Three Years of Operational Prediction of Forecast Skill at NMC

RICHARD L. WOBUS* AND EUGENIA KALNAY

Development Division, National Meteorological Center, Washington, D.C.

(Manuscript received 28 January 1994, in final form 3 January 1995)

ABSTRACT

In real time since 1990, the National Meteorological Center (NMC) has been running a system to predict the forecast skill of the medium-range forecasts produced by the NMC global spectral model. The predictors used are the agreement of an ensemble consisting of operational forecasts from various centers, the persistence in the forecast, and the amplitude of the anomalies. These predictors are used in a stepwise regression scheme, with the last 60 days used as training period, and the regional anomaly correlation of the 0000 UTC NMC global forecast is predicted from days 1 to 6. By far the most important predictor of skill is the agreement between the NMC global forecast started at 0000 UTC, out to 6 days, and four other 12-h “older” forecasts (Japan Meteorological Agency, United Kingdom Meteorological Office, and the European Centre for Medium-Range Weather Forecasts, as well as the average of the NMC forecast at 0000 UTC with the previous day’s forecast). The other predictors have been selected to add to the predictive capability of the agreement alone, and together they quantify the factors that forecasters use subjectively when evaluating the available forecasts. These predictions are available to NMC forecasters on workstations and to outside users through the Internet.

The predictive ability of this system compares favorably with recent theoretical and experimental studies. The correlation between predicted and verifying forecast skill seems to be best in regions where forecast skill varies significantly. The seasonal variation in predicting the skill is small except in the Tropics. The overall performance shows that these predictors include enough information about forecast skill to justify further development of skill predictions based on larger forecast ensembles and on more sophisticated statistical techniques.

1. Introduction

Since 1979, when the Global Weather Experiment took place, the skill of the operational forecasts has increased substantially in the short, medium, and even longer ranges, both in the Northern and the Southern Hemispheres and in the Tropics (e.g., Kalnay et al. 1990; Bengtsson 1991). Nevertheless, forecast skill remains highly variable from day to day, region to region, and season to season. It is clear that the utility of the numerical forecasts would be considerably enhanced if, for example, a human forecaster could know that today’s forecast will remain skillful for more than 6 days, or, conversely, that today’s 3-day forecast will be much less reliable than normal. This fact prompted the often quoted statement of Tennekes et al. (1986) that “no forecast is complete without a forecast of the skill.”

Epstein (1969) was the first to attempt to develop a method to predict the skill of dynamical forecasts. He introduced a stochastic-dynamic method for predicting the probability distribution of the model variables. This scheme, which requires prognostic equations for the covariances of the model variables, is too computationally expensive to be applied to realistic models. Epstein also suggested the use of ensemble forecasting for this purpose. Leith (1974) showed that in simulated “Monte Carlo forecasting” (MCF), when the perturbations were generated by random errors representative of the analysis uncertainty, even a small number of forecasts could improve the skill of the forecast mean and provide an estimate of its error. Hoffman and Kalnay (1983) suggested the use of the “lagged average forecasting” (LAF) method, and showed that, at least for a simple model, LAF resulted in better predictions of the skill than the Monte Carlo method.

In principle, the theory of predictability provides the basis for prediction of the forecast skill. Lorenz (1963) showed that a deterministic forced, dissipative dynamical system (such as the atmosphere) can be infinitely predictable if it is always stable but that if all states are unstable it has a finite limit of predictability. From this we can infer that the more unstable the atmosphere is, the less skillful our numerical forecasts of its future state are likely to be. In recent years there have been basically three approaches to the prediction of the forecast skill, all of which attempt to relate forecast skill to some measure of atmospheric stability.

* Other affiliation: General Sciences Corporation, Laurel, MD.

---

* Corresponding author address: Dr. Richard L. Wobus, Development Division, WNMClC2:RW, National Meteorological Center, Washington, DC 20233.
(a) Ensemble forecasting exploits the relationship between forecast agreement and forecast skill (e.g., Kalnay and Dalcher 1987; Palmer and Tibaldi 1988; Murphy 1988, 1990; McCalla and Kalnay 1988; Kalnay and Ham 1989; Tracton et al. 1989; Baumhefner 1991, Ebisuzaki and Kalnay 1992; Wobus and Kalnay 1991; Molteni and Palmer 1993). In this approach it is assumed that if the model is sufficiently realistic, the same instabilities that increase the distance between a forecast and the real atmosphere will also increase the "spread" among the members of the forecast ensemble. Two methods to generate perturbations for ensemble forecasting were recently implemented at the National Meteorological Center (NMC) (Tracton and Kalnay 1993; Toth and Kalnay 1993) and at the European Centre for Medium-Range Weather Forecasts (ECMWF) (Palmer et al. 1992; Molteni and Palmer 1993; Buizza 1994) but are not yet routinely used for prediction of the skill.

(b) Dependence of forecast skill on atmospheric regime is related to the variable stability properties of the large-scale flow (e.g. Palmer 1988; Palmer and Tibaldi 1988; Tracton et al. 1989; Tibaldi and Molteni 1990; Molteni and Palmer 1991). Similar to this approach are studies relating forecast skill to atmospheric persistence (e.g., Branstator 1986; Chen 1989) or to the Lorenz index (Kimoto et al. 1992).

(c) Regional prediction of maximum error growth is based on the use of the adjoint of the forecast model (e.g., Barkmeijer 1993; Houtekamer 1993).

