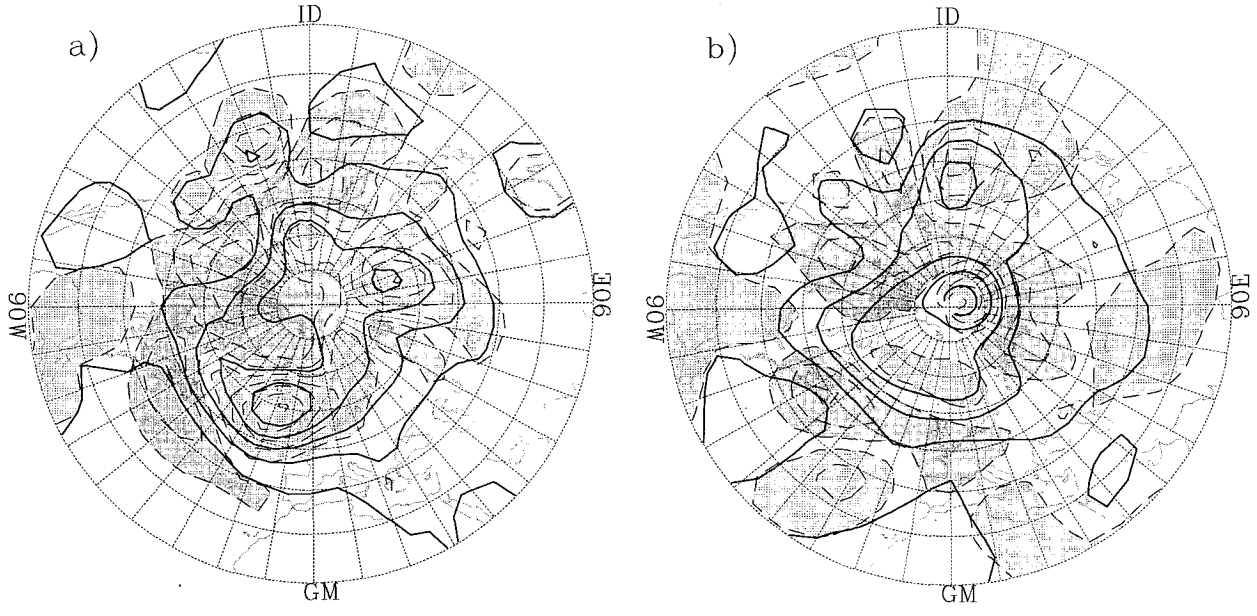

FIG. 3. As in Fig. 2 but for (a) the 14 August 1989 analysis and (b) the 5-day lead DERF90 forecast valid 14 August 1990.

casts verified over limited regions like North America (where the number of spatial as well as dynamical degrees of freedom resolved by the data is relatively small), the distribution of AC values tends to be broader, making chance occurrence of high (or very negative) AC forecasts even more likely. The Southern Hemisphere distribution of AC also differs from that for the Northern Hemisphere; several SH ACs of over 0.70 (and less than −0.70) occur for DERF90 forecasts verified against observations from the wrong year. The largest NH ACs have absolute magnitudes of slightly more than 0.6.

The preceding examples illustrate the hazards of evaluating skill without some control with which to establish statistical significance. While such a control is useful in the evaluation of individual forecasts as above, it can provide more information when applied to the distributions of ACs for a large set of forecasts.

The method applied here can be summarized as follows: suppose g is some measure (assumed to be scalar here although this is not essential) of the similarity between a forecast and an analysis; G is defined as the distribution of values of g for some set of forecasts f compared to the appropriate verifying analyses v; G^* is a control distribution of values of g for the same set of forecasts f but verified against some set of unrelated observations v_{false}. Forecasts can be considered skillful only if the distribution of elements in G is different from (and presumably better than) the distribution of G^*; G^* will be the most rigorous test if the unrelated observations v_{false} are for the same variable, same level, same region, and same time of year. Hence, v_{false} is chosen to be identical to v except that the year of the analyses is different from that of the forecasts.

