Trends in the Quality of National Weather Service Forecasts

ALLAN H. MURPHY AND THOMAS E. SABIN

Department of Atmospheric Sciences, Oregon State University, Corvallis, OR 97331

(Manuscript received 9 September 1985, in final form 21 January 1986)

ABSTRACT

This paper describes the results of a study of trends in the quality of National Weather Service (NWS) forecasts from 1967 to 1985. Primary attention is focused on forecasts of precipitation probabilities, maximum temperatures, and minimum temperatures. A skill score based on the Brier score is used to verify the precipitation probability forecasts, whereas the temperature forecasts are evaluated using the mean absolute error and percentage of errors greater than 10°F. For each element, trends are examined for objective forecasts produced by numerical-statistical models and for subjective forecasts formulated by NWS forecasters. In addition to weather element, type of forecast, and verification measure, results are stratified by season (cool and warm), lead time (three or four periods), and NWS region (four regions and all regions combined).

At the national level, the forecasts for these three weather elements exhibit positive and highly significant trends in quality for almost all of the various stratifications. Exceptions to this general result are associated solely with the minimum temperature forecasts, primarily for the 60 h lead time. These national trends are generally stronger for the objective forecasts than for the subjective forecasts and for the cool season than for the warm season. Regionally, the trends in quality are almost always positive and are statistically significant in a majority of the cases. However, nonsignificant trends occur more frequently at the regional level than at the national level. As a result of the positive trends in performance, current levels of forecast quality for these weather elements are markedly higher than the levels that existed 15-20 years ago.

1. Introduction

Over the last 20 years, the meteorological community has devoted a substantial fraction of its resources to basic and applied research programs designed to improve the weather forecasting system. These programs have led to the development and implementation of new data acquisition systems, greatly enhanced facilities to assimilate and display data, and improved numerical and numerical–statistical forecasting models. However, since such programs generally have been justified on the basis of anticipated improvements in the quality of weather forecasts, it seems appropriate to investigate trends in forecast quality over this period. The results of such investigations should be of considerable interest to a variety of individuals in the meteorological community as well as to actual and potential users of the forecasts.

Several studies of trends in forecast quality have been undertaken in recent years. For example, Charba and Klein (1980), Cook and Smith (1977), Glahn (1985), Murphy and Brown (1984), Ramage (1982) and Zurnedorfer et al. (1979) have investigated trends in the quality of forecasts produced by the U.S. National Weather Service (NWS). In addition, Bosart (1983) and Sanders (1986) have analyzed trends in the quality of forecasts formulated in conjunction with synoptic laboratory programs at their respective universities. Moreover, interest in this topic is not limited to the United States—Stuart et al. (1983) have studied trends in the quality of weather forecasts produced by the Atmospheric Environment Service in Canada, and Stern (1980) has investigated trends in the quality of temperature forecasts formulated by weather forecasters in Australia.

Conclusions forthcoming from these studies have varied widely from “little if any positive trend in performance” to “statistically significant improvements in quality.” Since the studies have generally been based on different time periods, weather elements, types of forecasts, regions or locations, lead times, and/or verification measures, such contradictory results should not be too surprising. However, controversies have arisen regarding the existence of significant trends in quality even in situations involving essentially the same elements, types of forecasts, and time periods. For example, Ramage (1982) and Glahn (1985) reached quite different conclusions from their studies of trends in the quality of NWS subjective precipitation probability forecasts over the periods from 1966 to 1978 and 1967 to 1982, respectively. Thus, it would appear to be desirable to investigate trends in the quality of NWS forecasts in greater detail by (i) updating the studies conducted by Ramage and Glahn, (ii) analyzing trends in the quality of objective precipitation–probability forecasts produced by the model output statistics (MOS) system, and (iii) extending this work to include studies of such trends for other important variables contained in NWS forecasts.

The primary purpose of this paper is to describe the
results of a study of trends in the quality of several types of NWS forecasts over the period from 1967 to 1985. Primary attention is focused on objective and subjective forecasts of precipitation probabilities and maximum and minimum temperatures. Brief mention is also made of some results related to cloud amount and wind speed forecasts. Section 2 contains a description of the verification data employed in this study. A short discussion of the measures used to verify the forecasts appears in section 3, which also briefly outlines the method used to estimate the trends in quality. Sections 4 and 5 contain the main results of the study, with the former devoted to the precipitation probability forecasts and the latter focused on the maximum and minimum temperature forecasts. A discussion of the results is included in section 6, and section 7 consists of a brief conclusion.

2. Verification data

Two series of reports were consulted to obtain the verification statistics for the objective and subjective forecasts examined in this study. For the period April 1967 through September 1980, these statistics were taken from NOAA technical memoranda in the NWS/FCST series (e.g., Polger, 1983), and for the period, October 1980 through March 1985, the scores were taken from the six-monthly office notes published by the NWS/Techniques Development Laboratory (TDL) (e.g., Carter et al., 1985). Exceptions to this procedure included (1) the verification statistics for the greater than 10°F temperature errors from the 1980–81 cool season through the 1984 warm season, which were provided by G. M. Carter of TDL and (2) the scores for the 60 h minimum temperature forecasts for the period 1977–85, which were available only in the TDL office notes.

Objective MOS probability of precipitation (PoP) forecasts were introduced in January 1972 (Glahn and Lowry, 1972); thus, the time series of scores for the objective PoP forecasts begin with the 1972 warm season. Verification statistics for the objective numerically-statistical temperature forecasts first appeared in the NWS/FCST technical memoranda for the 1970 warm season (Derouin and Cobb, 1972). These objective temperature forecasts were based on the perfect “prog” system prior to October 1973 and on the MOS system beginning with the 1973–74 cool season. Scores for the subjective precipitation probability and maximum and minimum temperature forecasts have been recorded for both cycle times (i.e., 0000 and 1200 GMT) since April 1967.

In the case of objective forecasts, the verification statistics taken from the NWS/FCST technical memorandum relate to the so-called “final” forecasts based on the primitive-equation and trajectory models. From October 1980 through March 1985, the scores for the objective forecasts involve the so-called “early” forecasts based on the limited-area, fine-mesh model (except for the PoP forecasts in the 1980–81 cool season—these scores involve the final forecasts). Once again, the 60 h minimum temperature forecasts represent an exception; in this case, the verification statistics relate to the early forecasts for the entire eight-year period.

