Abstract

During the past several years, there has been an increasing call for the automation of public forecasts issued by the National Weather Service (NWS). This call is the result of national verification statistics which show that, when all forecasts are averaged together, the improvement of NWS forecasters over computer generated forecasts of maximum/minimum temperature and probability of precipitation is small. However, grouping all forecasts in such a manner can hide certain trends which have been evident to field forecasters for many years, namely, that computer generated forecasts are excellent and hard to beat when the weather is seasonably normal, but field forecasters do much better when the weather is unusual. This paper will present the results of a verification study, which shows a significant correlation between abnormal temperature patterns, and the ability of field forecasters at the Albany, New York forecast office to improve upon computer generated forecasts of both the temperature and the probability of precipitation.

1. Introduction

The National Weather Service (NWS) has produced a Model Output Statistics (MOS) guidance package (Glahn and Lowry 1972) since the early 1970's. For nearly two decades, MOS guidance was based on output from the Limited-area Fine-Mesh (LFM) model (Newell and Deven 1981) and was known as the FPC guidance (National Weather Service 1983). The FPC guidance quickly became the standard used to measure local forecast performance. Overall, most local forecasters had little difficulty improving upon the FPC forecasts, as was shown by the initial verification results and by NWS APOS era Verification (AEV) results (Dagostaro 1985). Since the late 1980's, another MOS guidance package has been produced based on output from the Nested-Grid Model (NGM; Hoke et al. 1989) and it is known as the FWC guidance (National Weather Service 1992). Overall, the NGM has been much better than the LFM model, and this resulted in the FWC forecasts being better than the FPC forecasts once there was a sufficient database of NGM data to use for MOS equation development (Jacks et al. 1990). For many years the FPC and FWC guidance packages were produced simultaneously. During much of this time, the FPC guidance remained the standard used to measure local forecast performance. However, in 1993, the FWC guidance became the standard for comparison and the FPC guidance was discontinued shortly thereafter.

Since 1993, verification results indicate, overall, that the skill of the local forecasts of probability of precipitation (PoP) has been about the same as the skill of the FWC forecasts, and the local 12-h maximum/minimum temperature (TEMP) forecasts have been a little better than the FWC forecasts (Dagostaro and Dallavalle 1997). These verification results might appear to suggest that local forecasters add very little value to the 6- to 60-h general public forecasts, and that these forecasts could now be automated through the use of computer worded forecasts (Glahn 1979) based on MOS. However, a verification study carried out at the Albany, New York forecast office, shows that local forecasters at Albany significantly improve upon the FWC guidance when large temperature anomalies occur. Of course, the weather is of considerable interest to the general public during periods when the regime is anomalous compared to the average conditions expected at a given time of the year. The public's attention to weather information increases greatly during periods of unusually cold or hot conditions, unusually wet or dry periods, or when major storms approach. This study will show that the NWS forecasters at Albany were able to add considerable value to public forecast products during those periods when unusually cold or hot conditions occurred. In contrast, the ability of local forecasters to make significant improvements to MOS guidance during unusually wet or dry periods was not conclusive, and no effort was made in this study to quantify forecaster improvement over guidance for individual major storm events.

2. Definitions

PoP and TEMP forecasts were examined for Albany, New York and for Burlington, Vermont by using AEV data for the 45-month period of July 1993 through March 1997. Specifically, PoP forecasts for 12-h periods were verified for the first (12-24 h), second (24-36 h), and third (36-48 h) periods from the 0000 and 1200 UTC cycles. For TEMP, maximum/minimum temperature forecasts were verified for the same 12-h periods as for PoP, but were also verified for the fourth (48-60 h) period from the 0000 and 1200 UTC cycles. For each month, the Frequently and Effectively Departs Significantly (FEDS) score (Maglaras 1991) was used to determine the local forecast
improvement over MOS for TEMP, PoP, and TEMP/PoP forecasts combined. This score is based on the premise that one of the most desirable overall verification measures is to determine how frequently local forecasters deviate substantially from MOS, and how effective they are when they do so. Thus, for each month, the FEDS score is calculated by multiplying the monthly frequency (in percent) of significant changes (F), by the monthly percent improvement over MOS (I) when significant changes are made, and then dividing by ten. To this total, the overall monthly percent improvement over MOS (OI) is then added. Hence:

\[ \text{FEDS} = \left( \frac{F \times I}{10} \right) + OI \]