In the next section, the DERF90 forecasts will be verified against a set of no-skill forecasts that verify the DERF90 forecasts against analyses from the years 1987–89 and 1991. Table 1 provides a description of this "false verification" forecast control set. Figures 2 and 3, discussed above, are individual verifications from the 1987 and 1989 false verification control sets. With the exception of interannual variability in the observed climate, the "degrees of freedom" in the false verification set for each other year is identical to that for the DERF90 verification set.

The DERF90 forecasts can be said to have no skill if they have a distribution of ACs that is statistically indistinguishable from the distribution for the control

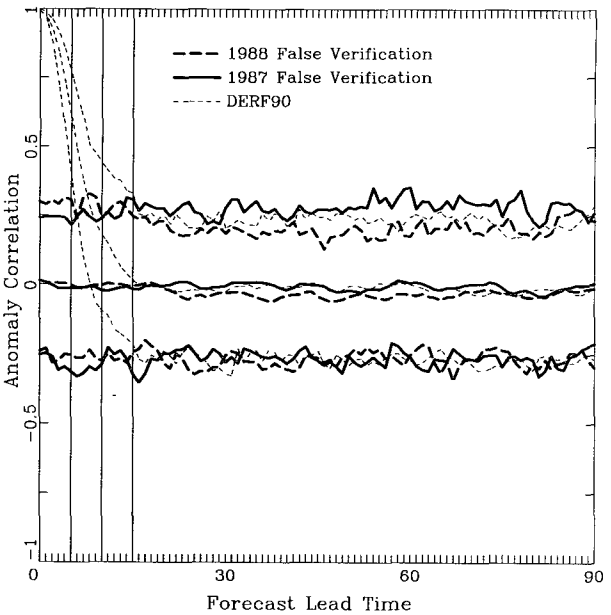

FIG. 4. The 95% threshold (upper curves), mean (middle), and 5% threshold (lower) for NH ACs as a function of lead time for DERF90 (thin dash) and the 1987 (solid) and 1988 (thick dash) false verification control sets.

forecast set. If the DERF90 AC distribution does differ significantly from the distribution for the false verification controls, the DERF90 forecast may be skillful. However, there are a number of other possible causes for a difference in the AC distribution. The climate for some false verification years may have a different number of dynamical degrees of freedom leading to a difference in the AC distribution width. The mean AC of the distribution may also change if, by chance, a given verification year has a time-mean anomaly from the long-term climate mean that is similar to the DERF90 model bias. Given a sufficiently large number of false verification control years, even these effects can be ruled out with statistical confidence.

5. Skill of DERF90 forecasts

This section compares the AC of the DERF90 forecasts to those for the false verification controls. Figure 4 shows the NH ACs for DERF90 and the 1987 and 1988 false verification sets as a function of forecast lead

TABLE 1. The forecast and verification fields for DERF90 and the no-skill control forecast set. The lower right box indicates a field that is different from the DERF90 case.

Forecast set name	Forecast fields	Verification fields
DERF90	DERF90 0–90-day lead forecasts.	DERF90 period observations (1990)
False verification control	DERF90 0–90-day lead forecasts.	Observations from DERF90 period but from different year (1987–89, 1991)

time; for each lead time there are 128 AC values. The upper and lower groups of curves are the 95% and 5% points of the AC distribution, while the middle group of curves are the mean values. The DERF90 ACs are clearly much better for short lead times when the MRF forecasts are known to have skill. The individual curves in each of the three groups become essentially indistinguishable (to our eyes at least) at leads of approximately 14 days. A similar result also holds for curves of the maximum and minimum AC values as a function of lead time (not shown). The maxima and minima curves have average absolute values slightly greater than 0.5 for leads past 15 days. This is consistent with the findings of Lorenz (1969) and Ruosteenoja (1988) that good analogs for large regions like the NH are extremely rare.