Barker (1991) performed a large number of Monte Carlo ensemble predictions using a simple but realistic primitive equations model (Roads 1987) and perturbations chosen at random from a long sample of the model output. His results were somewhat discouraging: he used a large number of ensemble members, of the order of 100, and a "perfect model" simulation (i.e., one in which the same model was used to simulate the atmospheric evolution and the model forecasts). However, even under these favorable conditions, the correlation between the hemispheric forecast rms error and the "rms spread" among the Monte Carlo ensemble members varied between only 0.35 and 0.55 for the first 10 days of the forecast and was even smaller later. This suggests that there is an upper limit for the predictability of the skill obtainable from ensemble forecasting. This is because, given a perfect model and a perfect knowledge of the statistical uncertainty in the initial conditions, the spread of a large ensemble of trajectories can perfectly predict how fast another large ensemble of trajectories (one of which could be the atmospheric trajectory) will drift apart. However, the ensemble spread cannot perfectly predict how a single forecast will compare with a single verification (H. van den Doel 1992, personal communication); likewise, the stochastic–dynamic method of Epstein (1969) can only predict well the skill of a collection of many forecasts. Moreover, Barker pointed out that when using randomly chosen Monte Carlo perturbations, even if they represent perfectly well the statistics of the simulated random analysis errors, the prediction of the skill has zero correlation between predicted and verifying rms errors at the initial time. It is only after the growing perturbations organize themselves and dominate the error growth that the correlation between forecast agreement and forecast skill starts to increase. For this reason, his correlation between predicted and verifying skill was 0.0 at the initial time and only 0.35 at day 1.

Palmer and Tibaldi (1988) and Molteni and Palmer (1991) developed an experimental system to forecast the skill of the operational ECMWF forecasts. They use several predictors based on the ECMWF operational forecast to predict the rms error directly and use statistical inference based on anomaly amplitude to predict the anomaly correlation. More recently, ensemble forecasting has been implemented at NMC (Toth and Kalnay 1993; Tracton and Kalnay 1993) and at ECMWF (Molteni and Palmer 1993; Mureau et al. 1993; Buizza 1994), with its main purpose to provide the basis for probabilistic forecasting. The two systems differ in the way the initial perturbations are created.

The purpose of this paper is to review the results of a different method, also developed at NMC, for operational prediction of the regional forecast skill. This method, which has been extensively tested, is based on the use of forecasts from multiple centers and other predictors of skill (Wobus and Kalnay 1991). The NMC system for skill prediction has been running in real time since 1990 and has actually resulted in better correlations between predicted and verifying skill than those obtained by Barker (1991). Its output is available in real time to any interested user through anonymous file transfer protocol (ftp) in the Internet (see the appendix). Section 2 contains a description of the method; section 3 presents daily predictions of the skill for a recent season (spring of 1993); section 4 contains comprehensive statistical verifications for all seasons and all regions of the world; and section 5 is a summary and discussion of future plans.1

2. The NMC system for prediction of the skill

As indicated in the previous section, predictions of the skill based on ensembles have been generally based on either MCF or LAF. For example, Dalcher et al. (1988) tested the LAF method with a 100-day dataset

1To keep the discussion as clear as possible, we use the term "forecast" to refer to a forecast of a future state of the atmosphere, as produced by a numerical model, and the term "prediction" to refer to a prediction of the skill of a forecast as measured by the anomaly correlation statistic.
from ECMWF and showed that the method showed promise in predicting the skill. However, the decay of the forecast skill itself with time makes the use of LAF with once-a-day forecasts undesirable because forecasts "older" by two or more days have initial errors much larger than those of the "younger" forecasts\(^2\) (Tracton et al. 1989).

To avoid this problem, as well as the need to perform additional LAF or MCF forecasts, we have developed a method in which the main predictor of skill is the average of the agreement between the NMC global forecast and other centers' operational forecasts (McCalla and Kalnay 1988; Kalnay and Ham 1989; Wobus and Kalnay 1991). We take advantage of the fact that the NMC medium-range forecasts (MRF) are started daily at 0000 UTC, whereas forecasts from three other global forecasting centers [United Kingdom Meteorological Office (UKMO), Japan Meteorological Agency (JMO), and ECMWF] are started at 1200 UTC. As a result, by the time the NMC MRF forecasts from 0000 UTC become available to the users, forecasts from UKMO, JMA, and ECMWF started 12 h earlier are also available. Although several other nations are producing or plan to produce global forecasts, none are currently available at NMC in time to be used for this purpose.

---

\(^2\)This problem is eliminated in "scaled LAF" (Ebisuzaki and Kalnay 1992). Here the LAF perturbations (differences between forecasts from previous analysis and the latest analysis) are divided by their "age" in, for example, units of 6 h (the 12-h LAF perturbation is divided by 2, the 24-h perturbation divided by 4, etc.). The SLAF perturbations all have similar sizes and can be added and subtracted from the control analysis, thus doubling the number of perturbations for ensemble forecasting.

