5.8 OBJECTIVE INTERPRETATION OF NUMERICAL WEATHER PREDICTION MODEL OUTPUT – A PERSPECTIVE BASED ON VERIFICATION OF TEMPERATURE AND PRECIPITATION GUIDANCE

J. Paul Dallavalle* and Valery J. Dagostaro
Meteorological Development Laboratory
Office of Science and Technology
National Weather Service/NOAA
Silver Spring, Maryland

1. INTRODUCTION

In the late 1960's and early 1970's, the National Weather Service (NWS) began implementing guidance products that objectively interpreted the output of numerical weather prediction (NWP) models. Initially, these products were for weather elements such as maximum (max) and minimum (min) temperature or the probability of precipitation (PoP) during 12-h periods. The first interpretive products were based on the "perfect prog" (Klein and Lewis 1970) statistical technique. By 1973, the Model Output Statistics (MOS) approach (Glahn and Lowry 1972) had superceded the perfect prog method, and the number of interpretive guidance products increased rapidly. Since then, the MOS approach has been used by the NWS to generate guidance products from the Primitive Equation, Limited-area Fine-mesh, Nested Grid, Eta, and Global Forecast System models.

Much of the improvement in NWP and in the statistical interpretation system can be tracked by the verification of the weather element guidance. Since 1966, the NWS has maintained a national verification program in which selected NWS forecast products are routinely verified as an indicator of the quality of public and aviation weather services provided by the NWS. Forecasts produced by each NWS forecast office are evaluated for one or more cities in the office's area of responsibility, and the skill of the human forecast is compared to that provided by the guidance for the same weather element. Thus, verifications of the max/min temperature and PoP provide one look at the evolution of skill in the NWP models and the NWS public weather forecasts from 1966 to the present.

Earlier studies have examined trends in the skill of the NWS forecasts. Zurndorfer et al. (1979) looked at verifications of guidance and local public weather forecasts for the period of 1970 through 1977 and found improvements in both the PoP and max/min temperature forecasts. Charba and Klein (1980) found a substantial increase in skill over a 10-year period when verifying the PoPs of the local NWS offices for projections of 24-36 and 36-48 hours in advance. Ramage (1982) contradicted these results and claimed that, except in certain regions of the U.S. in winter, any increase in accuracy of the PoPs from 1966 to 1978 was negligible. Glahn (1985), however, found significant improvement in the local NWS PoPs over the 15-year period from 1967 to 1982. Carter and Polger (1986) showed that the national skill scores for both the local NWS and the guidance PoP had significantly increased since 1966 for all seasons, both forecast cycles, and all projections examined. Murphy and Sabin (1986) confirmed the

**Corresponding author address:* J. Paul Dallavalle, 1325 East-West Highway, Station 11306, Silver Spring, MD 20910-3283; e-mail:paul.dallavalle@noaa.gov

Carter and Polger study and found that on the national level, the NWS PoP and max/min temperature forecasts were of significantly higher quality by the end of the approximately 15-20 year period studied than in 1966.

In this paper, we examine the skill of the official NWS max/min and PoP forecasts compared to the guidance for forecast periods out to approximately 60 hours in advance. After a brief description of the evolution of the NWS verification program for temperature and PoP, we describe changes made to the guidance system during the period of record, summarize the methodology used to verify the guidance and local forecasts, and show a number of verification time series beginning in the late 1960's. As expected, the skill of the NWS max/min temperature and PoP forecasts, as well as that of the guidance, has increased significantly since the late 1960's and early 1970's. While the increase in skill of the day 1 forecast has leveled off during the last decade, the skill of the day 2 forecast provided to the public is now comparable to that of the day 1 forecast provided 20 years ago.

2. PUBLIC WEATHER VERIFICATION

The National Weather Service program to verify forecasts issued to the public is described in Carter and Polger (1986) and Dagostaro et al. (1989). Established in 1966, the verification program was designed to evaluate the quality of the max/min temperature and PoP forecasts issued by the local NWS forecast offices and to compare those forecasts with the guidance provided to the offices. In 1966, the guidance was provided by forecasters at the National Meteorological Center (NMC), now the National Centers for Environmental Prediction (NCEP). Within several years, objective guidance obtained by applying statistical methods to NWP model output replaced the subjective guidance.

Because the NWP models were, at first, run only from 0000 and 1200 UTC initial conditions, and because the forecasts issued to the public were based on these runs, only forecasts from two issuance times were verified. The public forecasts were disseminated at approximately 0400 and 2000 local time (LT) and contained information for three forecast periods. From the 0400 LT issuance (0000 UTC model cycle), the public weather forecasts were valid for today, tonight, and tomorrow. For PoP, these definitions corresponded to forecasts valid 12-24, 24-36, and 36-48 hours after 0000 UTC. For the max/min temperature, the forecasts were valid for today's max, tonight's min, and tomorrow's max, which nominally were valid approximately 24, 36, and 48 hours after 0000 UTC. Similarly, from the 2000 LT issuance (1200 UTC cycle), the public weather forecasts were available for tonight, tomorrow, and tomorrow night. For the PoPs, these projections were defined exactly as for the 0000 UTC cycle, namely, valid for 12-h periods ending 24, 36, and 48 hours after initial model time. For the max/min temperature, the forecasts were for tonight's min, tomorrow's max, and tomorrow night's min, and nominally were valid approximately 24, 36, and 48 hours after 1200 UTC. A fourth period for the max/min temperatures was added in October 1975 so that an additional min (max) temperature became available for the 0000 (1200) UTC cycle. The fourth period PoP was not available to the verification system until summer 2002.

During the verification period, four different approaches to data collection were used. From 1966 until March 1974, extensive human labor, either at the forecast office or at NWS Headquarters, was required to record the local forecasts and verifying observations on a form and transfer these values to punched cards. The subjective guidance was also obtained from a form and punched cards; when

the guidance became objective, the appropriate guidance was extracted from the NWS central computer system. In March 1974, a second phase was initiated in which more extensive use was made of machine-based data extraction methods. For 15 months, the local forecasts and verifying observations were entered on mark sense cards which could be read directly by a computer. By July 1975, data collection became totally automated when the NWS implemented software to collect the local max/min and PoP forecasts from the Coded City Forecast bulletin and the verifying observations from the Selected Cities Summary prepared by the National Climatic Data Center. The guidance continued to be taken from files on the NWS central computer system.

A third era of data collection began after the NWS introduced the Automation of Field Operations and Services (AFOS) system into its field operations. In October 1983, the NWS implemented the AFOS-era verification (AEV) program. This program, run locally at each NWS forecast office, created a verification database in which the appropriate local forecasts for the max/min temperature and PoP, the verifying observations, and the objective guidance were stored. As before, the local forecasts were extracted from the Coded City Forecast bulletin. The verifying observations were obtained from a local database of hourly and synoptic observations. The guidance was obtained from the appropriate MOS alphanumeric bulletin. The AEV program was run twice daily to pick up the appropriate forecast information and to collate forecasts with the verifying observations. The local forecaster had an editing capability so as to modify any local forecasts or observations which were stored erroneously in the database. Within 5 to 7 days of the generation of the original forecast, an alphanumeric bulletin containing the collated local forecasts, objective guidance, and verifying observations was transmitted back to the NWS central computer system. Subsequently, staff from the Techniques Development Laboratory (now the Meteorological Development Laboratory or MDL) processed the data, did quality-control, and provided periodic verifications and analyses. Ruth and Alex (1987) described the AEV program; Dagostaro (1985) described the system used to process the AEV data.