For extended lead times, past the loss of DERF90 initial skill, the mean ACs are approximately 0. The mean AC of the DERF90 forecasts is generally bracketed by the 1987 and 1988 controls. It is interesting to note that the 95% AC curve for 1987 is almost always higher than those for DERF90 or the 1988 control. This skew toward larger ACs in the 1987 false verification control suggests that the climate anomaly of that year was somehow more similar to the DERF90 long-lead-time model bias than was the case for other observed years including 1990 itself. In contrast, the 5% curves are more similar.

The differences in behavior demonstrated by the 5%, 95%, and mean curves in Fig. 4 suggest that more information on DERF90 might be obtained by examining

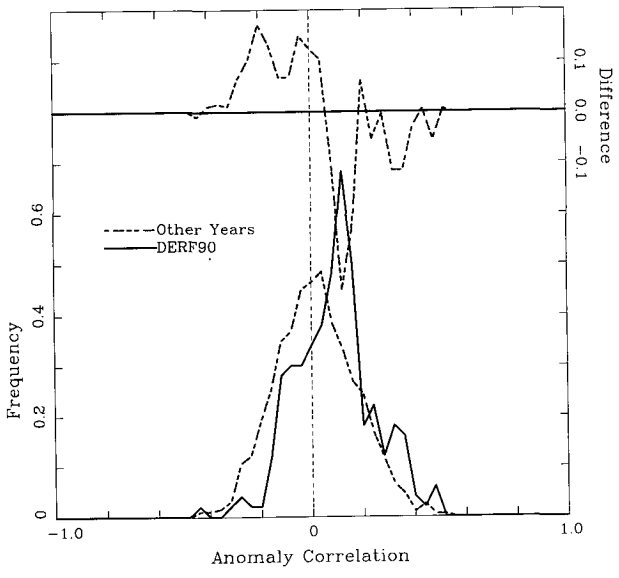

FIG. 5. Northern Hemisphere AC distribution for 12- and 13-day-lead DERF90 forecasts (solid), and the false verification control (dashed). The lower curves show the AC distributions, while the upper curve plots the value of the false verification distribution minus the DERF90 value.

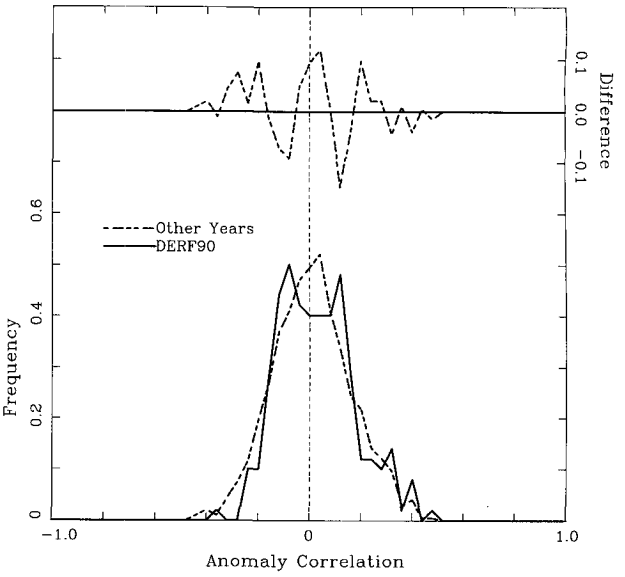

FIG. 6. Distribution of NH ACs for 14- and 15-day-lead DERF90 forecasts (solid) and the false verification controls (dashed). The upper curves plot the difference as in Fig. 5.

the entire distribution of DERF90 ACs as a function of lead time. Figure 5 displays the complete distribution of NH ACs for DERF90 forecasts with leads of 12 and 13 days and the distribution for the corresponding false verification sets for all four alternate years. Both Figs. 5 and 6 use a total of 51 bins to plot the results. Two lead times are included in the AC distribution plots because plots for a single lead are difficult to evaluate by eye due to sampling noise. In Fig. 5, the DERF90 forecasts clearly have many more positive AC forecasts than the controls, and it seems that the forecasts are still skillful at this lead time. Section 6 will present a statistical method to look more closely at this comparison.