The NMC operational prediction of forecast skill is based on a stepwise regression scheme using a 60-day training period. Because it is desirable to provide the forecaster with regional rather than hemispheric or global skill guidance, we have chosen to predict the regional 500-hPa anomaly correlation of the 0000 UTC NMC global forecast for regions of 30\(^\circ\) latitude by 60\(^\circ\) longitude covering the whole globe (Fig. 1). The most important predictor of skill is the agreement between the NMC medium-range global forecast started at 0000 UTC and four other 12-h older forecasts. These are the 1200 UTC forecasts from JMA, UKMO, and ECMWF, as well as an average of the current NMC forecast from 0000 UTC with that from the previous day, which simulates a forecast out of date by 12 h. The comparison forecasts are all less skillful than the current NMC forecast since they were started from analyses 12 h earlier. Since the members of the ensemble of forecasts have been chosen essentially randomly (they are all the 12-h older forecasts that happen to be available at NMC) and are obtained at no additional cost, this predictor can be considered to be based on a "poor person's" Monte Carlo ensemble. On the other hand, it could be also considered as nearly the most sophisticated and advanced ensemble that could be possibly created, since the ensemble members are made with state-of-the-art forecast models developed and improved over many years by many scientists, and the differences in their initial conditions truly reflect the uncertainties in our knowledge of the real atmosphere. Although all the systems use the same observations, without introducing additional random observational errors, this does not reduce the realism of the ensemble. Random errors in the analysis introduced by the use of observations at the analysis time have a negligible effect on the error growth when compared to the effect of the fast-growing
dynamical structures present in the operational analysis errors through the use of the analysis cycle (Toth and Kalnay 1993). It is the random differences in the fast-
growing modes among the initial states of the forecast models that contribute to our skill prediction.

We produce predictions of the regional skill for the MRF forecasts verifying at 1200 UTC. Since the MRF
is started at 0000 UTC, the forecasts of the skill are for 12 h, 36 h, 2.5 days, and so on up to 5.5 forecast
days. The upper limit of 5.5 days is forced by the fact that the UKMO forecasts are only 6 days long.

As mentioned before, the predictand is the regional 500-hPa anomaly correlation (AC) between the MRF
forecast and the analysis, that is, the regional pattern correlation between the forecast minus climatology and
the analysis minus climatology. We have chosen the anomaly correlation as a measure of skill because it is
the most widely used and easiest to interpret. The AC starts from an initial value of 1.0 (when the forecast is
identical to the analysis) and, on the average, decays monotonically toward zero. The value of AC = 0.6,
generally considered to be the minimum value for a forecast to retain useful skill, is attained on the average
at about 6–8 days on a hemispheric basis and somewhat earlier on a regional basis. However, the day-to-
day variability of the AC is very large, with many cases of forecasts maintaining an AC above 0.9 for over a
week or dropping below 0.0 in only 2 or 3 days (see, for example, Figs. 4, 7, 8).

We are currently using only three predictors of skill in stepwise regression (Draper and Smith 1981), all
computed daily for each region and for each forecast length.

1) Forecast agreement (denoted as AGR) is defined as the regional anomaly correlation between the MRF
forecast and each of the other 12-h older forecasts (UKMO, ECMWF, and JMA, and the average of the
latest two MRF forecasts). The individual ACs between the MRF forecast and the other four forecasts are
then averaged to create the forecast agreement predictor.

2) Forecast rms anomaly amplitude (RMSA) is defined as the regional rms amplitude of the MRF forecast
anomaly with respect to climatology.

3) Forecast persistence (PERS) is defined as the regional AC between the MRF forecast and the initial
analysis.

These three predictors are always offered to the statistical algorithm, but frequently only one or two are
chosen. The first predictor is always the one with the strongest correlation with forecast skill during training,
and in each subsequent step the predictor with the strongest correlation with the residuals is added if that
correlation is statistically significant. Many other potential predictors of skill were also tested by Kalnay
and Ham (1989), including regional values of baro-
clinic instability, Pacific–North American (PNA) pattern index, regional values of the height and zonal and
meridional wind values, etc., but they were shown to be less useful for the prediction of 0–6-day forecast
skill than the three predictors chosen above. This does not mean that their correlation with forecast skill was
lower than RMSA and PERS, but that they resulted in less reduction of variance when combined with AGR.
The stepwise regression procedure used in these tests and in the current skill prediction procedure attempts
to combine the predictors that contribute the most independent information to the regression.

All of the predictors have been computed using only numerical forecasts and climatology within the region
of interest. To some extent, the effect of upstream propagation and amplification of errors is accounted for by
the use of forecast agreement and anomaly amplitude at the verifying time as a predictor. However, in using
the daily skill predictions we observe that the verifying skill is sometimes closely related to the predicted skill
in a neighboring region (not necessarily in the “upstream” direction), but the relationship to neighboring
regions is intermittent and nonlinear and therefore not easily incorporated into the present linear scheme.

Every day we develop skill prediction equations for each region and forecast length by stepwise regression.
We use as training data the same predictors and predictands computed from the forecasts corresponding to
the previous 60 days. We do not constrain the regression equations to contain the same predictors from fore-
cast day to forecast day or from region to region. Thus, for each cycle, there are up to 60 forecasts available
from dependent data and one from independent data. It should be pointed out that the most important predictor,
by far, is the forecast agreement, which is selected over 95% of the time, compared to rms amplitude and forecast
persistance, selected about 70% and 45% of the time, respectively. With respect to the length of the training
period, experiments done by Kalnay and Ham (1989) indicated that using 60 days was better than using 30 days and that the results were similar or better than those obtained using a longer 90-day training pe-
riod. During our tests we have found that when 30–45-
day training periods are used there are more frequent
large errors in skill prediction associated with regime changes, particularly in the Tropics. On the other hand,
the advantage of using a recent series of forecasts is that the system quickly adapts to changes in the oper-
tional models or analysis systems employed as predictors.

The results of the regression provide, for each region and for all forecast lengths (from 0.5 to 5.5 days), the
predicted regional anomaly correlation. In addition, the system provides the average regional AC and the stan-
dard deviation of AC for the training period and the expected error of the predicted AC, obtained through
the reduction of variance from the dependent sample.
In addition, once it is available, the verifying AC is also provided.