From 1967 through 1972 the NWS/FCST reports contained verification statistics for two types of subjective forecasts: “FP” forecasts and “local” forecasts. The FP forecasts were produced by area forecast centers for their own location and usually one or more other locations, thereby creating a data base consisting of approximately 100 stations. Local forecasts were produced by subsidiary offices and were generally derived from the FP forecasts with relatively little modification. The number of stations contained in the data base for the local forecasts varied considerably from year to year and these data were not included in the reports after the 1972–73 cool season. Moreover, the FP stations correspond closely to the current NWS forecast offices. Thus, since the FP verification statistics represent a more coherent data set, they were used in this study.

Verification data are available for the 48 contiguous states of the United States and for four NWS administrative regions: the Eastern, Southern, Central and Western regions. The national results are based on overall verification statistics for approximately 100 stations prior to 1973 and between 80 and 90 stations from 1973 to the present. Each region generally consisted of 20 to 25 stations with the exception of the Western Region, which frequently contained only 15 to 20 stations. At the national level, sample sizes for each season/lead time combination ranged from 9000 forecasts to 23 000 forecasts.

The verification statistics are computed for 6-month periods corresponding to the warm and cool seasons. The warm season consists of the months April through September, whereas the cool season includes the months October through March. An exception to this rule occurred in 1970–71 when the warm season was defined as May through October and the cool season was defined as November through April. This temporary change in the definitions of the seasons should have little effect on the results.

Both objective and subjective forecasts are formulated twice each day, in conjunction with numerical model simulations made at the cycle times of 0000 and 1200 GMT. The objective forecasts, based primarily on numerical model output, are disseminated a few hours after these cycle times. On the other hand, the subjective forecasts are issued daily at approximately 1000 and 2200 GMT (i.e., about 10 h after the respective cycle times). Thus, the forecasters are able to consult the objective forecasts and other recent data (in addition to numerical model output) prior to formulating their forecasts.

The PoP forecasts are produced for three 12 h periods—namely, 12–24 h, 24–36 h, and 36–48 h—with
the first period starting 12 h after the corresponding cycle time. Glahn (1985) found that the scores for the PoP forecasts differ significantly between the cycle times, a result that may be related to the fact that the corresponding periods involve different times of the day (i.e., daytime and nighttime). In any case, the verification statistics for the two cycle times were combined here, as in Glahn (1985), to produce a single score for each combination of season and lead time.

Maximum and minimum temperature forecasts corresponding to the 0000 GMT cycle time include tomorrow’s maximum temperature, tomorrow night’s minimum temperature, the second day’s maximum temperature, and the second night’s minimum temperature. Forecasts associated with the 1200 GMT cycle time specify tonight’s minimum temperature, tomorrow’s maximum temperature, tomorrow night’s minimum temperature, and the second day’s maximum temperature. Thus, the lead times for these temperature forecasts are approximately 24, 36, 48, and 60 h.

The subjective temperature forecasts relate to nighttime minima and daytime maxima, whereas the objective forecasts involve temperatures valid for a calendar day. From April 1967 through March 1975 both types of forecasts were verified using 12 h maximum and minimum temperatures (i.e., temperature almost hourly occurring between 1200 and 0000 GMT and minimum temperatures occurring between 0000 and 1200 GMT). On the other hand, calendar-day maximum and minimum temperatures were used to verify these forecasts from April 1975 through September 1983. Obviously, the verification procedure favored the objective forecasts during this latter period, especially when the daily maxima or minima occurred outside the period for which the subjective forecasts were valid (e.g., in the case of a nighttime cold front passage).

From October 1983 through September 1984 the temperature forecasts were once again verified using 12 h maximum and minimum temperatures. Although this verification procedure appears to favor the subjective forecasts, nighttime minimum temperatures in the western United States often occur after 1200 GMT (0400 PST) and daytime maxima in this region may occur after 0000 GMT (1600 PST), thereby producing systematic errors and unwarranted errors in the forecasts. Such errors should have been eliminated with the start of the 1984—85 cool season, when proper nighttime maxima and nighttime minima were first used to verify both types of forecasts. The effects of these factors on the verification statistics associated with the temperature forecasts will be discussed in section 6.

3. Measures of quality and methods of analysis

The results of studies of trends in performance depend upon many factors, including the verification measures used to determine forecast quality. We attempted to select the most appropriate verification statistics available for the respective forecast elements. In the case of the PoP forecasts, the measure of quality employed here is the skill score. It represents the improvement (expressed in percent) in the accuracy of the forecasts of interest over the accuracy of forecasts based solely on the long-term climatological probabilities. Accuracy is measured by the Brier score (Brier, 1950), which represents the mean square error of probabilistic forecasts. We use the skill score rather than the Brier score to measure the quality of the forecasts because the former is much less sensitive than the latter to year-to-year variations in the relative frequency of the event in question. The climatological probabilities have been determined separately for each month at each location.

In the case of the maximum and minimum temperature forecasts, the basic measure of quality is the mean absolute error. In addition to this familiar verification statistic, we also examined trends in the percentage of errors in the temperature forecasts exceeding 10°F. Both of these statistics are measures of accuracy and, as such, they are subject to year-to-year fluctuations due simply to the interannual variability in day-to-day temperature values. In general, increases in such variability will lead to decreases in forecast quality.

As noted in section 1, the primary objective of this paper is to investigate trends in the quality of NWS precipitation probability and maximum and minimum temperature forecasts. Specifically, we want to obtain quantitative estimates of the trends in performance and to determine their statistical significance. A simple linear regression model is used as a method of analyzing these trends. In this model, the values of the dependent variable are the average scores associated with the appropriate verification measure, and the values of the independent variable are the corresponding years in which these scores were recorded. The slope of the regression line provides an estimate of the rate of change of the average score with time (i.e., the trend in quality).