Figure 6 shows a comparison between the 14- and 15-day-lead NH AC distribution for DERF90 and the corresponding distribution for the false verification control. As shown by the difference plot at the top of the figure, the DERF90 forecasts are no longer clearly skillful at leads beyond approximately 12 days since they do not have more high AC forecasts than the control. However, it is important to note that there are many forecasts in both Figs. 5 and 6 with AC greater than 0. Even the mean of the distributions is positive; this makes a control essential to accurately assess skill.

Even with a fairly large number of samples, it is difficult to compare distributions like those in Fig. 6 by eye. The next section presents a robust statistical method that can compare such distributions.

6. Kolmogorov–Smirnov tests

Because of sampling noise in the AC distributions, two lead times had to be combined to produce the rea-

sonably smooth AC distribution plots displayed in the previous section. Statistical tests are available that can compare samples from two distributions, despite large amounts of noise or small sample sizes. In this section, the Kolmogorov–Smirnov (KS) test (Knuth 1981; Conover 1980) is applied to compare DERF90 forecasts to the false verification control forecast set for each lead time. The KS test produces a probability that two samples of arbitrary sizes are taken from identical distributions. The cumulative distribution functions for the two samples $D(x)$ and $F(x)$ are compared over the range of the functions. The largest value of $G = |F(x) - D(x)|$ is used to evaluate the similarity of the two samples. The value of G can be compared to a known distribution to evaluate the probability that the two samples came from the same distribution. Here the KS test was computed using the International Mathematical and Statistical Libraries, Inc. statistical routine KSTWO.

Figure 7a shows the results of the KS test as a function of lead time for the DERF90 forecasts compared to the false verification control "forecasts." Kolmogorov–Smirnov values are shown for the NH, SH, and NA AC distributions. For lead times less than 10 days, the KS test gives values of very nearly 0, indicating that the DERF90 forecasts have a very different AC distribution from that for the control. Results from the previous section suggested that these differences are generally a result of much more frequent positive ACs in the DERF90 short lead forecasts.

Figure 7a shows that the KS test first becomes significantly nonzero (greater than 0.1, the 90% confidence threshold for distinct distributions) at leads of 13, 14, and 16 days for the NA, NH, and SH regions, respectively. At these leads, the DERF90 forecasts have apparently become unskillful since their AC distribution is statistically indistinguishable from the distribution for control forecasts with no skill. At shorter lead times, the DERF90 forecast AC distribution differs from that for the controls. Examination of the complete individual lead time distributions suggests that DERF90 is producing skillful forecasts in all three regions up to the time at which the KS tests become nonzero. The SH retains a small amount of skill for several days longer than the NH, an unexpected result given the more rapid decay of SH skill for short lead times. What part of the SH forecasts is actually maintaining the small amount of skill is currently unknown.

A closer examination of the KS values in Fig. 7 shows that the values fluctuate quite a bit for leads past 15 days. Figure 7b shows 5-day mean values of the KS test for leads out to 90 days. There are certain periods past 20 days lead in Fig. 7a, and even in the average values of Fig. 7b, when the KS values are quite small. This leaves open the possibility that some sort of return of skill might be occurring in the DERF90 forecasts, a possibility that is examined in the next section.