For TEMP forecasts, a significant change is defined as those cases where the local forecast deviated from MOS by 3 °F, or more, and the percent improvement over MOS is determined from the Mean Absolute Error (MAE) score. For PoP forecasts, a significant change is defined as those cases where the local forecast deviated from MOS by 20% or more, and the percent improvement over MOS is determined from the Brier score. Forecasters who frequently deviate significantly from MOS guidance, and who are also effective when they do so, will have the highest FEDS scores. Forecasters who do not deviate frequently or who are not effective when they do so, or both, will have lower FEDS scores.

For each month in the sample, the TEMP and PoP FEDS scores were calculated for all forecast periods and both forecast cycles combined. The combined TEMP/PoP FEDS score for each month was calculated by adding the corresponding TEMP and PoP FEDS scores. Each month was comprised of about 540 individual PoP forecasts and 720 TEMP forecasts, which results in a combined TEMP/PoP forecast total of about 1260. In general, local forecasters made significant changes to MOS TEMP forecasts 15% to 35% of the time, and significant changes were made to MOS PoP forecasts 10% to 20% of the time.

In order to determine how much the weather deviated from normal in terms of temperature, the Average Daily Temperature Departure (ADTD) was calculated for each month. The ADTD is defined as the sum of the absolute values of the daily temperature departures, divided by the number of days in the month,

\[ \text{ADTD} = \left( \sum_{n=1}^{m} |(T_O - T_A)| \right) / m \]

where
\[ m \] = number of days in the month
\[ T_O \] = the observed daily average temperature
\[ T_A \] = the climatological daily average temperature.

The temperature departure from normal for each month can sometimes mask extreme weather changes that occur during the course of the month. For example, at both Albany and Burlington, on a daily basis, the temperature averaged 10 to 15 °F below normal for about the first half of January 1996, and the weather also featured major snowstorms. For the last half of January 1996, the temperature averaged 10 to 15 °F above normal and there were heavy rainfalls accompanied by record, or near-record flooding. The temperature departure from normal at Albany (Burlington) for the month was zero °F (+1.2 °F), and masked the extreme nature of the weather that month. On the other hand, the ADTD was 12.4, which confirmed that January 1996 was the second most anomalous month of the 45-month sample in terms of temperature.

An example of the improvements over MOS temperature forecasts by Albany local forecasters during the cold part of the month was evident in the forecasts made from the 1200 UTC cycle on 5 January 1996, and are shown in Table 1. The forecaster on duty this day was able to improve on MOS guidance by a total of 42 °F, or an average of more than 5 °F per forecast period. MOS was in error by an additional 56 °F for this set of forecasts, or a total of 98 °F. The observed maximum/minimum temperatures during this period were about 20 °F to 30 °F below normal.

<table>
<thead>
<tr>
<th>Period</th>
<th>ALB</th>
<th>BTV</th>
</tr>
</thead>
<tbody>
<tr>
<td>MOS FCST</td>
<td>-3</td>
<td>20</td>
</tr>
<tr>
<td>LOCAL FCST</td>
<td>-10</td>
<td>17</td>
</tr>
<tr>
<td>OBS TEMP</td>
<td>-19</td>
<td>6</td>
</tr>
<tr>
<td>IMP over MOS</td>
<td>+7</td>
<td>6</td>
</tr>
</tbody>
</table>

Large improvements over MOS guidance were also made during the warm part of the month. For example, on 19 January 1996, the high temperature at Albany reached 60 °F, and at Burlington the high reached 65 °F. These temperatures were 30 °F to 35 °F above normal. The second period MOS forecasts that were made for this day from the 1200 UTC cycle on 18 January 1996, were 44 °F and 45 °F, respectively, for Albany and Burlington. The corresponding local forecasts were 54 °F and 60 °F. This resulted in an improvement over MOS for each station of 10 °F and 15 °F, respectively. The improved local temperature forecasts were not only significant because they let the public know how unusually warm it would be that day, it also meant that forecasters on duty were anticipating that snow melt would be a significant factor in the record-breaking flooding that would eventually occur on 19 and 20 January.