Under the assumption that the residuals (i.e., the differences between the actual scores and the scores estimated by the regression model) are independent and normally distributed, it is possible to determine the statistical significance of the slope or trend in quality. In particular, we can test the null hypothesis that this slope is zero (no trend) against the alternative hypothesis that the slope is nonzero (positive or negative trend). A t-statistic involving the estimated value of the slope and its standard error can be computed as a basis for performing such a test. In fact, it is possible to compute the probability that a value as large as or larger than the actual value of the t-statistic would be observed under the assumption that the true slope is zero. If this probability value (p-value) is small—say, less than 0.05—then it is not unreasonable to conclude that the slope is nonzero. This method of analysis has been applied separately to each time series of scores (i.e.,
the data have been stratified by element, forecast type, season, region, lead time and verification measure).

4. Probability of precipitation forecasts

Trends in the skill of objective (OBJ) and subjective (SUB) PoP forecasts are examined over the period from 1967 to 1985. When comparing these trends for various combinations of region, forecast type, season, and lead time, several factors should be kept in mind. First, NWS regions are subject to different weather regimes, and such differences can significantly influence the skill scores. In addition, national skill scores involve larger samples of forecasts and observations than regional skill scores; thus, the national scores generally exhibit less year-to-year variability than the regional scores. Second, PoP forecasts usually exhibit greater skill in the cool season than in the warm season (for a given lead time). Third, skill generally decreases as lead time increases; in this regard, the skill of the 36–48 h PoP forecasts was quite modest when the verification statistics were first computed in the late 1960s. Fourth,
the skill of the OBJ forecasts was relatively low compared to the skill of the SUB forecasts when the former initially became available in the early 1970s. In this regard, it should be noted that lower levels of skill at any point in time obviously provide the possibility of greater improvements in skill over any subsequent period.

a. National results

Time series of the national skill scores for the OBJ and SUB forecasts are depicted in Figs. 1 and 2, respectively, with separate curves for each combination of forecast type, season and lead time. The slopes of the regression lines and the associated p-values for these time series are included in Table 1. For the SUB forecasts, slopes and p-values have been determined for the entire 18-year period and for a 13-year period corresponding to the period for which the OBJ forecasts have been evaluated.

The results presented in Table 1 indicate that the PoP forecasts possess statistically significant positive trends in skill for all combinations of forecast type, season, and lead time (trends are considered to be statistically significant when the p-values are less than 0.05). In the case of the OBJ forecasts, the slopes are highly significant (p-values less than 0.01) for both seasons and all lead times. Comparison of the slopes for the cool and warm seasons reveals that the latter are somewhat larger than the former. Differences in slopes as a function of lead time are relatively modest, with the 12–24 h forecasts exhibiting the strongest rate of improvement in skill in both seasons.

Trends in the skill of the SUB forecasts over the entire 18-year period are also positive and highly significant. These trends are stronger in the cool season than in the warm season, and the respective slopes increase as lead time increases in both seasons. The slopes for the 18- and 13-year periods exhibit similar magnitudes in the warm season, whereas the former are larger than the latter in the cool season.

Comparison of the trends in skill of the two types of forecasts over the 13-year period indicates that the OBJ forecasts possess a stronger rate of improvement than the SUB forecasts for almost all combinations of season and lead time. The differences in the magnitudes of the respective slopes are particularly large for the 12–24 h forecasts. In these cases, the OBJ slopes are more than twice as large as the SUB slopes.

Table 1. Trends in skill of the objective (OBJ) and subjective (SUB) precipitation probability forecasts over the period 1967–85 in the (a) cool season and (b) warm season, as indicated by the slopes (%/year) and p-values* of regression lines fitted to the corresponding time series of skill scores.

<table>
<thead>
<tr>
<th>Lead time (h)</th>
<th>Type of forecast</th>
<th>Period** (years)</th>
<th>National</th>
<th>Eastern</th>
<th>Southern</th>
<th>Central</th>
<th>Western</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>Slope</td>
<td>p-value</td>
<td>Slope</td>
<td>p-value</td>
<td>Slope</td>
</tr>
<tr>
<td>(a) Cool season</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>12–24</td>
<td>SUB</td>
<td>1967–85</td>
<td>0.519</td>
<td>0.0000</td>
<td>0.529</td>
<td>0.0029</td>
<td>0.295</td>
</tr>
<tr>
<td></td>
<td>SUB</td>
<td>1972–85</td>
<td>0.401</td>
<td>0.0127</td>
<td>0.238</td>
<td>0.372</td>
<td>0.376</td>
</tr>
<tr>
<td></td>
<td>OBJ</td>
<td>1972–85</td>
<td>0.842</td>
<td>0.0021</td>
<td>0.569</td>
<td>0.0477</td>
<td>0.915</td>
</tr>
<tr>
<td>24–36</td>
<td>SUB</td>
<td>1967–85</td>
<td>0.826</td>
<td>0.0000</td>
<td>0.747</td>
<td>0.0001</td>
<td>0.737</td>
</tr>
<tr>
<td></td>
<td>SUB</td>
<td>1972–85</td>
<td>0.574</td>
<td>0.0005</td>
<td>0.369</td>
<td>0.1139</td>
<td>0.748</td>
</tr>
<tr>
<td></td>
<td>OBJ</td>
<td>1972–85</td>
<td>0.683</td>
<td>0.0004</td>
<td>0.452</td>
<td>0.0489</td>
<td>0.932</td>
</tr>
<tr>
<td>36–48</td>
<td>SUB</td>
<td>1967–85</td>
<td>0.937</td>
<td>0.0000</td>
<td>0.969</td>
<td>0.0000</td>
<td>0.852</td>
</tr>
<tr>
<td></td>
<td>SUB</td>
<td>1972–85</td>
<td>0.763</td>
<td>0.0004</td>
<td>0.663</td>
<td>0.0263</td>
<td>0.993</td>
</tr>
<tr>
<td></td>
<td>OBJ</td>
<td>1972–85</td>
<td>0.750</td>
<td>0.0001</td>
<td>0.670</td>
<td>0.0147</td>
<td>0.897</td>
</tr>
<tr>
<td>(b) Warm season</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>12–24</td>
<td>SUB</td>
<td>1967–84</td>
<td>0.460</td>
<td>0.0003</td>
<td>0.659</td>
<td>0.0004</td>
<td>0.366</td>
</tr>
<tr>
<td></td>
<td>SUB</td>
<td>1972–84</td>
<td>0.445</td>
<td>0.0306</td>
<td>0.754</td>
<td>0.0204</td>
<td>0.319</td>
</tr>
<tr>
<td></td>
<td>OBJ</td>
<td>1972–84</td>
<td>0.971</td>
<td>0.0003</td>
<td>1.250</td>
<td>0.0009</td>
<td>0.890</td>
</tr>
<tr>
<td>24–36</td>
<td>SUB</td>
<td>1967–84</td>
<td>0.605</td>
<td>0.0000</td>
<td>0.889</td>
<td>0.0000</td>
<td>0.483</td>
</tr>
<tr>
<td></td>
<td>SUB</td>
<td>1972–84</td>
<td>0.600</td>
<td>0.0018</td>
<td>0.952</td>
<td>0.0031</td>
<td>0.482</td>
</tr>
<tr>
<td></td>
<td>OBJ</td>
<td>1972–84</td>
<td>0.730</td>
<td>0.0010</td>
<td>1.058</td>
<td>0.0043</td>
<td>0.713</td>
</tr>
<tr>
<td>36–48</td>
<td>SUB</td>
<td>1967–84</td>
<td>0.624</td>
<td>0.0000</td>
<td>0.908</td>
<td>0.0000</td>
<td>0.563</td>
</tr>
<tr>
<td></td>
<td>SUB</td>
<td>1972–84</td>
<td>0.677</td>
<td>0.0002</td>
<td>0.913</td>
<td>0.0007</td>
<td>0.714</td>
</tr>
<tr>
<td></td>
<td>OBJ</td>
<td>1972–84</td>
<td>0.853</td>
<td>0.0000</td>
<td>1.024</td>
<td>0.0002</td>
<td>0.884</td>
</tr>
</tbody>
</table>