7. Return of skill

Although the results of section 6 indicate that overall there is no significant skill at lead times in the midteens, this does not entirely rule out the possibility of return of skill. For instance, the individual forecast NA AC

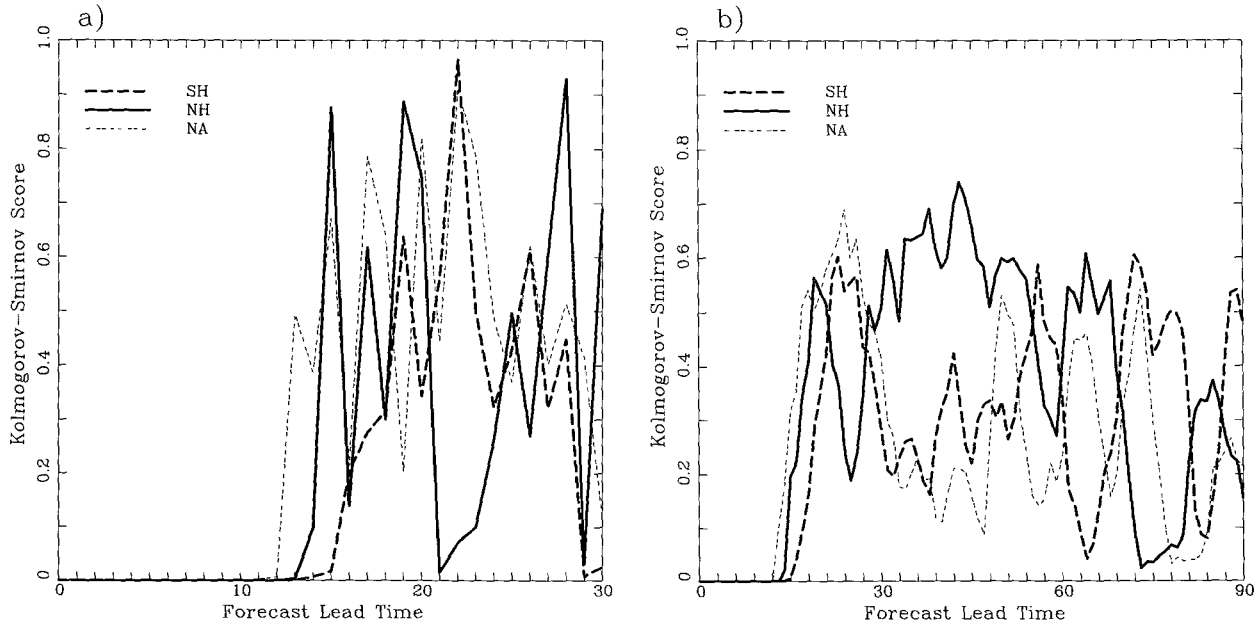

FIG. 7. Values of the Kolmogorov–Smirnov test comparing the DERF90 AC distribution and the false verification AC distribution as a function of lead time for the NH, SH, and NA domains. Daily values are shown in (a), and 5-day means in (b).

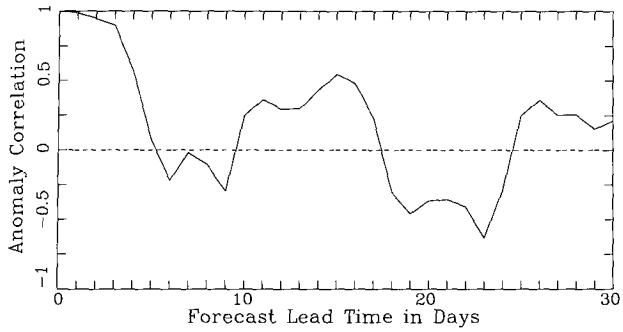

FIG. 8. Anomaly correlation as a function of lead time for the individual 165th DERF90 forecast initialized on 14 October 1990.

series shown in Fig. 8 has two apparent returns of skill, the first in the range where slightly skillful forecasts are still being produced on average. This is not an isolated example; similar return of skill can be found for other initial conditions and for the NH and SH regional ACs. This section will examine the return of skill phenomenon with the goal of determining if there is any *statistically significant* return of skill in the DERF90 forecasts. If so, further questions about the utility of such forecasts and the physics involved would be in order.