In January 2000, the AEV program was transferred to the NWS Advanced Weather Interactive Processing System (AWIPS). The AWIPS verification program (AVP) was analogous to the AEV software. Further changes in the NWS methodology of generating forecast products, specifically, the introduction of the Interactive Forecast Preparation System (IFPS) and the generation of local forecasts on grids, began to eliminate the local use of the AVP software. This fourth era of data collection was starting to draw to a close by the spring of 2004.

3. GUIDANCE AVAILABILITY

When the verification program began in 1966, NMC forecasters provided subjective PoP and max/min temperature guidance to the local forecast offices. Until April 1969, for the 12-24 h PoP, the guidance was categorical, that is, the PoP was equal to 0 or 100%, for no rain or rain, respectively. The PoPs for the other forecast projections were probabilities that ranged from 0 to 100%. From April 1969 to the present, the PoP guidance for all projections presented a range of probabilities. In January 1972, the subjective PoP guidance produced by human forecasters was replaced by a MOS PoP system (Lowry and Glahn 1976) developed from the Primitive Equation (PE) and trajectory models. Since that time, the official PoP guidance issued to the local forecasters has been based on the MOS approach applied to one of the NWP models.

Similarly, for the max/min temperature, subjective guidance was issued by NMC forecasters until 1970. In April 1970, the NWS implemented objective max/min temperature guidance based on the perfect prog approach applied to the Primitive Equation (PE) model. In August 1973, the perfect prog guidance was replaced by a MOS max/min temperature guidance system developed from the PE and trajectory models. Like the PoP guidance, the official max/min temperature guidance since that time has been generated by the MOS approach applied to one of the NWP models.

As noted, the NWS was using the MOS approach extensively by 1973. In subsequent years, different packages of MOS guidance were often available, depending on the availability of NWP models. For many of the last 30 years, the forecasters have had access to two or more guidance packages when producing the public forecast. For instance, in 2004, forecasters can use for the first four forecast periods max/min temperature and PoP guidance that is based on the Nested Grid Model (NGM), the Global Forecast System (GFS) model, or the Eta model. Obviously, this plethora of guidance complicates comparative verification. For the NWS verifications presented in this paper, however, only the "official" guidance is used. The choice of the official MOS guidance product was dictated by the software collecting the MOS guidance, and not by the newest or most accurate guidance package available. As noted above, the official MOS guidance was initially PE-based. In April 1980, the MOS guidance based on the Limited-area Fine-mesh model (LFM) became the official guidance package. The NGM-based guidance replaced the LFM-based guidance in June 1993 as the official package. Finally, during the summer of 2002 (the precise date varied from station to station because of the difficulty of implementing software in a distributed processing system like AWIPS), the GFS-based MOS guidance became the official guidance for the purpose of verification.

One last change in the characteristics of the guidance must be noted. Although forecasters were generating public weather temperature forecasts that referred to daytime and nighttime periods, the original MOS guidance for max/min temperature was valid for a calendar day max and min. This valid period was dictated by the availability of the observations used in the MDL developmental system. In November 1985, MDL implemented a new MOS system that predicted daytime max and nighttime min temperatures (Erickson and Dallavalle 1986). As will be seen in the verification time series, this particular change resulted in substantial improvement in the accuracy of the guidance.

4. VERIFICATION METHODOLOGY

In doing this latest verification study, we've verified the local forecasts and the guidance since April 1966. A matched sample of local forecasts and guidance was used. Verifications were done for two seasonal stratifications, namely, for warm (April 1 – September 30), and cool (October 1 – March 31) seasons. We evaluated the local forecasts and guidance for both cycles (0000 and 1200 UTC) and all projections. In the results shown here, the verification scores for the PoPs for a specific projection, for example, 12-24 hours after 0000 or 1200 UTC, are combined to produce one set of scores. Though verification results are highly variable between the two seasons, we did not find that the PoP scores varied as much between the daytime and nighttime period and so combined them. In contrast, the verifications of the max and min temperatures vary substantially and so we did not combine the scores for those two elements.

The PoP forecast represents the probability of measurable precipitation (defined as 0.01 inches or more) occurring within a specific 12-h period ending at either 0000 or 1200 UTC. Thus, the verifying observation is available by summing two 6-h precipitation amounts. In contrast, the max (min) temperature forecast represents the high (low) temperature occurring during the local day (night). Observations of the max/min temperature taken at standard observing sites coincide with a UTC interval or with a local calendar day. Thus, no direct observation is made of the daytime max or nighttime min. Prior to the 1984-85 cool season, the daytime max was estimated by using the maximum temperature reported for the 1200-0000 UTC period. The nighttime min was estimated by the 0000-1200 UTC report of the minimum for stations in the eastern and central U.S., and by the 0000-1800 UTC report of the minimum for stations in the western part of the U.S. Starting with the 1984-85 cool season, an algorithm was implemented in the AEV system to estimate the daytime max (defined as 7 a.m. - 7 p.m. Local Standard Time) and the nighttime min (defined as 7 p.m. - 8 a.m. Local Standard Time). Enhancements to this algorithm were made in subsequent years as observing standards were modified.

During the period from 1966 to 2004, the number of stations available for verification varied as a function of the NWS organizational structure and the process being used to collect the data. To establish a relatively constant set of sites, we checked station availability during three periods of time, namely, the pre-AEV era (1966-1983), the AEV era (1983-1999), and the AVP era (2000-2004). To be included in the long-term verifications, a station had to be available during all three periods, and we had to have approximately 50% or more of all possible data from each of the three periods. Under these constraints, forecasts for 80 stations in the contiguous U.S. (Fig.1) were verified. Results are shown here for all stations combined.

A number of verification measures were computed. For the max/min temperature, we calculated mean absolute error, bias or mean algebraic error, root mean square error, and the percent improvement of the forecasts compared to a forecast based on the normal max/min temperature. We used the 1961-90 normals obtained from the National Climatic Data Center as our climatic standard. Only the mean absolute error statistics are discussed here.

For PoP, the Brier score (Brier 1950) and the relative improvement over the Brier score obtained by using the appropriate monthly climatic relative frequency as the forecast were calculated. Note that the Brier score is equal to the mean square error for a probabilistic forecast and a verifying observation of either 0 or 1 for no precipitation or precipitation, respectively. The climatic relative frequencies used for the percent improvement score were based on precipitation records for the period of 1972 through 1985 (Jensenius and Erickson 1987). In addition, we calculated the reliability for the PoP which measured the correspondence between a particular probability forecast and the relative frequency of the event.

5. RESULTS

Figures 2 and 3 show the mean absolute error of the local and guidance forecasts for today's (24 hr) and tomorrow's (48 hr) max temperature for the warm and cool seasons, respectively. As discussed previously, these forecasts were generated during the 0000 UTC cycle. Several features can be noted. First, the cool season errors are greater than the warm season errors for the same projection. Second, the cool season errors seem to exhibit