* The p-value represents the probability associated with a test of a null hypothesis that the regression coefficient (i.e., the slope) is equal to zero.
** The period for the cool season extends from 1967–68 or 1972–73 through 1984–85, whereas the period for the warm season extends from 1967 or 1972 through 1984.
Examination of the time series of OBJ and SUB skill scores in Figs. 1 and 2, respectively, reveals some interesting results. For example, the skill of the OBJ forecasts (Fig. 1) in the cool season increased markedly in 1976–77 (plotted as 1976) but quite modestly thereafter. In the warm season, increases in the skill of the OBJ forecasts appear to have occurred somewhat more gradually, with periods of definite improvement in the mid-1970s (especially for the 12–24 h lead time) and in the early 1980s.

The skill scores for the SUB forecasts (Fig. 2) in the cool season increased noticeably between 1970–71 and 1972–73 and again in 1976–77, with only modest trends toward higher scores in the last six years. Year-to-year variations in skill are larger in the warm season than in the cool season; nevertheless, the time series for the former reveals increases in skill between 1971 and 1974 and again in 1978. Relatively little improvement in performance is evident for these forecasts since the late 1970s, although the levels of skill attained in the last two years are as high or higher for all three lead times than any levels reached heretofore.

b. Regional results

Slopes and p-values associated with regression models for the four NWS regions are also included in Table 1 (diagrams containing the corresponding time series of skill scores have been omitted to conserve space). The regional results are generally similar to the national results. For example, regional trends in skill are positive for all combinations of forecast type, season, and lead time, and the associated p-values indicate that a vast majority of these trends are statistically significant. Moreover, as in the case of the national results, the regional trends generally increase as lead time increases for the SUB forecasts and are usually stronger in the cool season than in the warm season for both types of forecasts. In addition, the rates of improvement in the regions are most often stronger for the OBJ forecasts than for the SUB forecasts. Differences in trends among regions are generally not large, although the Eastern Region exhibits relatively weak trends in skill in the cool season for the OBJ forecasts and relatively strong trends in skill in the warm season for both types of forecasts.

5. Maximum and minimum temperature forecasts

In the case of the temperature forecasts, we analyzed time series for two verification measures, mean absolute error (MAE) and percentage of errors greater than 10°F (PEG10). However, since the PEG10 results are similar to the MAE results, the former have been omitted to conserve space. Moreover, discussion of the PEG10 results will be limited to those situations in which they differ substantially from the MAE results. As emphasized previously in connection with the PoP forecasts, care should be exercised when comparing the trends in accuracy of the maximum or minimum temperature forecasts for the various combinations of region, forecast type, season and lead time. It should also be noted that, unlike the PoP forecasts, the accuracy of temperature forecasts is generally greater in the warm season than in the cool season.

As indicated in section 2, the method of verifying the temperature forecasts has been modified several times over the 18-year period. These modifications—especially the modification made in October 1983—are associated with noticeable changes in the values of the verification measures. As a result, we have analyzed the trends in accuracy of the temperature forecasts for the period prior to October 1983 (referred to as the “reduced period”) as well as for the entire period. Finally, since the OBJ and SUB forecasts relate to different temperatures (calendar-day maxima and minima for the former and daytime maxima and nighttime minima for the latter), the various modifications in the verification method might be expected to affect the two types of forecasts differently.

a. National results

1) Maximum temperature

Time series of national MAE statistics for the OBJ and SUB maximum temperature ($T_{max}$) forecasts are depicted in Figs. 3 and 4, respectively (60 h $T_{max}$ forecasts are available only for a relatively short period and have not been considered here). To facilitate the interpretation of the results, we have added dashed vertical lines to these diagrams (and to Figs. 5 and 6) to indicate when modifications were made in the method of verification. These lines divide the overall period into the following subperiods in which the indicated types of observed temperatures were used to verify the forecasts: period A, 12 h temperatures; period B, calendar-day temperatures; period C, 12 h temperatures; and period D, daytime maxima and nighttime minima. The estimated slopes of the regression lines and the corresponding p-values for the MAE statistics are included in Table 2.

The results in Table 2 reveal that the $T_{max}$ forecasts possess highly significant positive trends in accuracy for all combinations of forecast type, season and lead time. For the OBJ forecasts, the rate of increase in accuracy is greater in the cool season than in the warm season for all three lead times. Differences in slopes as a function of lead time are relatively small, although the slopes associated with the PEG10 statistics in the cool season increase as lead time increases.