a. Methodology

As in the previous section, the return of skill in the DERF90 forecasts will be compared to return of high AC in a set of no-skill control forecasts. In the DERF90 forecasts, an initial AC minimum is required as a precursor to potential return of skill. For each DERF90 forecast, the initial skill minimum is defined as the first day for which the AC is a local minimum as a function of lead time and is also below a threshold criterion (AC = 0.25 in the cases presented here although results for other thresholds show no qualitative difference). The days following this initial AC minimum are candidates for a return of skill. For instance, in Fig. 8, the first AC minimum occurs at day 6, and all days after that are potential return of skill days.

The false verification forecast set is used as a standard to which the DERF90 return of skill can be compared. For each false verification forecast (the 0–90-day lead forecasts from an individual DERF90 initial condition verified against another year's analyses), a local AC minimum as a function of lead time is located at leads of 10 days or greater. Ten days is chosen so that the mean lead time of the false verification control and DERF90 initial AC minima are approximately the same. A local false verification AC minimum will be accepted only if it is bounded by the largest and smallest values of the AC found for initial skill minimum days in the DERF90 forecasts. Again, the purpose is to make the mean value of the false verification AC minima as close as possible to the mean for the DERF90 case. This is an attempt to produce as fair a comparison as possible between forecasts following the false verification AC minima and forecasts following the DERF90 AC minima. If there is significant return of skill in the DERF90 forecasts, the distribution of ACs following the AC minimum should be different for DERF90 than for the false verification controls.

b. Results

Figure 9 shows the distribution of ACs for the first 20 days after the initial NH AC minimum for DERF90 and for the false verification controls using a total of 201 bins for plotting. The two distributions are nearly identical, with a KS value of 0.62. In fact, if only 51 bins had been used in Fig. 9, the curves would be almost indistinguishable. There is no evidence at all of a statistically significant return of skill in the DERF90 forecasts.

Results for values of the initial AC minimum threshold other than 0.25 are similar. If the value is made much larger than 0.25, some small dips in AC occurring in the first few days of DERF90 forecasts are picked up as candidates for return of skill. In a strict sense, these are true return of skill cases but they occur only at very short lead times. The return of skill of interest here would occur at times later than 2 or 3 days lead time, when ACs have generally dropped below 0.25.

Results for return of skill in the SH and NA regions (not shown) are similar. Again, there is no significant difference between the AC distribution for DERF90 and that for the false verification controls immediately following an AC minimum.

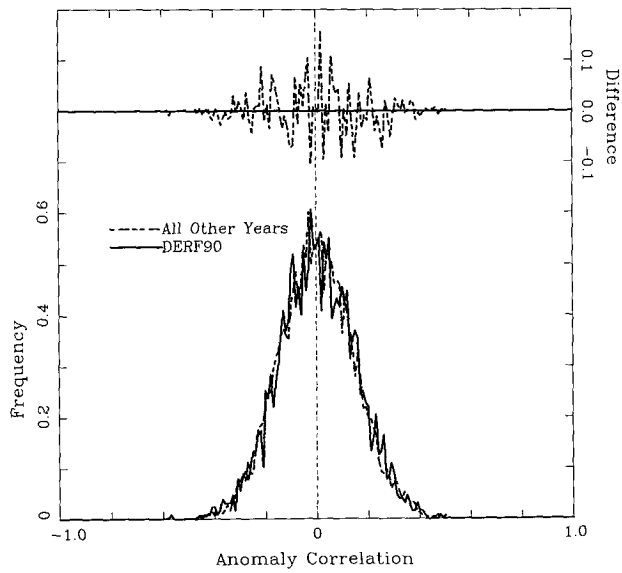

FIG. 9. Distribution of NH ACs for DERF90 forecasts (solid) and the false verification controls (dashed). The upper curves plot the difference as in Fig. 5. Anomaly correlations are for first 20 days following an initial AC minimum of less than 0.25.