Precipitation anomaly measures were also calculated. For each of the 45 months that comprised the data sample, the absolute value of Monthly Precipitation Amount Departure from normal (MPAD) was calculated (in percent), and was used to determine the anomaly for precipitation amount. In order to determine how anomalous each month was in terms of the frequency of measurable (≥ .01 inches) precipitation events, the absolute value of Monthly Precipitation Frequency Departure (MPFD) was also calculated (in percent).
3. Results

Figure 1 is a scatter diagram of the combined TEMP/PoP FEDS score and the ADTD for each month in the verification data sample. Figure 1 also shows the corresponding best fit linear regression line which was calculated using the Statistical COrelation and REgression program (SCORE) (Wooldridge and Burrus 1997). These results reveal a strong correspondence between the improvement of Albany local forecasts over FWC guidance, and the departure of temperature from normal. As the ADTD increases (greater temperature anomalies) the local forecast improvement over guidance increased rapidly.

Past experience with the combined TEMP/PoP FEDS score has shown that a value over 50 was a good score, and a value of 100 or more was an outstanding score. Based on the regression line shown in Fig. 1, when the ADTD is around 4, the TEMP/PoP FEDS score will be around zero, which indicates little or no overall improvement over MOS. When the ADTD is around 7.5, the TEMP/PoP FEDS score will be around 50 (a solid improvement over MOS). Finally, when the ADTD is around 11, the TEMP/PoP FEDS score will be around 100, which is an outstanding improvement over MOS.

In order to help the reader compare the use of the FEDS score to more traditional verification scores, the combined TEMP/PoP FEDS score for January 1996 (the highest FEDS score in the 45-month sample), will be calculated as an example. Overall, for all forecasts combined, the MAE of MOS TEMP forecasts for January 1996 was 5.3 °F, while for local forecasters it was 4.6 °F. This gave an overall percent improvement over MOS for the month (OI) of 14.1. During this month, forecasters deviated from MOS TEMP forecasts by 3 °F or more, 37.7% of the time. Thus, the frequency of significant changes in percent (F) was 37.7. When local forecasters made significant changes, the MAE of those forecasts for MOS was 6.5 °F, while local forecasters had a MAE of 4.7 °F, which gave a monthly percent improvement over MOS when significant changes were made (I) of 27.9. Thus, the TEMP FEDS score for January 1996 was 119.3. The Brier score for all MOS PoP forecasts was .122, while local PoP forecasts had a Brier score of .108, and resulted in an overall percent improvement over MOS for the month (OI) of 11.4. Local forecasters deviated from MOS PoP forecasts by 20% or more, 11.6% of the time, so the frequency of significant changes in percent (F) was 11.6. The percent improvement over MOS when significant changes were made (I) was 44.5. Thus, the PoP FEDS score for January 1996 was 63.0, and the combined TEMP/PoP FEDS score was 182.3.

Figures 2 and 3 are the same as Fig. 1, except they show the relationship of the TEMP FEDS score and the PoP FEDS score, respectively, to the ADTD. Figure 2 shows a correspondence between the improvement of Albany local TEMP forecasts over MOS forecasts, and the departure of temperature from normal. As the ADTD increases, the local TEMP forecast improvement over
guidance increases considerably. In addition, Fig. 2 also reveals that local TEMP forecasts were as good as or better than MOS for nearly every month in the sample.

Figure 3 also shows a correspondence between the improvement of Albany local PoP forecasts over MOS forecasts, and the departure of temperature from normal. As the ADTD increases, the local PoP forecast improvement over guidance increases, but not as greatly as for TEMP forecasts. In fact, overall, Fig. 3 shows that local and MOS PoP forecasts are about equal. Local forecasters did better than MOS for only about half of the months in the sample. However, the trend of increasing local forecaster improvement over guidance with increasing temperature departure from normal is still evident.