A typical example of the regression equation, averaged for the spring of 1993, corresponding to region N11 (denoted “North America”), and for the 3.5-day forecast is

\[ AC_{\text{pred}}(\text{N11, 3.5 days}) = 0.66 \text{AGR} + 0.12 \text{RMSA} + 0.06 \text{PERS} + 0.08. \]

In this case, the average value of the agreement was 0.83, the average rms anomaly (in units of 100 m) was 0.98, and the average persistence was 0.29. Figure 4a, discussed in the next section, shows the daily predicted and verifying AC obtained for this period. It is clear from the coefficients in the equation and the similar size of the predictors that the agreement and the rms amplitude dominate the contribution to the prediction of the anomaly correlation.

To show the relative contributions of the various predictors, we present in Table 1 the annual average of seasonal correlations (summer 1992 through spring 1993) between predicted and verifying anomaly correlation averaged over the six regions of each of two Northern Hemisphere latitude bands. These correlations indicate that nearly all of our ability to predict skill comes from agreement alone, with very small contributions from the other two predictors on the average.

Figure 2 presents two regional predictions of the skill for the forecast from 0000 UTC 8 March 1993, which included the “great blizzard of 1993,” whose maximum amplitude was observed on 1200 UTC March 13 on the east coast of North America. The predictions of the skill correspond to region N11 (North America) and N9 (Japan) in Fig. 1. The average AC in each

---

**TABLE 1.** One-season correlations between predicted and verifying AC, averaged over one year (summer 1992 through spring 1993) and over the six regions of one latitude band, for all forecast lengths and for various combinations of predictors.

<table>
<thead>
<tr>
<th>Predictors</th>
<th>0.5</th>
<th>1.5</th>
<th>2.5</th>
<th>3.5</th>
<th>4.5</th>
<th>5.5</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>NH midlatitudes</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AGR</td>
<td>0.60</td>
<td>0.65</td>
<td>0.56</td>
<td>0.55</td>
<td>0.50</td>
<td>0.44</td>
</tr>
<tr>
<td>RMSA</td>
<td>0.55</td>
<td>0.52</td>
<td>0.46</td>
<td>0.41</td>
<td>0.35</td>
<td>0.27</td>
</tr>
<tr>
<td>AGR RMSA</td>
<td>0.62</td>
<td>0.65</td>
<td>0.57</td>
<td>0.56</td>
<td>0.51</td>
<td>0.45</td>
</tr>
<tr>
<td>PERS</td>
<td>0.32</td>
<td>0.21</td>
<td>0.12</td>
<td>0.06</td>
<td>0.04</td>
<td>0.02</td>
</tr>
<tr>
<td>AGR RMSA PERS</td>
<td>0.61</td>
<td>0.65</td>
<td>0.57</td>
<td>0.55</td>
<td>0.51</td>
<td>0.45</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>NH Tropics</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>AGR</td>
<td>0.57</td>
<td>0.67</td>
<td>0.68</td>
<td>0.64</td>
<td>0.57</td>
<td>0.47</td>
</tr>
<tr>
<td>RMSA</td>
<td>0.58</td>
<td>0.59</td>
<td>0.57</td>
<td>0.54</td>
<td>0.48</td>
<td>0.39</td>
</tr>
<tr>
<td>AGR RMSA</td>
<td>0.63</td>
<td>0.69</td>
<td>0.69</td>
<td>0.66</td>
<td>0.59</td>
<td>0.50</td>
</tr>
<tr>
<td>PERS</td>
<td>0.34</td>
<td>0.12</td>
<td>0.04</td>
<td>0.07</td>
<td>0.04</td>
<td>-0.04</td>
</tr>
<tr>
<td>AGR RMSA PERS</td>
<td>0.61</td>
<td>0.69</td>
<td>0.68</td>
<td>0.64</td>
<td>0.58</td>
<td>0.48</td>
</tr>
</tbody>
</table>

---

**Fig. 2.** Skill predictions with verification for regions (a) N11 (North America) and (b) N9 (Japan) for MRF forecasts from 0000 UTC 8 March 93. The “great blizzard of 1993” had its maximum amplitude on the East Coast on 13 March 1993. The training AC is shown by the thinnest line and its standard deviation by the shaded area; the forecast of the AC and its expected standard deviation are shown by the medium solid and dashed lines; and the thicker line is the verifying AC of the MRF forecast.
region and its standard deviation during the training period are shown with a thin line and vertical shading, and the predicted AC and the standard deviation of its distribution, as computed by the regression scheme, are superimposed. Also shown is the verifying AC, which was of course not available in real time. From the information available in real time, the forecaster can observe how today’s predicted forecast skill compares with the average skill of the model during the previous 60 days. Since the predicted skill plot includes a representation of the expected error (dashed lines) based on the reduction of variance during the training period, the forecaster can see how much confidence should be placed in the prediction of skill. A skill prediction with a standard deviation that is small compared to the training standard deviation implies that the regression has achieved a significant reduction of variance from the training sample, while a skill prediction with a standard deviation similar to the training standard deviation indicates that the regression has not extracted not much information from the training sample.

This figure (and the predictions of skill for the subsequent forecasts, not shown) indicates that throughout the period 8–13 March 1993, the MRF forecast of the storm in North America was generally excellent and that this performance was also correctly predicted, enhancing the confidence in the forecast of this unusual event. In region N9 (Japan) in the same forecast cycle, the regional AC was correctly predicted to be low. Figures 3a and 3b present the 5.5-day forecast from 0000 UTC 8 March 1993 and the corresponding verification that is valid on 1200 UTC 13 March 1993, and Fig. 3c shows the error in the 500-hPa field at the verification time. It is interesting to note that the actual errors in the North America and Japan regions are rather similar: they are both dominated by a dipole of a high and a low oriented in the east–west direction. The amplitude of these dipoles is only slightly larger in Japan than in North America, and in fact the rms error is quite similar in both regions. Nevertheless, since the actual amplitude of the anomaly was huge in North America and relatively small in Japan, the general direction of the flow was well predicted in North America but poorly predicted in Japan. This is reflected in the prediction of skill in both regions (compare Figs. 2a and 2b), and both predictions of skill verified well in this case.