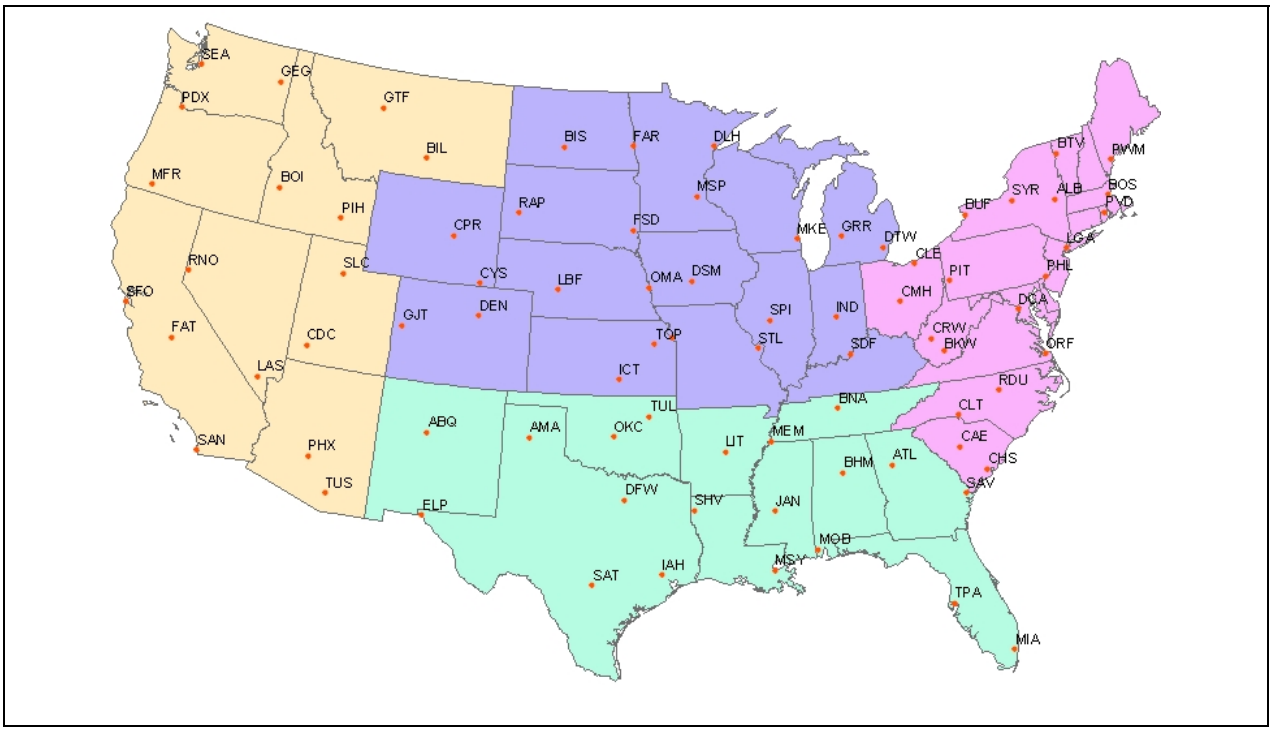

Figure 1. Sites for which local and guidance forecasts were verified. Shading denotes the NWS administrative regions.

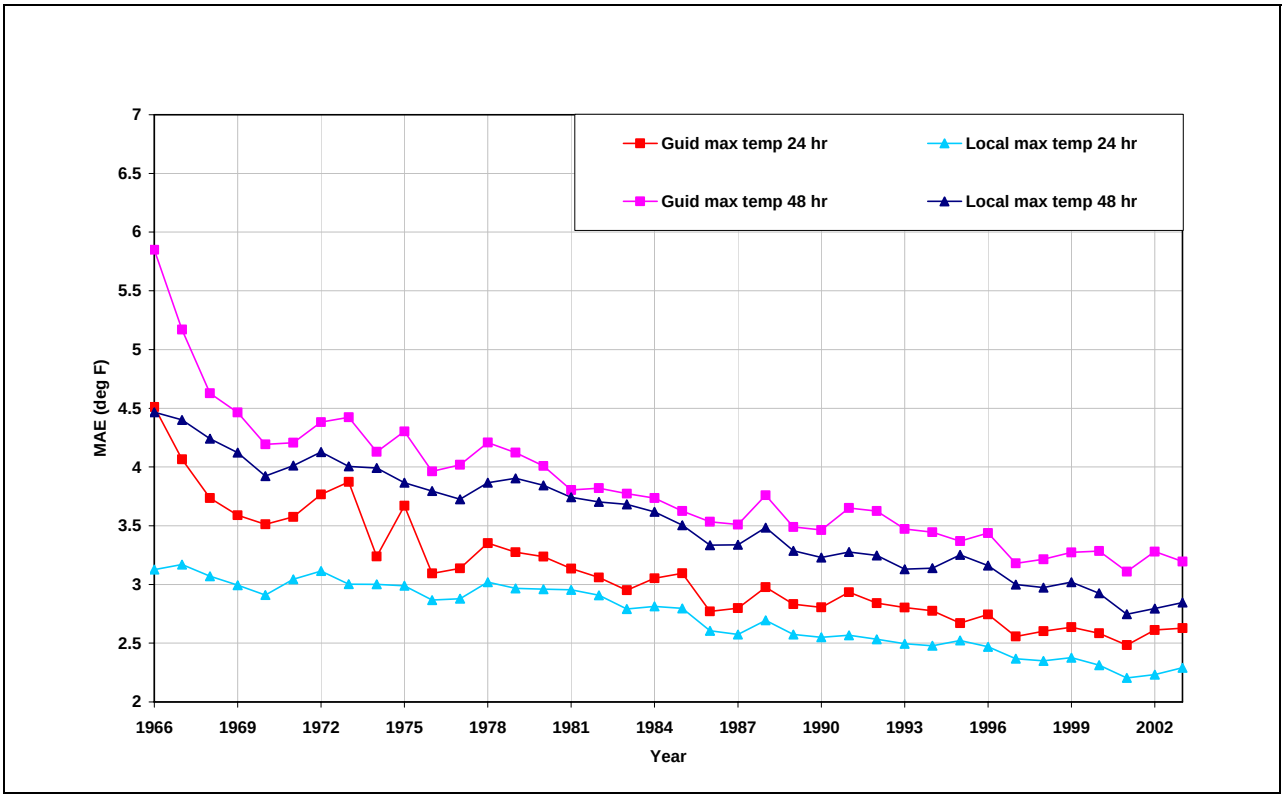

Figure 2. Mean absolute error for warm season local and guidance forecasts of today's (24-hr) and tomorrow's (48-hr) max temperature generated during the 0000 UTC cycle.