8. Conclusions

NMC's DERF90 forecasts have been examined to establish the longest lead times for which the existence of forecast skill can be claimed. At lead times past 10 days, these forecasts can be expected to have sporadic and/or extremely limited amounts of skill. Examples have been presented to demonstrate the hazards of equating positive ACs of individual extended-range forecasts with forecast skill since high ACs can also be generated by chance. To evaluate whether a set of dynamical forecasts has statistically significant skill, the distribution of AC for the entire set should be examined and compared to some reasonable control distributions.

A set of control forecasts, known a priori to have no skill, has been developed to help evaluate the DERF90 set of extended-range dynamical forecasts. The control set retains the same forecasts as the DERF90 experiment but substitutes verification fields from years other than that for which the DERF90 forecasts were made.

The results presented compare the distribution of AC as a function of lead time for the DERF90 forecasts and the control set. Although not presented here, all the tests were also repeated using the rms difference (section 3) to verify individual forecasts. The rms is a good complement to the AC because, as opposed to the AC, it does not depend on the choice of a climatological field. All results presented were reproduced qualitatively using rms instead of AC.

Comparisons between DERF90 and the control set were performed for three different regions. Differences in the dynamical degrees of freedom resolved in each region can lead to very different distributions of AC. The dependence of AC distributions on verification region makes a priori bounds on the AC for skillful forecasts almost impossible to define. Hence, it is always necessary to compare DERF90 AC distributions to the control distribution for the corresponding region.

The distributions of AC for DERF90 were compared to those for the control as a function of lead time. Visual inspection of the distributions was used to estimate the lead times at which the DERF90 forecasts were no longer clearly superior to the controls. The Kolmogorov–Smirnov test, which evaluates the probability that two samples are taken from identical distributions, was also used to examine the leads at which the DERF90 AC distributions become statistically indistinguishable from the controls. This occurs at leads of 13, 14, and 16 days for the NA, NH, and SH regions, respectively.

Although significant skill for the complete DERF90 set is lost at these leads, it is still possible that individual forecasts have skill at larger leads. It is not hard to find individual forecasts in the DERF90 set with apparent return of skill after an initial AC minimum. Comparison to the no-skill controls shows that just as many individual forecasts with return of skill occur in the controls. Hence, there is no evidence of any statistically significant return of skill in the DERF90 forecasts.

There appears to be little skill in the DERF90 forecasts at leads greater than 14 days. If it is possible to identify a priori those forecasts that maintain (or return) skill at longer leads, forecasts could be issued only in those cases. If this were possible, it would also be possible to identify all other forecasts as less skillful on average than a control forecast. However, our results suggest that forecasts with skill at extended ranges do not exist.

Another potential way to improve the lead times at which DERF forecasts retain skill is the reduction of model bias or drift. As shown by Anderson (1993), the MRF used in DERF90 has a significant drift that may be dominating the forecast errors for extended leads. Johannson and Saha (1989) in a simple model, and Saha (1992) in the MRF, have clearly demonstrated the potential skill increases that can result from correcting model bias. How much effect such techniques would have on extended-range forecast skill is an interesting and as yet unanswered question.

Great care is always needed when searching for skill in forecasts that are expected to have little or none. It is vital to establish the statistical significance of skill results. This paper has presented a method that can better evaluate the amount of skill in extended-range forecasts than many traditional methods. It is essential to gauge accurately the small amounts of useful skill available in extended-range dynamical forecasts in order to evaluate further model improvements. With this in mind, more robust and sophisticated methods for evaluating skill of slightly skillful forecasts need to be developed. Studies evaluating skill for forecasts with very limited amounts of skill should include comparisons to adequate controls. Some procedure similar to that described here could be applied in all studies, especially those where intricate manipulations (averages, filters, etc.) are performed on forecasts and analyses before skill is evaluated. Preliminary application of the procedure to time- or ensemble-averaged forecasts has produced some interesting results that will be discussed in a note to be published at a later date. It is hoped that the use of good controls will lead to a more accurate assessment of the current state of extended-range forecast skill and subsequent improvements in this skill.