This example brings up the main advantage of the use of AC as a measure of skill when compared with rms error. As pointed out by Palmer and Tibaldi (1988), the AC is related to the amplitude of the anomaly, that is, the signal that is being forecast. When the anomaly is smaller than usual, the AC is usually small; but when the anomaly is larger than usual, the AC is frequently high. The size of the anomaly plays the same role in the agreement among forecasts, and therefore the agreement and the AC tend to be both high when the anomaly is large, and vice versa. Thus, the AC is more of a relative measure of skill than the rms error and therefore is more useful in the case of large forecast anomalies, in which case the rms error is likely to be large. The rms error is a more absolute measure of skill and therefore is more meaningful in the case of small anomalies when the AC tends to be poor because the signal-to-noise ratio is small. Since the forecaster has access to the actual forecast and its anomaly with respect to climatology, he or she can judge the significance of the predicted AC. These predictions of forecast skill are available to the NMC forecasters as shown above in graphical form, but they are also available in digital form through the Internet (see the appendix).

The example presented above is an exceptionally good example of skill prediction, associated with an excellent forecast over North America (and a poor forecast over eastern Asia). In the next section we compare the predicted and verifying AC for a complete season for a few selected regions, and in section 4 we present comprehensive statistics for all regions, forecast lengths, and seasons.

3. Daily results for the spring of 1993

We have shown in section 2 an example of how the prediction of the regional forecast anomaly correlation is presented to the forecaster on a workstation or personal computer. In this section we present a comparison between 90 days of daily predicted and verifying forecast AC for several representative midlatitude Northern Hemisphere regions, as well as for one extratropical region in the Southern Hemisphere and one tropical region. As will be shown in the next section, the high-latitude (60°–90°) statistics are very similar to those of the midlatitudes (30°–60°), therefore, in the interest of space, we do not present high latitude examples. In the following figures the correlation ρ between predicted and verifying AC for the whole season is also indicated.

Figures 4a–d present the predicted and verifying regional AC for the 3.5-day forecasts in regions N11, N7, N9, and N1 (North America, Europe, Japan, and Africa, respectively), verifying during March, April, and May of 1993. The prediction of skill in North America is excellent (ρ = 0.73), with most of the major maxima and minima in forecast skill well predicted. It should be noted that the scheme seems to predict quite well almost all the extreme events during this period, which is remarkable for a linear regression algorithm. In Europe, where ρ = 0.48, the variability in skill is much lower than in North America, and the forecast of the skill is able to capture the low-frequency variability in skill but not much of the day-to-day variation. In Japan, ρ = 0.72, and the prediction of skill, as in North America, captures most of the major maxima and minima but misses some of the daily variability. The seasonal trend has not been subtracted from the ACs, but these figures (and those for the other seasons) do not show
Fig. 3. The 500-hPa MRF (a) 5.5-day forecast, (b) verification analysis, and (c) error, valid 1200 UTC 13 March 1993.
an appreciable effect of the seasonal cycle. In Africa the variation of skill is larger and is generally well predicted.

These qualitative observations are quantitatively supported by Fig. 5, where we present spectra of the variance of predicted and verifying AC and of the correlation between them. In North America the correlation is above 0.8 for all the timescales except the shortest, while in Europe the variance is relatively small and correlates poorly for all shorter timescales. In Africa there is much larger variance in skill, especially at longer timescales, and the correlation is good at most timescales, especially at the lowest frequencies.

Figures 6a–c present the daily predictions of skill and observed anomaly correlations for 5.5-day forecasts. In this case the correlation $\rho$ in Europe is only 0.34, and in this area the skill prediction scheme is not as useful on a day-to-day basis, although it does seem to capture the intraseasonal variability in skill. In North America, the correlation is 0.45, and although the scheme captures some of the low-frequency variability in skill, it generally underestimates it. A notable exception is the 6-day period around 10 March when the scheme consistently and correctly predicted high skill associated with the forecast of the East Coast blizzard of 1993. The 5.5-day prediction in Japan, on the other hand, continues to be quite skillful, with the scheme still capturing both high- and low-frequency variability and a correlation $\rho$ of 0.61.

Finally, Figs. 7a and 7b show the 3.5-day prediction of skill for the midlatitude region S11 ("southern cone") and for the tropical region S15 (Australia). For the southern cone, $\rho = 0.61$, and the prediction is quite good, even on a daily basis. For Australia, $\rho = 0.61$, and most of the skill is attained for the low-frequency variations, as is generally the case for the Tropics.

Subjective evaluation of many time series and spectra like these, considering the usual criteria for judging the skill of a forecast by its AC, suggests that the pre-
diction of skill should be useful, at least for the low-frequency variability in skill, if the correlation between predicted and verifying anomaly correlation is above 0.4. When it is above 0.6, it is also useful in the prediction of day-to-day variability in skill. To aid in evaluating lower-frequency variations in real time, we present to the forecaster and on Internet skill predictions for the most recent eight forecast days, including all available verifications.