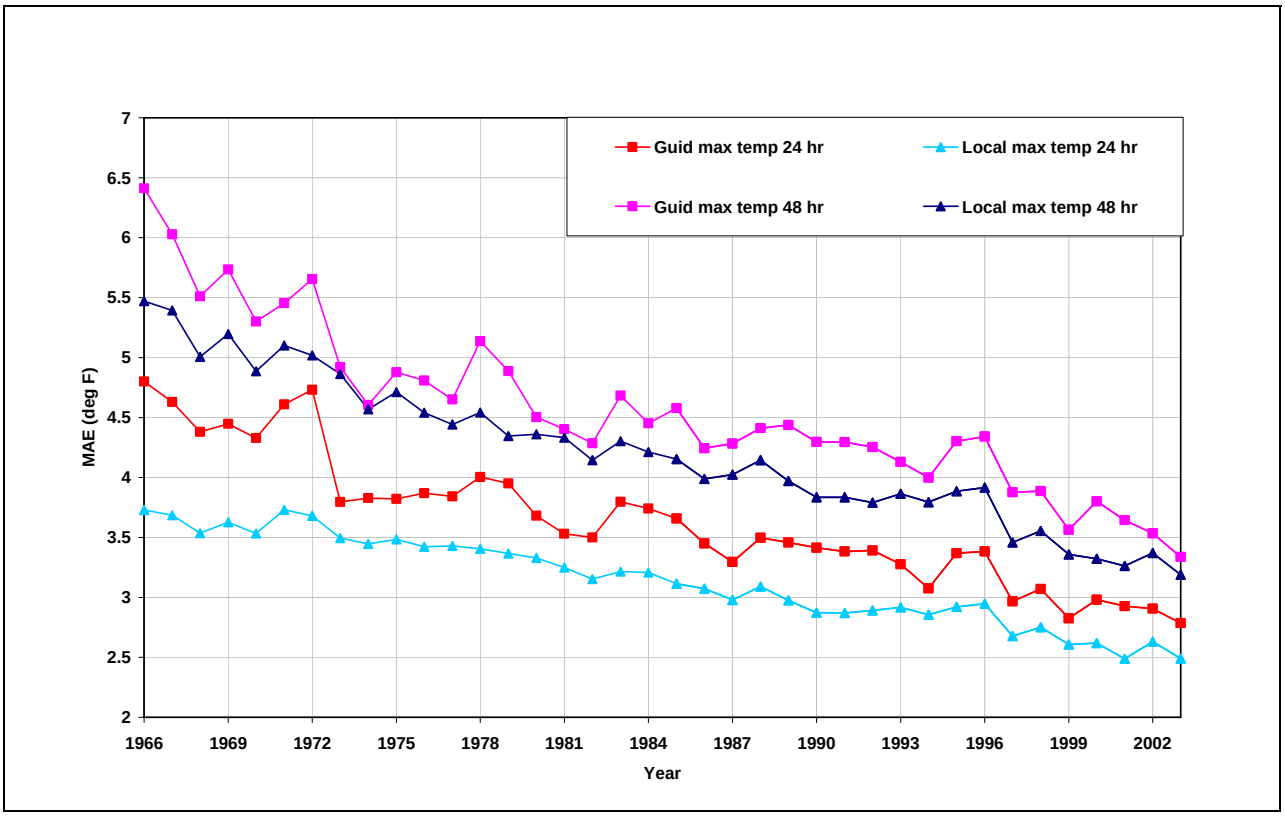

Figure 3. Same as Fig. 2, except for cool season verifications.

more variability from one season to another than the warm season errors. Third, local forecasters are consistently more accurate than the guidance, and this improvement is relatively consistent in the warm season. In the cool season, the variability in the improvement of the local forecast relative to the guidance is somewhat greater. Fourth, the general improvement in both the local forecasts and the guidance over the period of record is evident. For instance, the accuracy of the local forecast for tomorrow's max is now as accurate as the forecast for today's max was 20 years ago. The guidance during the cool season for tomorrow's max is now as accurate as the guidance for today's max was 10 years ago. The downward trend in error is greater in the cool season. Finally, the improvement in the guidance due to the implementation of the MOS approach (August 1973) is quite obvious, particularly during the cool season.

Figures 4 and 5 show for the warm and cool seasons, respectively, the verification time series for tomorrow's (36-hr) and the day after tomorrow's (60-hr) max temperature forecasts generated during the 1200 UTC cycle. The results are similar to those of Figs. 2 and 3. Note the striking improvement in the accuracy of the 60-h max temperature guidance for the 2002-03 and 2003-04 seasons coincident with the use of the GFS guidance as the "official" product.

Mean absolute errors for tonight's (24 hr) and tomorrow's (48 hr) min temperature forecasts generated during the 1200 UTC cycle are shown in Figs. 6 and 7, respectively, for the warm and cool seasons. As with the verifications for the max temperature, several features are notable. First, the improvement in the quality of the local forecasts and the guidance over the years is evident. For instance,

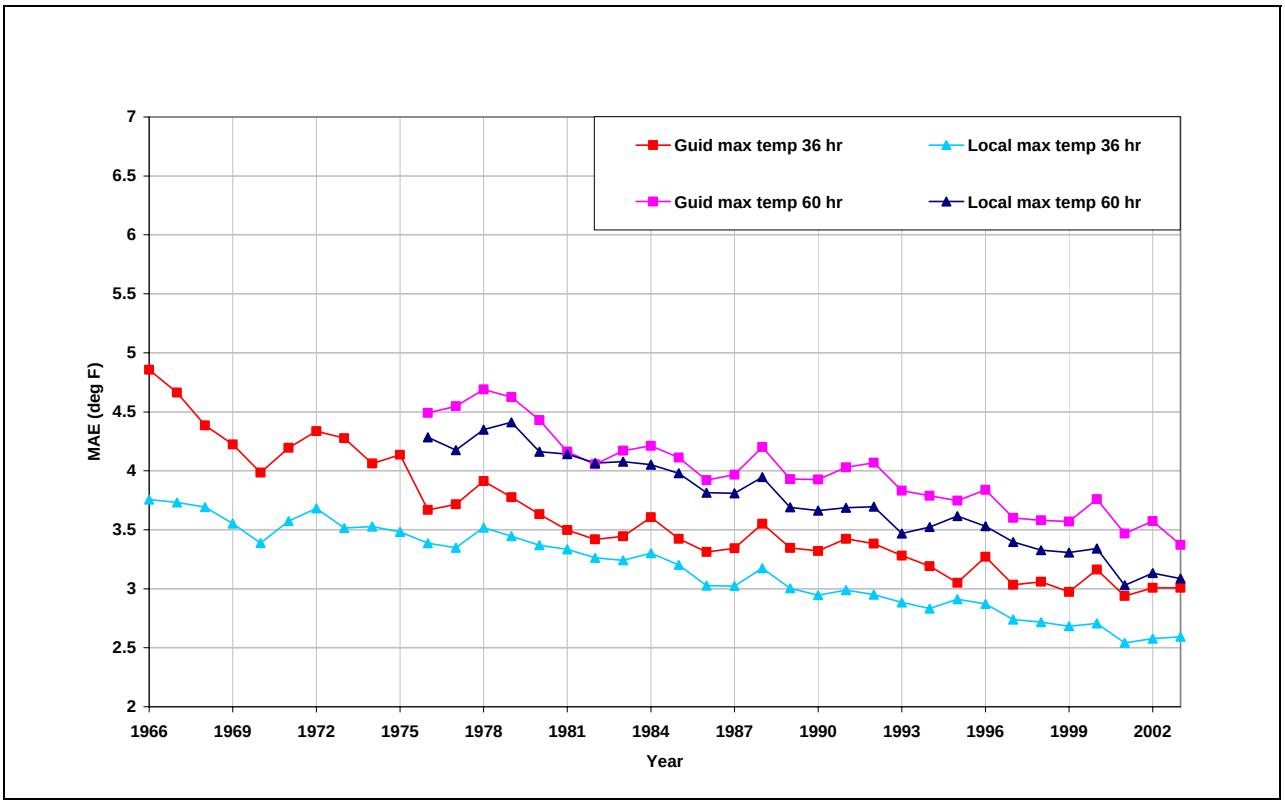

Figure 4. Same as Fig. 2, except for tomorrow's max (36-hr) and the day after tomorrow's max (60-hr) generated during the 1200 UTC cycle.