Acknowledgments. The authors are indebted to S. Saha for helping with data acquisition, to A. Barnston for his statistical insight, and to the many others at CAC and NMC's Development Division that contributed to the DERF90 project.

REFERENCES

Anderson, J. L., 1993: The climatology of blocking in a numerical forecast model. *J. Climate,* **6,** 1041–1056.

Conover, W. J., 1980: *Practical Nonparametric Statistics.* John Wiley and Sons, 493 pp.

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

Hollingsworth, A., K. Arpe, M. Tiedtke, M. Capaldo, and H. Savijärvi, 1980: The performance of a medium-range forecast model

in winter: Impact of physical parameterization. *Mon. Wea. Rev.*, **108,** 1736–1773.

Johannson, Å., and S. Saha, 1989: Simulation of systematic error effects and their reduction in a simple model of the atmosphere. *Mon. Wea. Rev.*, **117,** 1658–1675.

Kalnay, E., M. Kanamitsu, and W. E. Baker, 1990: The NMC global forecast system. *Bull. Amer. Meteor. Soc.*, **71,** 1410–1428.

Kanamitsu, M., K. C. Mo, and E. Kalnay, 1990: Annual cycle integration of the NMC MRF model. *Mon. Wea. Rev.*, **118,** 2543–2567.

Knuth, D. E., 1981: *The Art of Computer Programming. Vol. II: Seminumerical Algorithms.* Addison Wesley, 688 pp.

Lorenz, E. N., 1969: Atmospheric predictability as revealed by naturally occurring analogues. *J. Atmos. Sci.*, **26,** 636–646.

Miyakoda, K., G. D. Hembree, R. F. Strickler, and I. Shulman, 1972: Cumulative results of extended forecast experiments. Part I: Model performance for winter cases. *Mon. Wea. Rev.*, **100,** 836–855.

——, T. Gordon, R. Caverly, W. Stern, J. Sirutis, and W. Bourke, 1983: Simulation of a blocking event in January 1977. *Mon. Wea. Rev.*, **111,** 846–869.

Murphy, A. H., and R. L. Winkler, 1987: A general framework for forecast verification. *Mon. Wea. Rev.*, **115,** 1330–1338.

Palmer, T. N., 1993: Extended-range atmospheric prediction and the Lorenz model. *Bull. Amer. Meteor. Soc.*, **74,** 49–66.

Ruosteenoja, K., 1988: Factors affecting the occurrence and lifetime of 500 mb height analogues: A study based on a large amount of data. *Mon. Wea. Rev.*, **116,** 368–376.

Saha, S., 1992: Response of the NMC MRF model to systematic-error correction within integration. *Mon. Wea. Rev.*, **120,** 345–360.

——, and H. M. Van den Dool, 1988: A measure of the practical limit of predictability. *Mon. Wea. Rev.*, **116,** 2522–2526.

Tracton, M. S., K. Mo, W. Chen, E. Kalnay, R. Kistler, and G. White, 1989: Dynamical extended range forecasting (DERF) at the National Meteorological Center. *Mon. Wea. Rev.*, **117,** 1604–1635.

Van den Dool, H. M., 1989: A new look at weather forecasting through analogues. *Mon. Wea. Rev.*, **117,** 2230–2247.

——, 1993: Long range weather forecasting through numerical and empirical methods. *Ocean Atmos. Dyn.*, **20,** in press.

White, G. H., and P. M. Caplan, 1991: Systematic Performance of the NMC medium-range model. *Ninth Conf. on Numerical Weather Prediction,* Boulder, CO, Amer. Meteor. Soc., 806–809.