Trends in the accuracy of the SUB forecasts are also stronger in the cool season than in the warm season, and the rate of increase in accuracy increases as a function of lead time in both seasons. Comparison of the trends in accuracy of the SUB forecasts for the entire period and reduced period reveals that slopes are larger for the latter than for the former in the cool season.
the warm season, the two periods have similar slopes for the MAE statistics, whereas the PEG10 statistics exhibit larger slopes over the entire period than over the reduced period.

When the trends in accuracy of the OBJ and SUB forecasts are compared over the reduced period, it is found that the former possess a stronger rate of improvement than the latter for all combinations of sea-
The time series for the SUB forecasts in Fig. 4 exhibit trends toward greater accuracy that are less erratic than the corresponding time series for the OBJ forecasts (cf. Figs. 3 and 4). Substantial increases in accuracy for these forecasts occurred in the cool season during the period 1972–74 and in the warm season during the period 1967–70. The trend toward increasing accuracy in the SUB forecasts is relatively modest in the 1975–80 period, but the early 1980s witnessed subjective $T_{\text{max}}$ forecasts that were more accurate for both seasons and all lead times than heretofore. Variations in the MAE values associated with these forecasts in the period 1983–85 may be related to the aforementioned modifications in the methods used to verify the temperature forecasts.

2) Minimum Temperature

The national MAE time series for the OBJ and SUB minimum temperature ($T_{\text{min}}$) forecasts are presented...
in Figs. 5 and 6, respectively. Estimated slopes of the regression lines and the corresponding $p$-values for these MAE statistics are contained in Table 3. Examination of these results reveals that statistically significant positive trends in accuracy (i.e., negative slopes) exist for most combinations of forecast type, season and lead time. Two notable and consistent exceptions to this general result are (1) the 60 h forecasts and (2) the OBJ forecasts for the 1970–85 period in the cool season. Significant trends did not occur for any stratifications in the case of the 60 h forecasts, a result that is undoubtedly due in part to the limited length of record available for these forecasts and to the changes in the verification procedure. The lack of significance in the case of the cool season OBJ forecasts for the entire period results from the marked increases in MAE val-

**Fig. 5.** Time series of national mean absolute errors for objective $T_{\text{max}}$ forecasts for four lead times in (a) cool season over period 1970–85 and (b) warm season over period 1970–84. See text for additional details.

**Fig. 6.** Time series of national mean absolute errors for subjective $T_{\text{min}}$ forecasts for four lead times in (a) cool season over period 1967–85 and (b) warm season over period 1967–84. See text for additional details.
### Table 3. As in Table 2 but for minimum temperature.*

<table>
<thead>
<tr>
<th>Lead time (h)</th>
<th>Type of forecast</th>
<th>Period (years)</th>
<th>National Slope</th>
<th>p-value</th>
<th>Eastern Region Slope</th>
<th>p-value</th>
<th>Southern Region Slope</th>
<th>p-value</th>
<th>Central Region Slope</th>
<th>p-value</th>
<th>Western Region Slope</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>24</td>
<td>SUB</td>
<td>1967–85</td>
<td>−0.011</td>
<td>0.1674</td>
<td>−0.005</td>
<td>0.6807</td>
<td>−0.017</td>
<td>0.0219</td>
<td>−0.019</td>
<td>0.1656</td>
<td>−0.020</td>
<td>0.0309</td>
</tr>
<tr>
<td></td>
<td>SUB</td>
<td>1970–83</td>
<td>−0.017</td>
<td>0.2515</td>
<td>−0.016</td>
<td>0.3881</td>
<td>−0.019</td>
<td>0.0667</td>
<td>−0.019</td>
<td>0.4080</td>
<td>−0.036</td>
<td>0.274</td>
</tr>
<tr>
<td></td>
<td>OBJ</td>
<td>1970–83</td>
<td>−0.107</td>
<td>0.0000</td>
<td>−0.119</td>
<td>0.0000</td>
<td>−0.096</td>
<td>0.0000</td>
<td>−0.131</td>
<td>0.0000</td>
<td>−0.102</td>
<td>0.0001</td>
</tr>
<tr>
<td></td>
<td>OBJ</td>
<td>1970–85</td>
<td>−0.048</td>
<td>0.0699</td>
<td>−0.050</td>
<td>0.1022</td>
<td>−0.048</td>
<td>0.0390</td>
<td>−0.062</td>
<td>0.0542</td>
<td>−0.048</td>
<td>0.0753</td>
</tr>
<tr>
<td>36</td>
<td>SUB</td>
<td>1967–85</td>
<td>−0.044</td>
<td>0.0000</td>
<td>−0.041</td>
<td>0.0045</td>
<td>−0.048</td>
<td>0.0000</td>
<td>−0.052</td>
<td>0.0000</td>
<td>−0.047</td>
<td>0.011</td>
</tr>
<tr>
<td></td>
<td>SUB</td>
<td>1970–83</td>
<td>−0.068</td>
<td>0.0000</td>
<td>−0.078</td>
<td>0.0009</td>
<td>−0.068</td>
<td>0.0000</td>
<td>−0.068</td>
<td>0.0006</td>
<td>−0.066</td>
<td>0.0025</td>
</tr>
<tr>
<td></td>
<td>OBJ</td>
<td>1970–83</td>
<td>−0.120</td>
<td>0.0000</td>
<td>−0.129</td>
<td>0.0003</td>
<td>−0.107</td>
<td>0.0001</td>
<td>−0.143</td>
<td>0.0008</td>
<td>−0.115</td>
<td>0.0002</td>
</tr>
<tr>
<td></td>
<td>OBJ</td>
<td>1970–85</td>
<td>−0.057</td>
<td>0.0664</td>
<td>−0.059</td>
<td>0.0947</td>
<td>−0.058</td>
<td>0.0274</td>
<td>−0.076</td>
<td>0.0435</td>
<td>−0.055</td>
<td>0.0718</td>
</tr>
<tr>
<td>48</td>
<td>SUB</td>
<td>1967–85</td>
<td>−0.054</td>
<td>0.0000</td>
<td>−0.049</td>
<td>0.0112</td>
<td>−0.065</td>
<td>0.0000</td>
<td>−0.063</td>
<td>0.0000</td>
<td>−0.052</td>
<td>0.0002</td>
</tr>
<tr>
<td></td>
<td>SUB</td>
<td>1970–83</td>
<td>−0.097</td>
<td>0.0000</td>
<td>−0.098</td>
<td>0.0012</td>
<td>−0.088</td>
<td>0.0000</td>
<td>−0.115</td>
<td>0.0021</td>
<td>−0.101</td>
<td>0.0014</td>
</tr>
<tr>
<td></td>
<td>OBJ</td>
<td>1970–83</td>
<td>−0.039</td>
<td>0.1633</td>
<td>−0.031</td>
<td>0.3225</td>
<td>−0.044</td>
<td>0.0386</td>
<td>−0.049</td>
<td>0.1595</td>
<td>−0.052</td>
<td>0.0724</td>
</tr>
<tr>
<td>60</td>
<td>SUB</td>
<td>1977–85</td>
<td>0.005</td>
<td>0.8822</td>
<td>−0.038</td>
<td>0.4396</td>
<td>−0.031</td>
<td>0.4316</td>
<td>−0.006</td>
<td>0.8424</td>
<td>0.023</td>
<td>0.6446</td>
</tr>
<tr>
<td></td>
<td>SUB</td>
<td>1977–83</td>
<td>−0.054</td>
<td>0.2657</td>
<td>−0.034</td>
<td>0.6629</td>
<td>−0.117</td>
<td>0.0408</td>
<td>−0.034</td>
<td>0.5354</td>
<td>−0.066</td>
<td>0.2787</td>
</tr>
<tr>
<td></td>
<td>OBJ</td>
<td>1977–83</td>
<td>−0.091</td>
<td>0.2678</td>
<td>−0.106</td>
<td>0.2844</td>
<td>−0.100</td>
<td>0.2209</td>
<td>−0.077</td>
<td>0.5360</td>
<td>−0.106</td>
<td>0.0499</td>
</tr>
<tr>
<td></td>
<td>OBJ</td>
<td>1977–85</td>
<td>0.048</td>
<td>0.4663</td>
<td>0.064</td>
<td>0.4249</td>
<td>0.037</td>
<td>0.5661</td>
<td>0.024</td>
<td>0.7455</td>
<td>0.004</td>
<td>0.9631</td>
</tr>
</tbody>
</table>