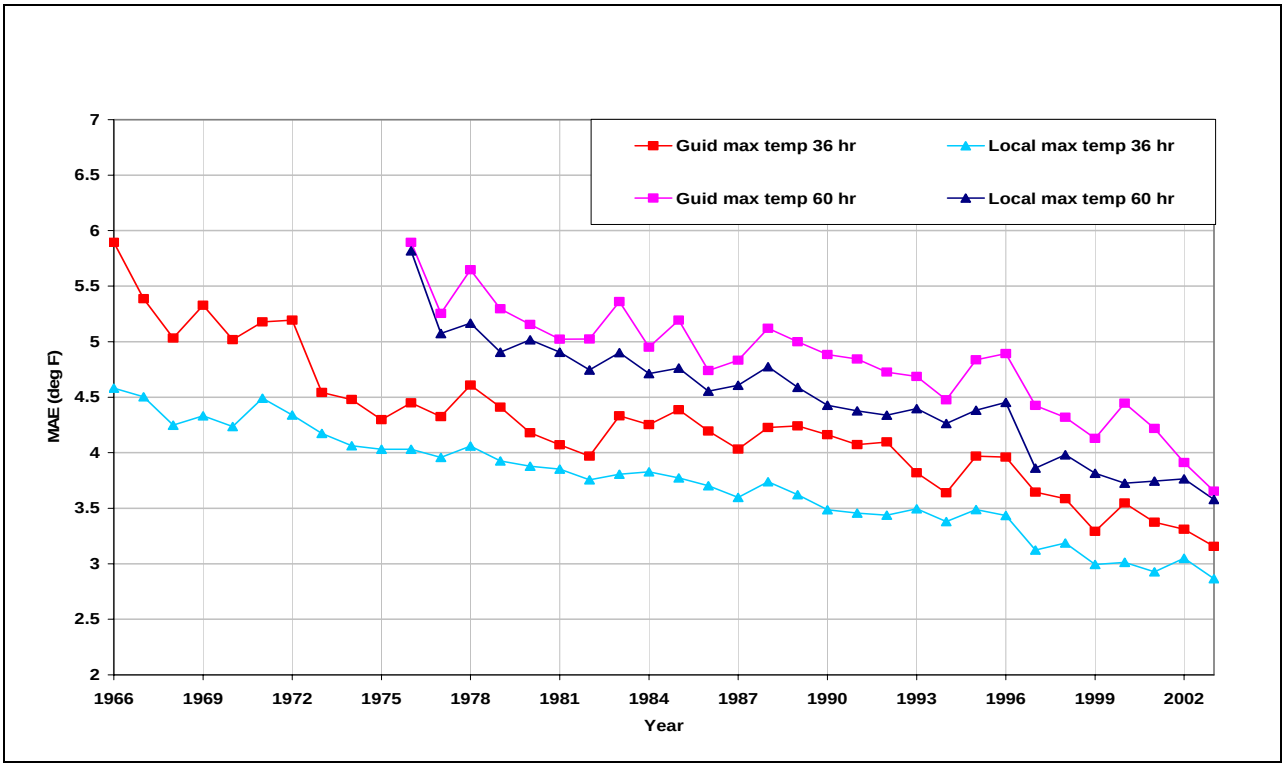

Figure 5. Same as Fig. 4, except for cool season.

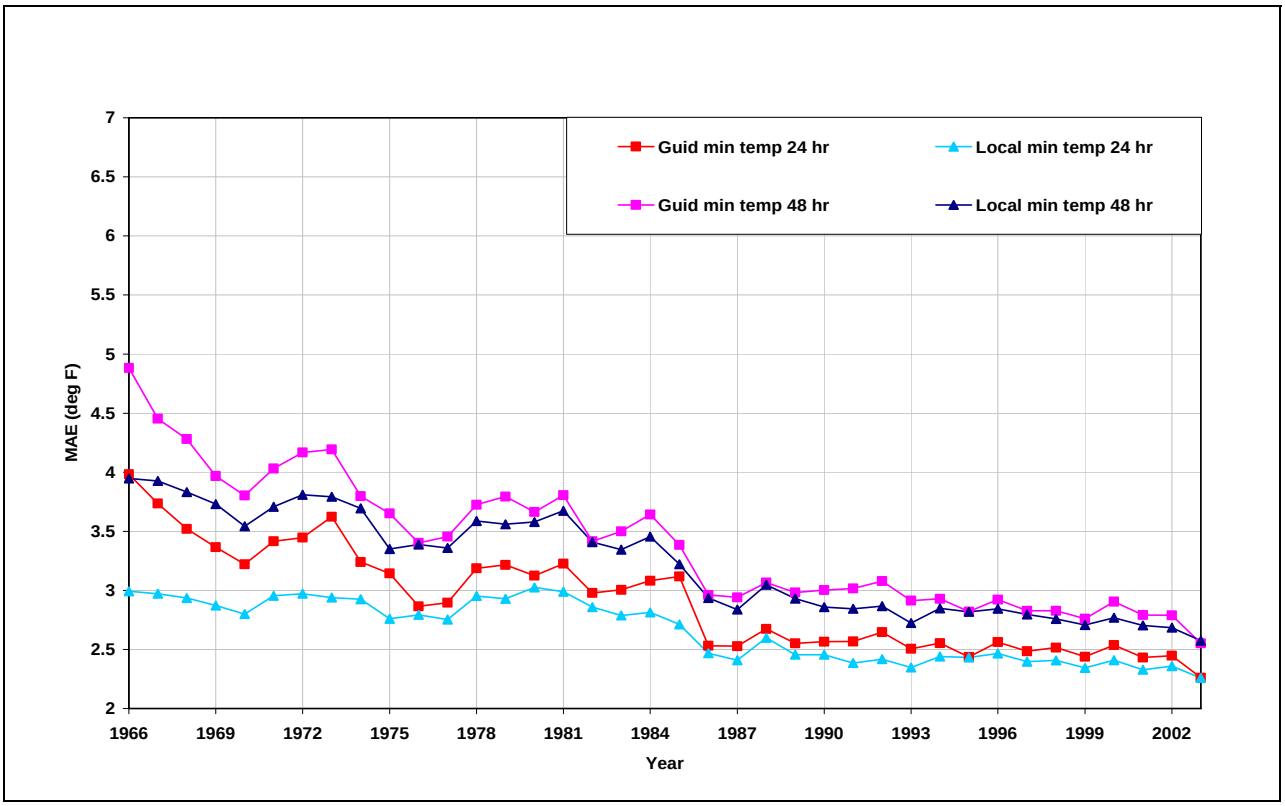

Figure 6. Mean absolute error for warm season local and guidance forecasts of tonight's (24-hr) and tomorrow night's (48-hr) min temperature generated during the 1200 UTC cycle.