4. Statistical verifications for all seasons, regions, and forecast lengths

In this section we present verification statistics for the three years for which the skill prediction system has been operationally available. In the statistics that follow, the correlations between predicted and verifying forecast skill and similar quantities have been computed by 3-month season, then averaged annually, in order to exclude the annual cycle. It should be noted that the analysis–forecast system used for the MRF model at NMC remained relatively stable during this period, which spans June 1991–May 1994. Although a number of relatively minor improvements were implemented, the most significant change was the replacement of the optimal interpolation analysis (OI) by the spectral statistical interpolation analysis (SSI; Parrish and Derber 1992; Derber et al. 1991), a three-dimensional variational analysis system, which took place on 25 June 1991. This change, which affected most dramatically the Tropics by eliminating or drastically reducing the model spinup, influenced the results for the summer season (JJA 1991) both directly and through the 60-day training period. A second major change took place in July 1993 when the vertical resolution was increased from 18 to 28 levels and the Kuo cumulus convection scheme was replaced by a simplified Arakawa–Schubert scheme (Pan and Wu 1994).

a. Average statistics for one year

We now present a number of statistical characteristics of the skill-prediction system, averaged by latitude band and over four recent seasons (summer of 1992 through spring of 1993). It should be noted that the annual average is obtained as the average of four individual seasonal averages, so that the correlation \( \rho \) between verifying and predicted ACs is not significantly affected by the seasonal cycle.
In Fig. 8 we present the verifying Northern Hemisphere regional anomaly correlations averaged for the four seasons for three latitude bands. It is apparent that the forecast AC of the NMC medium-range model computed regionally remains above 60% until about day 6, even in the annual average. The skill for mid- and high latitudes is virtually identical. It displays the initial “convex up” portion of the S-shaped curve characteristic of the AC of forecasts whose main source of errors is the unstable growth of initial errors rather than model deficiencies (Reynolds et al. 1994). On the other hand, the AC in the Tropics decreases initially much faster than in the extratropics, although there seems to be a crossover point at about day 6. The AC for the Tropics exhibits the “concave up” shape that, according to Reynolds et al. (1994), is characteristic of forecasts whose random errors are dominated by model deficiencies rather than atmospheric instabilities.

Figure 9 is like Fig. 8, but it shows the forecast agreement instead of the anomaly correlation. It bears a remarkable resemblance to Fig. 8. A careful comparison between the two figures shows that the forecast agreement in mid- and high latitudes is only marginally higher than the forecast anomaly correlation with the real atmosphere, whereas for the Tropics the forecast agreement is actually lower than the AC. The similarity in shape between the two figures shows that the divergence among forecasts is also dominated by instabilities in the extratropics, whereas in the Tropics it is the model differences, presumably due to different physical parameterizations, that dominate the forecast spread. These figures also show that if we compare the NMC model forecasts with an ensemble of 12-h older operational forecasts, the often repeated statement that “any two forecasts resemble each other more closely than any one of them resembles the real atmosphere” is no longer true with present state-of-the-art systems. This realistic divergence among forecasts coming from different centers is probably one of the reasons for the relative success of the present method for predicting skill, as will be discussed later.
Figure 10 presents the correlation \( \rho \) between predicted and verifying AC for the dependent (training) sample. It is somewhat surprising that the \( \rho \) is significantly higher for the tropical regions than for the extratropics, reaching a peak of 0.75 at day 2.5 in the Tropics. For mid- and high latitudes, the dependent sample correlation decreases slowly from about 0.7 at 0.5 days to less than 0.6 at day 5.5. Figure 10 also shows the same correlation \( \rho \) for the independent (actual forecast) samples. As could be expected, the scores for the independent samples are lower than those for the dependent samples by about 10\%–15\%. The mid-latitudes actual skill in predicting the AC decreases from 0.65 at day 1.5 to 0.45 at day 5.5. The high-latitude scores are similar to the midlatitude scores but about 5\% lower. The skill in predicting forecast skill is again higher in the Tropics, probably because the anomaly correlations themselves are much more variable and have much more low-frequency variability in the Tropics than in mid- or high latitudes (see Fig. 5). Therefore, there is more “room” to capture variability in the signal.

In Fig. 10 we also include a similar correlation obtained by Barker (1991) using a perfect model Monte Carlo ensemble rms spread to predict the rms forecast error. It is remarkable that our operational results are actually better than those of Barker’s perfect model Monte Carlo experiments, which used a much larger number of ensemble members (120 versus only 4 in the NMC system) and which was not encumbered by model deficiencies. The fact that at short forecast intervals his results are poorer than the NMC results.
should not be surprising since, as mentioned before, his Monte Carlo approach starts from random perturbations not dynamically related to the current state of the atmosphere and therefore has zero predictive skill at the initial time. The NMC operational system, on the other hand, starts from very "realistic" perturbations since each of the operational systems is a state-of-the-art system and each one of their analysis cycles is a "breeding ground" for the same type of fast-growing errors that plague all operational forecasts (Toth and Kalnay 1993). This is because errors not related to fast-growing modes are more strongly suppressed. Beyond day 3 the results are more similar between Barker's system and the NMC operational system, although the latter remains clearly superior. The difference is important since, as indicated before, skill predictions with correlations above 0.4 seem to have some usefulness, especially for low-frequency variability, and those above 0.6 appear to be quite useful.