*(a) Cool season

*(b) Warm season

*The period for the cool season extends from 1967–68 or 1970–71 (1977–78 for the 60 h forecasts) through 1982–83 or 1984–85, whereas the period for the warm season extends from 1967 or 1970 (1977 for the 60 h forecasts) through 1983 or 1984.*

ues in 1983–84 (see Fig. 5a); in this regard, note that the trends in accuracy for these forecasts for the reduced period are statistically significant.

Differences in the trends in accuracy of the reduced-period OBJ forecasts across lead times are relatively small (see Table 3). However, these trends are stronger in the cool season than in the warm season, and the PEG10 statistics reveal a tendency for the trends to increase as lead time increases in the warm season. On the other hand, the trends in accuracy of the SUB forecasts increase as lead time increases in both seasons, with the slopes in the cool season generally exceeding the slopes in the warm season. Comparison of the slopes of the regression lines for the OBJ and SUB forecasts over the reduced period reveals that the former are greater than the latter for all combinations of season and lead time. These differences in rates of improvement are particularly large for the 24 and 36 h lead times.

The time series of mean absolute errors for the $T_{min}$ forecasts exhibit some of the same characteristics as those identified previously in the case of the $T_{max}$ forecasts. With regard to the OBJ forecasts (Fig. 5), substantial increases in accuracy occurred during the first half of the 1970s in both the cool and warm seasons. The second half of the 1970s and the early 1980s witnessed much more modest improvements in the objective $T_{min}$ forecasts.

Recent marked increases in the MAE values associated with these forecasts are believed to be due primarily to the change made in October 1983 in the procedure used to verify the forecasts from calendar-day...
minimum temperatures to 12 h minimum temperatures. This factor evidently led to a particularly sharp increase in the MAE values for the objective $T_{\text{min}}$ forecasts in the cool season. In addition, unusual meteorological conditions in 1982–83 over substantial areas of the United States may have contributed to these larger MAE values.

With regard to the SUB forecasts (Fig. 6), the time series for the cool season exhibit much greater variability than the time series for the warm season. As in the case of the subjective $T_{\text{max}}$ forecasts, these $T_{\text{min}}$ forecasts reveal a systematic increase in MAE values in the 1983–84 cool season and in the 1984 warm season. Another change in the verification procedure prior to the 1984–85 cool season may have contributed to the small but consistent decrease in these MAE values.

3) COMPARISON OF $T_{\text{MAX}}$ AND $T_{\text{MIN}}$ FORECASTS

Comparison of the national trends in accuracy for the $T_{\text{max}}$ and $T_{\text{min}}$ forecasts (cf. Tables 2 and 3) indicates that the differences in the corresponding slopes of the regression lines are not large in most cases. The PEG10 statistics in the cool season represent an exception to this general statement; in this case, the trends in accuracy for the $T_{\text{max}}$ forecasts are markedly stronger than the respective trends for the $T_{\text{min}}$ forecasts. In addition, in the cool season for the MAE statistics and in the warm season for the PEG10 statistics, a tendency exists for the rate of improvement in the $T_{\text{max}}$ forecasts—objective and subjective—to exceed the rate of improvement in the $T_{\text{min}}$ forecasts. No noticeable or consistent differences in slopes are evident as a function of either forecast type (OBJ/SUB) or lead time.

b. Regional results

1) MAXIMUM TEMPERATURE

Slopes and $p$-values associated with the MAE statistics for the $T_{\text{max}}$ time series for the four NWS regions are included in Table 2. A great majority (84 out of 96) of these slopes are statistically significant. All of the nonsignificant slopes involve the Southern Region, with 9 of these 12 slopes associated with the warm season.

As in the case of the national results, the regional results for the OBJ forecasts exhibit stronger trends in accuracy in the cool season than in the warm season, and no systematic patterns in these slopes exist as a function of lead time. For the SUB forecasts (18-year period), the slopes are also generally larger in the cool season than in the warm season (the Eastern Region is an exception). Moreover, the trends in accuracy of the SUB forecasts increase as lead time increases for all regions in both seasons. When the slopes of the OBJ and SUB forecasts are compared over the reduced period, it is evident that the rate of increase in accuracy is greater for the former than for the latter, with the largest differences in the rates of improvement associated with the 24 h lead time. Comparison among regions reveals that these trends are generally strongest in the Central Region and weakest in the Southern Region.