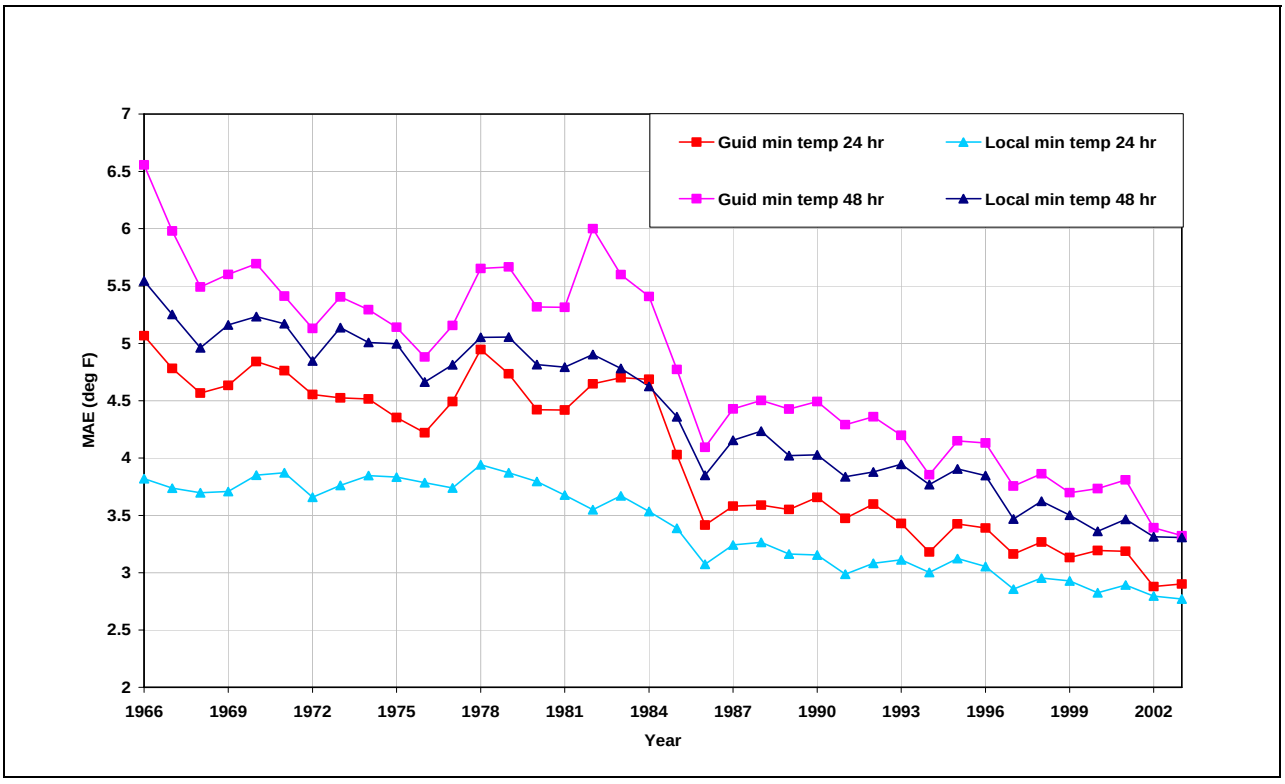

Figure 7. Same as Fig. 6, except for the cool season.

the accuracy of the local forecast of tomorrow night's min is now as good as the accuracy of tonight's min was 20 years ago. Similarly, the errors of the guidance for tomorrow night's min are now about the same as the errors of the guidance for tonight's min about 10 years ago. Second, unlike the max temperature forecasts, the difference in accuracy between the local forecasts and the objective guidance is very small during the warm season, and has become increasingly smaller during the cool season. Third, the implementation of the MOS approach in August 1973 seemed to improve the warm season guidance. A similar improvement was not evident in the cool season guidance. Fourth, the implementation of the nighttime min guidance in late 1984 seemed to have the biggest impact on the errors of both the local forecasts and the guidance. The average error was lower after that time, and the inter-annual variability decreased substantially. After the 1984 warm season, the accuracy of the local forecasts and the guidance seems relatively invariant from year to year, except for a slow downward trend in the errors. In contrast, during the cool season, the inter-annual variability in the errors is noticeably less after the 1985-86 cool season, and the improvement in the errors of the 48-h local forecasts and guidance is pronounced.

Figures 8 and 9 show the verification time series for tomorrow's min (36-h) and tomorrow night's min (60-h) generated during the 0000 UTC cycle for the warm and cool seasons, respectively. As with the 60-h max temperature guidance, the improvement in the guidance for tomorrow night's min is quite dramatic after changes were made during the 2002 summer to use the GFS-based MOS guidance as the official guidance package.

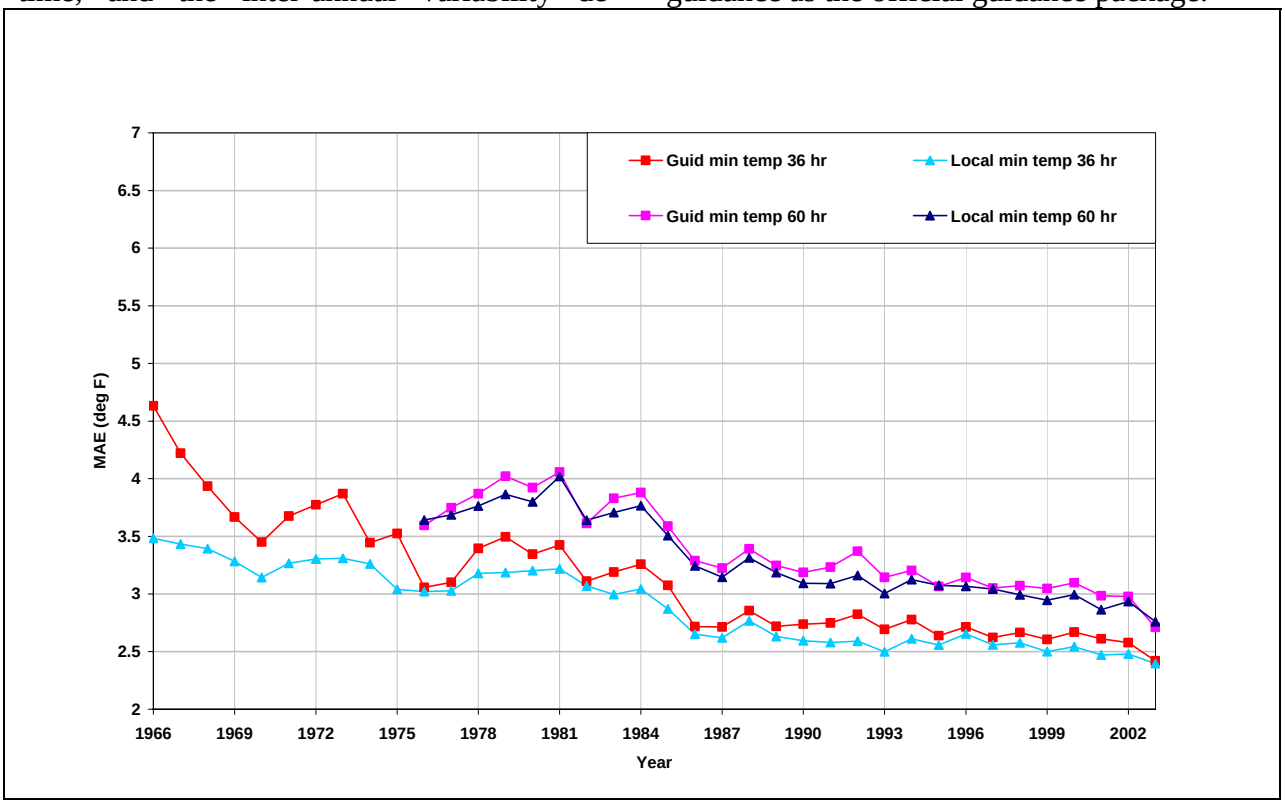

Figure 8. Same as Fig. 6, except for tonight's min (36-hr) and tomorrow night's min (60-hr) generated during the 0000 UTC cycle.