There are several additional possible explanations for such superiority: the NMC system predicts AC, while Barker (1991) predicts rms error. The NMC system uses other predictors in addition to forecast agreement, that is, forecast persistence, a proxy for atmospheric stability, and the rms amplitude of the anomaly. The last predictor is especially useful for predicting AC (Branstator 1986), since at low values of the anomaly the signal-to-noise ratio of the AC becomes small (Palmer and Tibaldi 1988). As shown in the example of the prediction of skill from the forecast from 8 March 1993, the advantage of higher predictability that the AC has when compared to rms error is not simply a statistical artifact: the rms errors in Japan and North America were rather similar in both shape and size, but the anomalous circulation was far stronger over the United States. Therefore, an rms error score would have indicated similarly poor forecasts, whereas the AC score correctly indicated that the anomalous circulation was not well captured over Japan but was very well represented over North America.

In addition to the advantage of using AC, it is possible that the "poor person's" Monte Carlo predictor used in NMC's system, having systems with many more degrees of freedom than the simpler model of Barker (1991), has more skill variability and therefore more room to predict the skill. As mentioned above, we see this effect in Fig. 7b (Australia, a tropical region) when compared to the midlatitude regions of Figs. 4 and 7a. This may also explain the difference between our results and those of Houtekamer and Derome (1994), who used a "breeding cycle" to generate analysis errors as well as ensemble perturbations instead of starting with random perturbations unrelated to the initial state. When they performed perfect model ensemble forecasts and used, as did Barker, the ensemble spread to forecast the rms errors, they obtained correlations between 0.39 and 0.47 throughout the first 10 days of the forecast. This relatively low value for a perfect model system may be due not only to the lesser predictability of the rms error but also to the use of a T21 quasigeostrophic model, with fewer degrees of instability than the real atmosphere or the operational models.

In Fig. 11 we compare the correlations $\rho$ between predicted and verifying AC for the Southern Hemisphere with those for the Northern Hemisphere. The scores for the Southern Hemisphere are about 10% worse for both the Tropics and the extratropics, possibly due to a lower general skill of the forecasts them-
selves, which is associated with fewer observations and therefore poorer initial conditions.

Figure 12a compares regional correlations $\rho$ for several regions of the Northern Hemisphere. Africa, an example of a tropical region, has a correlation between predicted and verifying AC varying between 0.6 and 0.73; North America and Japan have values of $\rho$ between 0.5 and 0.7; and Europe starts at 0.62 at day 0.5, but the skill decreases to only 0.3 at day 5.5. The low correlation in Europe and the high correlation in Africa are probably associated with a much better overall forecast skill in Europe than in Africa and therefore with much smaller forecast skill variability (as shown also in Fig. 5). The comparison is somewhat similar in the Southern Hemisphere (Fig. 12b), where the three regions Brazil, Australia, and southern cone maintain a correlation close to 0.6 up to day 3.5, while at the end of the period the three tropical regions including Peru remain above 0.4 longer than the extratropical southern cone, which had a higher average AC.

Figure 13 presents the annual average of the percentage of cases in which the three predictors—forecast agreement, forecast persistence, and rms amplitude of the anomaly—are selected as predictors by the stepwise regression during the training. It shows that agreement is selected between 90% and 100% of the predictions, rms amplitude between 65% and 80%, and forecast persistence in 40%–50% of the predictions. These percentages are essentially independent of the forecast length and are also remarkably independent of latitude. Recall that after the first predictor is chosen, additional predictors are selected based on their contribution of additional information in the regression.

b. Variability over 12 seasons

Next we present statistics indicating the interseasonal variability of the statistics presented above. The available data correspond to three years starting with the NH summer (JJA) of 1991 and ending with the spring (MAM) of 1994. For the sake of brevity, results are only presented for the forecast length of 3.5 days and for the Northern Hemisphere.

Figure 14 shows both the AC and the correlation $\rho$ for all seasons and for both midlatitudes and the Tropics. For the midlatitudes, the seasonal dependence is weak, with the average AC at day 3.5 above 0.85, except during the summers, when it is slightly lower. The correlation $\rho$ between predicted and verifying AC for midlatitudes varies between 0.5 and 0.6 and shows even less seasonal dependence. The AC in the Tropics, like the midlatitudes, has the poorest forecast skill during the summers, with higher forecast skill during the
other seasons, during which intrusions of midlatitude air masses make the tropical flow more predictable (G. White 1992, personal communication). In the Tropics the higher skill in 1992 may be associated with El Niño. The worst correlation ρ for the Tropics occurred during the first season, which was affected by the implementation of the SSI at the end of June 1991 and the corresponding inhomogeneity of the training period of 60 days previous to each forecast.

Figures 15a–d show, for different regions, the verifying average AC and the correlation ρ between predicted and verifying AC for both the training period (dependent sample) and the actual predictions of skill (independent sample). The individual midlatitude regions (North America, Europe, and Japan) do not exhibit a clear seasonal dependence of AC, except for lower values in the summer. Africa, like the tropical averages presented above, shows the lowest skill in the summer.

The correlation ρ in the same four regions has more variability, but the variation does not correlate strongly with season. It is very encouraging that there is a clear correlation over the 12 seasons between the correlations for the training (dependent) sample (which is the basis for the error standard deviations of the predicted AC in Fig. 2) and for the verifying (independent) predictions. This suggests that the average accuracy of the skill prediction can be approximately estimated from the dependent sample.

5. Discussion

We have presented a real-time statistical scheme to predict the regional anomaly correlation of NMC’s medium-range global forecasts. Based on earlier work, we have chosen three predictors to use in a stepwise linear regression scheme with the previous 60 days’ forecasts used as the training period. The most often chosen predictor is the average agreement of the current MRF with four other global forecasts, and the other predictors are the rms anomaly amplitude and the persistence of the forecast. This scheme has been in operational use for three years, and its predictions are available electronically through the Internet (see the appendix). While it is specifically designed to predict the skill of NMC’s medium-range forecast model’s forecasts, the predictors it uses are related to the overall difficulty of each regional forecast; therefore, the skill predictions should be useful with forecasts from other models.