2) MINIMUM TEMPERATURE

The regional slopes and $p$-values corresponding to the MAE statistics for the $T_{\text{min}}$ time series are included in Table 3. Trends in accuracy are positive for almost all combinations of region, forecast type, season, and lead time; however, many of the slopes are not statistically significant. For example, the regional trends in the accuracy of the 60 h forecasts are generally not significant. On the other hand, the trends in accuracy are statistically significant for almost all (other) lead times in the Southern and Western regions in the cool season and in the Eastern, Central and Western regions in the warm season.

Trends in the accuracy of the $T_{\text{min}}$ forecasts are generally stronger in the cool season than in the warm season. These trends increase as lead time increases for the SUB forecasts in both seasons, whereas this result holds for the OBJ forecasts in the warm season only. Comparison of the slopes for the OBJ and SUB forecasts (reduced period) reveals that the rates of increase in accuracy are greater for the former than for the latter for almost all combinations of season and lead time, with the largest differences in these rates of improvement associated with the 24 hour forecasts. When the trends in accuracy of the $T_{\text{min}}$ forecasts in the various regions are compared, it is evident that these trends are usually strongest in the Central Region in the cool season and in the Eastern and Western Regions in the warm season. The weakest trends generally occur in the Southern Region.

3) COMPARISON OF $T_{\text{MAX}}$ AND $T_{\text{MIN}}$ FORECASTS

As in the case of the national results, the differences among regional trends in the accuracy of the $T_{\text{max}}$ and $T_{\text{min}}$ forecasts (cf. Tables 2 and 3) are generally not very large. The slopes for the subjective $T_{\text{max}}$ forecasts exceed the slopes for the subjective $T_{\text{min}}$ forecasts in the Eastern Region in the warm season, in the Central Region in both seasons, and in the Western Region in the cool season. For the objective forecasts, the $T_{\text{max}}$ slopes exceed the $T_{\text{min}}$ slopes in the Western Region in both seasons, whereas the opposite relationship holds in the Eastern Region in the warm season and in the Southern Region in both seasons.

6. Discussion

The results of this study indicate that, on a national basis, statistically significant increases in forecast quality have occurred over the last two decades for precipitation probability and maximum temperature forecasts for all combinations of forecast type, season, lead time, and verification measure considered here. Similar
results also hold for most of these stratifications in the case of minimum temperature forecasts. Moreover, the regional results are similar to the national results in that the trends are statistically significant for most of the relevant combinations, with the fewest exceptions in the case of PoP forecasts and the most exceptions in the case of $T_{\text{min}}$ forecasts. These increases in forecast quality are undoubtedly due, in part, to improvements in the numerical and numerical-statistical models, as well as to “refinements” in other components of the weather forecasting system. However, it is difficult to unambiguously attribute specific increases to particular improvements or refinements.

Comparison of the results across seasons and lead times reveals that the trends in quality (i) are stronger in the cool season than in the warm season and (ii) increase as lead time increases in the case of subjective forecasts. Of course, improvements in numerical models would be expected to be relatively more effective in contributing to increases in forecast quality in the cool season and at longer lead times. In this regard, it is somewhat surprising that the objective forecasts do not exhibit a similar pattern of results as a function of lead time.

When trends in the quality of the objective and subjective forecasts are compared, the former are generally found to exceed the latter. This result is evidently due primarily to the relatively low quality of the objective forecasts when they were first produced in the early 1970s. As a result of stronger positive trends in quality, the OBJ forecasts have appreciably “narrowed the gap” between their performance and the performance of the SUB forecasts over the subsequent period.

No consistent patterns appear in the regional results. However, it is interesting to note that the Western Region seldom exhibits the strongest trends. Of course, improvements in numerical models would be expected to be relatively less effective in producing higher quality forecasts in a region significantly influenced by important topographic features.

With the exception of the overall positive trends in quality of the PoP, $T_{\text{max}}$ and $T_{\text{min}}$ forecasts, the most notable feature of the results is the increase in MAE values associated with the temperature forecasts in the 1983–84 cool season and the 1984 warm season (see Figs. 3–6). We believe that these increases in MAE values are due in large measure to a change in the method of verifying the temperature forecasts that was implemented in October 1983. In this regard, since changes (or differences) in the definitions of the forecasts and/or observations could affect the verification statistics, it is important to keep in mind the relationships between the definitions of these two quantities. Here, the OBJ forecasts relate to calendar-day maximum and minimum temperatures, whereas the SUB forecasts relate to daytime maxima and nighttime minima. Thus, the OBJ forecasts are (completely) consistent with the observations only during the period from the 1975 warm season through the 1983 warm season (period B) and the SUB forecasts are consistent with the observations only during the 1984–85 cool season (period D). This lack of correspondence between the definitions of the forecast and observed temperatures, together with the modifications in the verification procedure, obviously complicate the interpretation of the results. Specifically, it is difficult to distinguish between real increases or decreases in accuracy and changes in MAE values associated with these inconsistencies and modifications.

With these considerations in mind, note that the accuracy of the OBJ forecasts generally tended to increase between the 1974–75 and 1975–76 cool seasons and between the 1974 and 1975 warm seasons. Since these increases in accuracy coincide with a change from an inconsistent to a consistent verification procedure, they may be partly spurious (in the sense that the accuracy of the OBJ forecasts in period A may have been underestimated). The increases in MAE values associated with the transition from period B to period C in October 1983 have already been mentioned. Since this transition necessarily implied that the OBJ forecasts were no longer consistent with the verifying observations, the MAE values would be expected to increase. Moreover, differences between calendar-day and 12 h temperatures are generally larger for minima than for maxima and for the cool season than for the warm season, and the results are consistent with these relationships. Evidently, the cool season MAE values for the OBJ forecasts were not strongly affected by the transition from period C to period D (both of which are associated with “inconsistent” verification procedures). The difficulties encountered in interpreting the results for the OBJ forecasts will be greatly reduced beginning with the 1985–86 cool season when a new MOS system will be implemented. This system will provide forecasts of daytime maxima and nighttime minima and will be consistent with both the SUB forecasts and the verification procedure.