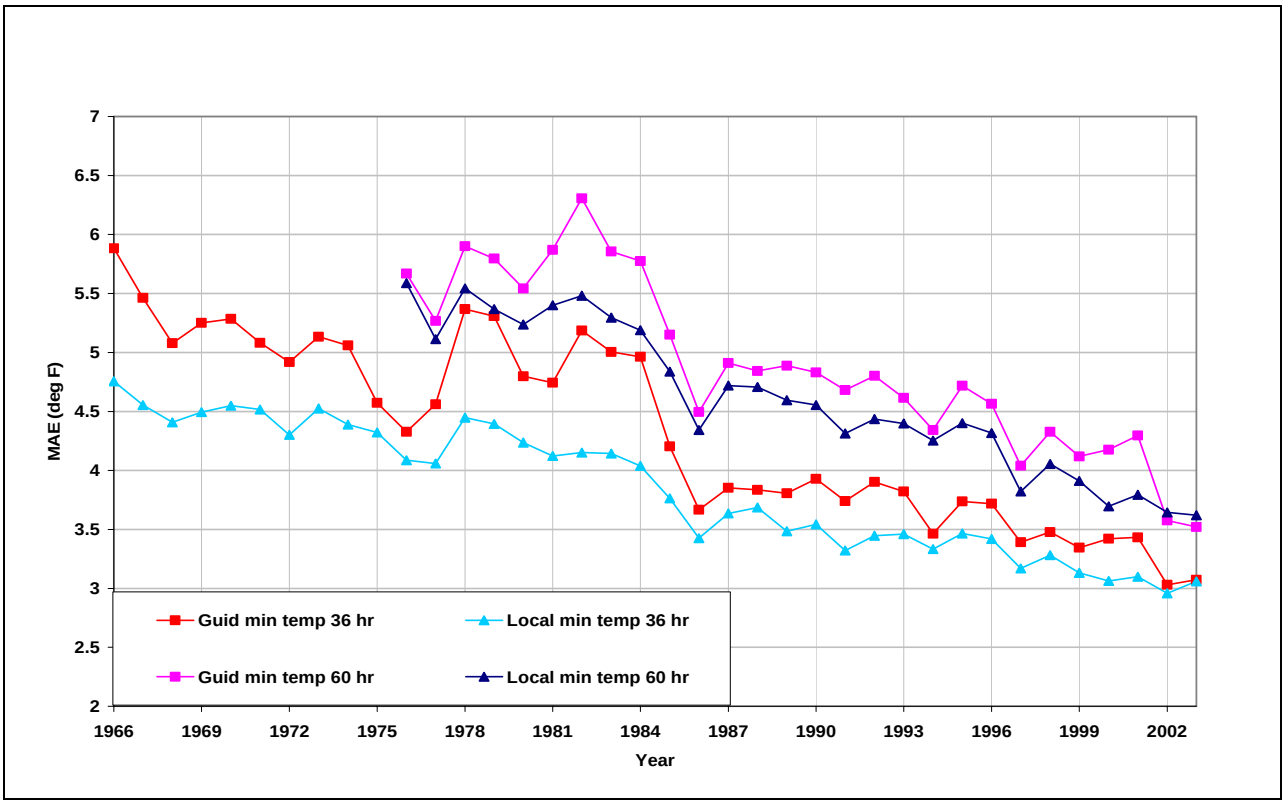

Figure 9. Same as Fig. 8, except for the cool season.

The Brier skill score (defined as the improvement in the Brier score of the forecasts or guidance, relative to a Brier score based on a prediction of the climatic relative frequency) is shown in Figs. 10 and 11 for the warm and cool seasons, respectively. Only the scores for the 12-24 and 36-48 h projections for the local and guidance forecasts are shown. Forecasts from both the 0000 and 1200 UTC cycles were combined in the verifications. Note that the Brier skill score for the 24-h guidance prior to 1969 is meaningless because only categorical guidance, rather than probabilities, were available. Several features of the verifications are notable. First, the skill scores are much higher in the cool season than in the warm season. Thus, by the 2003-04 cool season, the Brier skill scores for the 48-h PoP are actually higher than the warm season skill scores for the 24-h PoP. Second, the improvement in the skill scores for both the local forecasts and the guidance is quite dramatic, particularly in the cool season. As was seen in the temperature verifications, the skill of the local forecasts for the 48-h PoP is now about as great as the skill of the 24-h PoP was 20 years ago. Similarly, the skill of the 48-h PoP guidance is now about the same as the skill of the 24-h PoP guidance 10 years ago. Third, the correlation in skill between the local forecasts and the guidance is very strong. Fourth, while the local forecasts are generally more skillful than the guidance, in some seasons (for example, the 2002-03 and 2003-04 cool seasons and the 2003 warm season), the 48-h guidance is more skillful than the local forecasts. Lastly, the correlation between the improvement of the local forecasts or the guidance and a specific model implementation or change in the statistical approach does not seem high. Rather, the improvements tend to reflect the overall improvement in the NWP system.

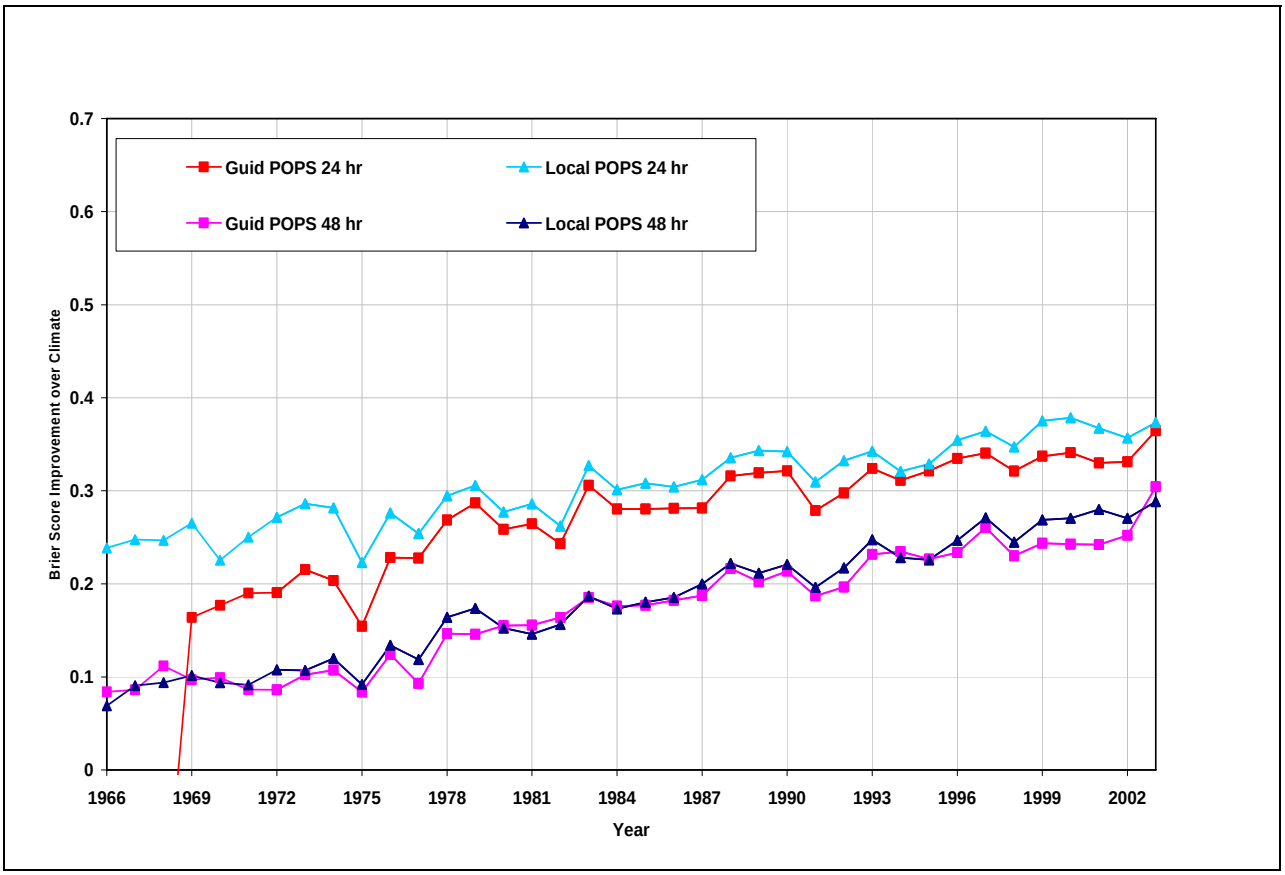

Figure 10. Improvement in Brier score of local and guidance PoP's for the 12-24 h (24hr) and 36-48 h (48 hr) forecasts during the warm season. Results from both cycles were combined.