The period for which we have shown results has been relatively homogeneous for the NMC model but includes major improvements in the analysis in June 1991 (the SSI, which affected mostly the Tropics and introduced temporary problems with the training) and in the vertical resolution and convection parameterization in July 1993. Here we have presented the results for the daily prediction of skill for a recent season, comprehensive summary regional statistics for all the regions and latitude bands, and their seasonal dependence over the three years.

We find that for most midlatitude regions, and most seasons, the correlation between observed and predicted skill at 3–4 days is between 0.4 and 0.7, whereas over the Tropics it is usually higher. Inspection of the daily plots indicates that the predictions of skill should be useful when the correlation between predicted and observed AC is above 0.4, especially in the prediction of the low-frequency variability in the skill. When the correlation is 0.6 or higher we find that there is significant skill in predicting the day-to-day variability in forecast AC. A 5-day running mean of the predicted and observed AC results in a considerable improvement of the correlations, typically of 0.1–0.2, confirming that there is considerable predictability of low-frequency variability in the skill (Kalnay and Ham 1989).

Our results are somewhat better than those obtained by Barker (1991) using large ensembles and by Houtekamer and Derome (1994) using small ones. They obtained correlations between predicted and observed skill of only about 0.3–0.5. Although our results are more encouraging than theirs, they also confirm that the prediction of skill is a very difficult problem. This is due to the fact that the actual trajectory followed by the atmosphere is only one of an ensemble of possible trajectories. It would be possible to perfectly predict the skill of a large ensemble of forecasts if we compared them against a large ensemble of possible verifications. In reality, however, the atmosphere goes through a single realization so that the prediction of the skill of a single individual forecast or even of an ensemble of forecasts is much less than perfect, as shown in the experiments of Barker (1991). This also suggests that in the future we should develop a system to predict
the reliability of probability forecasts, a task made possible by the implementation of ensemble forecasting at NMC (Tracton and Kalnay 1993).

The results of our system can then be considered as moderately encouraging even though we are using an extremely small ensemble (four members in addition to the base NMC forecast) compared to Barker (1991), who used large ensembles, and Houtekamer and Derome (1994), who used ensembles of two members. We attribute this to two principal factors.

First, we are using one of the most realistic ensembles possible: the members are each an operational forecast derived independently in different global forecasting centers, with initial conditions also derived from independent analysis cycles. Both the analysis cycles and the models are somewhat different from each other, and, given the friendly competition among centers to show good forecast skill, they each represent a slightly different state-of-the-art system. Therefore, the differences among the analyses and among the forecast models represent in the most realistic way the present uncertainties about our knowledge of the state of the atmosphere and the detailed laws that govern its evolution. For this reason, for example, our predictions of AC start and retain a correlation between predicted and verifying AC above 0.6 for several days, whereas Barker (1991), using a perfect model and a very large ensemble of forecasts (but starting with randomly chosen perturbations), started with zero correlation at the initial time.

Second, we use the anomaly correlation as a measure of skill rather than the rms error. This allows the use of an additional predictor of skill, the anomaly amplitude, which adds up to 10% (section 2 above; Kalnay and Ham 1991) to the correlation between predicted and verifying skill. As discussed in section 2, this additional information is not just a simple exploitation of the “signal-to-noise ratio” that occurs both in the fore-
cast agreement and in the verifying AC of the forecast. As shown in the example of the case of 13 March 1993, the rms errors in the Japan and North America regions were quite similar in shape and magnitude. Nevertheless, the anomaly itself was huge in North America (the "great blizzard of 1993") and relatively small in Japan. Therefore, the forecast in North America captured well the observed anomalous circulation, whereas in Japan the forecast of the regional circulation was very poor. As a result, even though the absolute regional rms errors were quite similar, the forecast AC was unusually high in North America and very low in Japan. The prediction of the AC captured correctly this difference, undoubtedly taking advantage of the additional information provided by the forecast of the anomaly amplitude (Branstator 1986). Finally, our third predictor, forecast persistence, makes a very small additional contribution to our skill.

The system described here has used the same ensemble members since the JMA forecasts became available at NMC in 1991. We plan to add more members to the ensemble as they become available to us in real time via global telecommunications, such as those from Canada, Australia, Germany, and the U.S. Navy, which should improve the short-range prediction of skill.

We are considering testing Kalman filtering and/or neural networks as an alternative to the linear regression method used here. These methods would eliminate the need for a hard cutoff to the training period and would probably simultaneously improve the timeliness of the prediction model and the response to regime transitions. The results presented here indicate that there is enough information in our predictors to use either of these methods effectively. We are also developing new predictors based on the breeding ensemble (Toth and Kalnay 1993), which will be used in a manner analogous to the predictors we now use but for forecast ranges to 7 days, as well as for the "week 2" forecasts soon to be implemented. If the ensemble forecasts are extended to a month, we may attempt to predict the skill of appropriate time averages.

APPENDIX

Access to Operational Prediction of MRF Skill Through the Internet

The skill-prediction files are available from NMC through the Internet. The file /pub/dd/mrfskill/readme, which can be downloaded from nic.fb4.noaa.gov, gives the names and descriptions of these files and instructions for downloading them. Typical commands to download this file using anonymous ftp:

type get command  get readme yourfn(cr)
type bye command  bye(cr),

where password is your e-mail address and yourfn is the name of the file on your computer into which to download the documentation.

REFERENCES