In the case of the SUB forecasts, a tendency can be noted for the MAE values to decrease in the transition between periods A and B and to increase in the transition between periods B and C. Recall that 12 h temperature observations were used to verify the forecasts in periods A and C, that calendar-day temperature observations were used to verify the forecasts in period B, and that neither of these procedures is consistent with the SUB forecasts. The results suggest that the errors introduced by using 12 h observed temperatures are larger and/or more frequent than the errors associated with using calendar-day maxima and minima. Since the set of NWS offices over which the verification statistics were calculated was modified in October 1983 in connection with the implementation of a new verification system (see Carter et al., 1985), this change also could have contributed to the decrease in MAE values at that time. Finally, it is encouraging to see a decrease in MAE values in the 1984–85 cool season (period D), since the SUB forecasts and the verification...
procedure first became consistent at that time. However, these latest results suggest that the accuracy of the temperature forecasts in the early 1980s may have been overestimated. Confirmation of this hypothesis must await additional results from the new verification system and/or more detailed analyses of the verification data.

As noted in section 1, we have also conducted a preliminary study of national trends in the quality of NWS cloud amount and wind speed forecasts over the period from the 1975–76 cool season through the 1983 warm season. Verification measures employed in these analyses were percent correct and Heidke skill score in the case of cloud amount; and mean absolute error, percent correct and Heidke skill score in the case of wind speed. The results of this study reveal that statistically significant trends in the quality of these forecasts exist for relatively few combinations of forecast type, season, lead time and verification measure. Of course, the fact that only eight years of such verification data are available undoubtedly reduces the likelihood that any positive trends in performance would be judged to be statistically significant.

Since a primary motivation for this study was to update and extend the studies of trends in the quality of subjective PoP forecasts undertaken by Ramage (1982) and Glahn (1985), it is of some interest to compare the results presented here with their results. In brief, our results correspond quite closely with the results reported by Glahn and strongly support his conclusion that “strong evidence (exists) that probability of precipitation forecasts have improved over the 15-year period from 1967 to 1982.” With regard to the possible reasons for the differences between their results, we generally support the arguments set forth by Glahn. For example, we believe that the use of percent correct as a measure of the quality of PoP forecasts is totally inappropriate and tends to mask any changes in the quality of probabilistic forecasts. A skill score based on a probabilistic measure of performance, such as the Brier score, is a much more suitable verification measure in this context. In a different vein, Glahn and Ramage reported trends in percent correct (with the effects of the relative frequency of precipitation taken into account) that were and were not significant, respectively. This difference may be due simply to the respective “lengths” of the time series—15 years for Glahn and 12 years for Ramage. Such a hypothesis is supported by the fact that the trends in the skill of the PoP forecasts reported here (and based on 18 years of data) are generally stronger and/or more highly significant than the trends reported by Glahn.

7. Conclusion

As a result of positive trends in performance over the last 15–20 years, the NWS precipitation probability, maximum temperature and minimum temperature forecasts are now of considerably higher quality than they were at the beginning of the period. For example, national skill scores for the 12–24 h subjective PoP forecasts in the cool (warm) season have increased from approximately 38% (24%) in 1967–68 (1967) to approximately 48% (34%) in 1984–85 (1984). Similar increases in skill have occurred for the other lead times and for the subjective PoP forecasts at all three lead times. Moreover, the 24–36 h (36–48 h) PoP forecasts are now as skillful as the 12–24 h (24–36 h) PoP forecasts were in the late 1960s.

Similar improvements over this period can be found in the $T_{\text{max}}$ and $T_{\text{min}}$ forecasts, although the magnitude of these improvements is more difficult to estimate because of the changes in the procedure used to verify the forecasts. Ignoring the recent effects of such changes, the mean absolute errors of the 24 h $T_{\text{max}}$ forecasts in the cool (warm) season have decreased from approximately 3.8°F (3.3°F) in 1967–68 (1967) to approximately 3.2°F (2.8°F) in 1984–85 (1984). Corresponding results for the $T_{\text{min}}$ forecasts in the cool (warm) season indicate decreases in MAEs from 3.9°F (3.1°F) in 1967–68 (1967) to 3.7°F (2.8°F) in 1984–85 (1984). Results for other lead times—and for the objective $T_{\text{max}}$ and $T_{\text{min}}$ forecasts—generally reveal even greater increases in accuracy over the respective periods. When the quality of the temperature forecasts for the various lead times is compared at the beginning and end of the periods, it is clear that the current 36 and 48 h forecasts are as accurate—or even more accurate—than the 24 and 36 h forecasts, respectively, were in the late 1960s (SUB) or the early 1970s (OBJ). Thus, substantial, as well as statistically significant, increases in the quality of the PoP, $T_{\text{max}}$ and $T_{\text{min}}$ forecasts have occurred over the last two decades.

With regard to future work in this area, many important issues concerning trends in the quality of NWS forecasts remain to be addressed. These issues include (i) identification of the scientific, technological and/or operational developments that account for, or contribute to, specific increases (or decreases) in forecast quality; (ii) determination of the extent to which the results presented in this paper are representative of trends in performance in other regions or at specific locations, for shorter or longer lead times, and for other weather elements; (iii) investigation of trends in quality using more appropriate verification measures (e.g., a skill score for temperature forecasts that accounts for the variability in observed temperatures) and/or more suitable “modeling” procedures than simple linear regression (e.g., an approach that would permit the detection of trends in relatively short time series involving “noisy” data); and (iv) evaluation of the actual or potential impact of such increases (or decreases) in forecast quality on users of the forecasts and on the economic value of such information. Studies devoted to these and other related issues should lead to more accurate and realistic assessments of recent trends in the quality of weather forecasts and of the operational significance of such trends.
Acknowledgments. The authors would like to thank Gary M. Carter and Harry R. Glahn for their assistance in acquiring and interpreting some of the verification data analyzed in this paper. This work was supported in part by the National Science Foundation (Division of Atmospheric Sciences) under Grant ATM-8507495.

REFERENCES