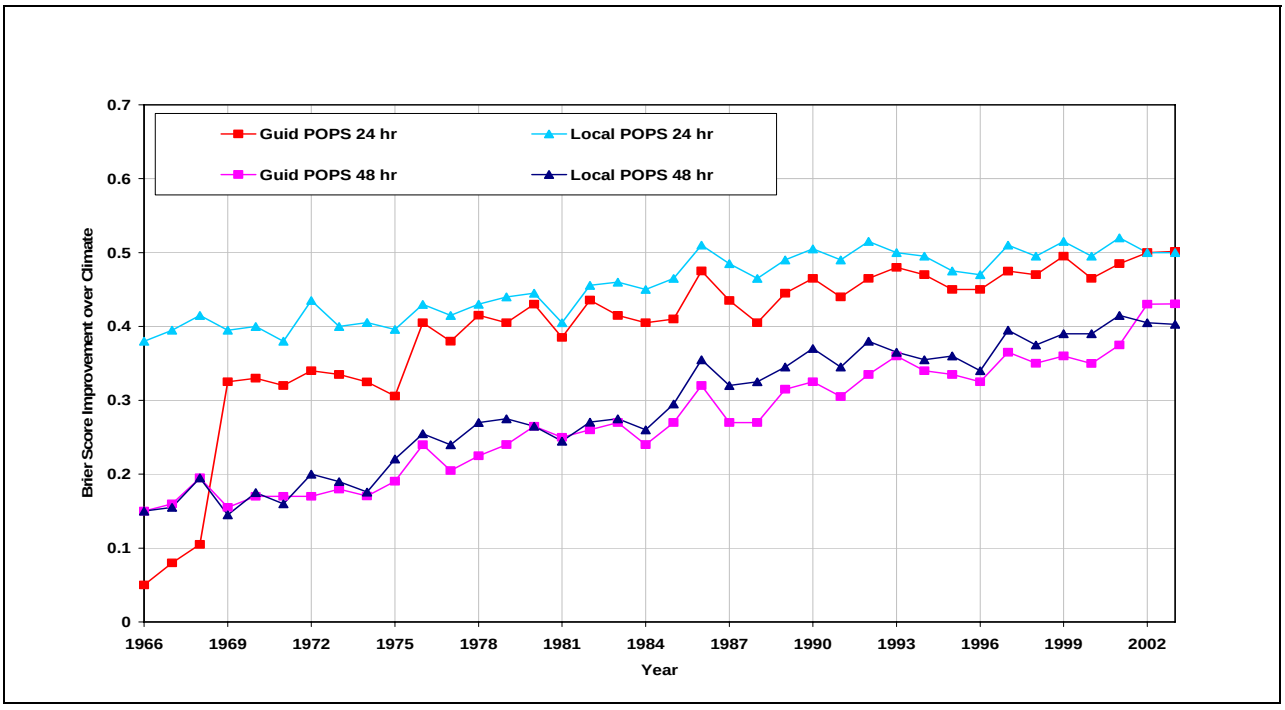

Figure 11. Same as Fig. 10, except for the cool season.

.

6. CONCLUSIONS

As the results indicate, the improvement in the quality of the local forecasts and the guidance is quite obvious in the verification time series. In general, the local forecasts for day 2 are as good as they were for day 1 about 20 years ago. The quality of the guidance for day 2 is about as good as the guidance for day 1 was 10 years ago. Improvements for the 48- and 60-h forecasts are particularly obvious. In general, temperature forecasts during the cool season have larger errors than during the warm season, but also show the greater improvement in skill over the period. The PoP forecasts have the greatest skill during the cool season when convection and mesoscale systems are less common.

To understand the improvement in the overall forecast system since the late 1960's, we will eventually add other verification measures to this study. For instance, the seasonal verification scores can be fit with regression lines to determine trends in improvement. The Brier score can be decomposed into its components to assess whether forecast reliability or resolution has increased. We plan to look at whether large errors in the temperature forecasts have decreased over the years. Finally, differences in the rates of improvement between regions of the country should indicate where improvements in the NWP models have had the greatest impact.

7. ACKNOWLEDGMENTS

Over the years, numerous NWS employees in NWS Headquarters, the forecast offices, and NCEP have contributed to the establishment of this verification database. We can not possibly name them all, but we are grateful for their efforts. We thank Chad Shafer who did many of the verification diagrams shown here, and Kari Sheets who provided the map indicating the verification sites.

8. REFERENCES

Brier, G. W., 1950: Verification of forecasts expressed in terms of probabilities. *Mon. Wea. Rev.*, **78,** 1-3.

Carter, G. M., and P. D. Polger, 1986: A 20-year summary of National Weather Service verification results for temperature and precipitation. NOAA Tech. Memo. NWS FCST-31, 50 pp.

Charba, J. P., and W. H. Klein, 1980: Skill in precipitation forecasting in the National Weather Service. *Bull. Amer. Meteor. Soc.*, **61,** 1546-1555.

Dagostaro, V. J., 1985: The national AFOS-era verification data processing system. TDL Office Note 85-9, 47 pp. [Available from NWS Meteorological Development Laboratory, 1325 East-West Highway, Silver Spring, MD 20910-3283.]

Dagostaro, V. J., C. L. Alex, G. M. Carter, J. P. Dallavalle, and P. D. Polger, 1989: Evolution of the NWS national verification system: Past, present, and future. Preprints *11th Conference on Probability and Statistics in Atmospheric Sciences*, Monterey, CA, Amer. Meteor. Soc., J41-J46.

Erickson, M. C., and J. P. Dallavalle, 1986: Objectively forecasting the short-range maximum/minimum temperature—A new look. Preprints *11th Conference on Weather Forecasting and Analysis*, Kansas City, MO, Amer. Meteor. Soc., 33-38.

Glahn, H. R., 1985: Yes, precipitation forecasts have improved. *Bull. Amer. Meteor. Soc.,* **66,** 820-830.

Glahn, H. R., and D. A. Lowry, 1972: The use of Model Output Statistics (MOS) in objective weather forecasting. *J. Appl. Meteor.,* **11,** 1203-1211.

Jensenius, J. S., and M. C. Erickson, 1987: Monthly relative frequencies of precipitation for the United States for 6-, 12-, and 24-h periods. NOAA Tech. Rep. NWS TDL-39, 262 pp.

Klein, W. H., and F. Lewis, 1970: Computer forecasts of maximum and minimum temperatures. *J. Appl. Meteor.*, **9,** 350-359.

Lowry, D. A., and H. R. Glahn, 1976: An operational model for forecasting probability of precipitation—PEATMOS PoP. *Mon. Wea. Rev.,* **104,** 221-232.

Murphy, A. H., and T. E. Sabin, 1986: Trends in the quality of the National Weather Service forecasts. *Wea. Forecasting*, **1,** 42-55.

Ramage, C. S., 1982: Have precipitation forecasts improved? *Bull. Amer. Meteor. Soc.*, **63,** 739-743.

Ruth, D. P., and C. L. Alex, 1987: AFOS-era forecast verification. NOAA Techniques Development Laboratory Computer Program NWS TDL CP 87-2, 50 pp. . [Available from NWS Meteorological Development Laboratory, 1325 East-West Highway, Silver Spring, MD 20910-3283.]

Zurndorfer, E. A., J. R. Bocchieri, G. M. Carter, J. P. Dallavalle, D. B. Gilhousen, K. F. Hebenstreit, and D. J. Vercelli, 1979: Trends in comparative verification scores for guidance and local aviation/public weather forecasts. *Mon. Wea. Rev.,* **107,** 799-811.