Based on Figs. 2, and 3, as temperature anomalies increase, forecaster improvement over MOS guidance shows a significant increasing trend not only for TEMP forecasts, but (to a lesser extent) for PoP forecasts as well. This is not surprising since temperature, precipitation, and other meteorological variables are not independent. For example, even if the air mass is not very cold, a cloudy and rainy day in the summer will usually result in a daytime maximum temperature that is 10 to 20 °F below normal. On the other hand, a calm, clear, precipitation free night in January when there is snow cover on the ground could result in a nighttime minimum temperature that is 10 to 20 °F below normal. In these scenarios, the abnormal surface temperature readings are the result of interactions with other meteorological variables. Frequently, especially when MOS has not adequately taken into account precipitation, clouds, and wind, MOS TEMP forecasts will not do well in such scenarios and local forecasters have a good opportunity to make significant improvements over MOS forecasts. However, before large improvements can be made, local forecasters must also correctly forecast precipitation, clouds, and wind. The combined TEMP/PoP FEDS score was used in this paper as an overall measure of forecaster performance. The fact that the improvement trend of the combined TEMP/PoP FEDS score is greater than for the individual elements (shown in Figs. 1-3) is a reflection of the fact that, generally, when large temperature deviations occur, local forecasters not only provide better forecasts of temperature than does MOS, but, in order to do so, they must also provide better forecasts of other meteorological variables as well. The verification data used in this study imply that this hypothesis is true. On a monthly basis, most of the time, when the local TEMP FEDS score was high and the ADTD was large, the local PoP FEDS score was also high. Conversely, when the local TEMP FEDS score was low and the ADTD was low, the local PoP FEDS score was low.

An analysis of the FEDS score and its relationship to the ADTD, MPAD, and MPFD anomaly measures was done using the SCORE program. The results of this analysis are shown in Table 2. For temperature departures, Table 2 reveals that the correlations of the TEMP, PoP, and combined TEMP/PoP FEDS scores to the ADTD were higher than for any other measure and were 54.6%, 48.2%, and 64.9%, respectively. Also, the statistical correlation between the combined TEMP/PoP FEDS score and the ADTD, which was 64.9% and is shown graphically in Fig. 1, was considerably higher than for the TEMP and PoP FEDS scores, individually. This adds support to the hypothesis discussed in the previous paragraph, that when large temperature deviations occur, local forecasters will usually make significant improvements over MOS for both elements, resulting in a higher combined FEDS score.

We tested the correlations in Table 2 for significance using the F-test. The F-test, as used in this study, tests the utility of the best fit linear regression line at predicting future values of the FEDS score, based on the value of a particular anomaly measure. The higher the F-score, the higher our confidence level that the linear regression line is useful for predicting the value of the FEDS score, and, thus, that the correlation of the data in the sample is significant. F-test results are usually considered significant when the confidence level is 95% or higher. The results of the F-test showed that the correlation of the TEMP, PoP, and TEMP/PoP FEDS score, respectively, to the ADTD were all significant at the 99% confidence level.

For precipitation amount departures, Table 2 shows that the correlations of the TEMP, PoP, and the TEMP/PoP FEDS scores to the MPAD were 27.3%, 12.5%, and 26.0%, respectively. For precipitation frequency departures, the correlations of the TEMP, PoP, and the TEMP/PoP FEDS scores to the MPFD were 8.0%, -34.1%, and -13.3%, respectively. The F-test showed that most of these correlations were not significant. The only significant correlation associated with precipitation anomalies was the correlation of the PoP FEDS score to the MPFD, which was -34.1%. However, the correlation of the MPFD to the PoP FEDS score was significant at the 95% confidence level, and was not as clear as the correlation of the ADTD to the TEMP, PoP, and the TEMP/PoP FEDS scores which were significant at the 99% confidence level. Even though the relationship is weaker, the verification data still suggest that as the MPFD increases, local forecaster improvement over MOS PoP guidance decreases. Although the results are not shown here, the relationship of forecaster performance to precipitation frequency deviation was based solely on the magnitude of the monthly precipitation frequency departure. When the monthly precipitation frequency departures were also stratified by wet (more frequent precipitation) and dry (less frequent precipitation) months, there was little or no correlation evident.

| Table 2. Correlations of the TEMP, PoP, and TEMP/PoP FEDS scores to the ADTD (average daily temperature departure), MPAD (monthly precipitation amount departure) and MPFD (monthly precipitation frequency departure) anomaly measures for the period from July 1993 through March 1997. |
|---------------------------------|-----------------|-----------------|
| ADTD                           | MPAD            | MPFD            |
| TEMP FEDS                      | +0.546          | +0.273          | +0.080          |
| PoP FEDS                       | +0.482          | +0.125          | -0.341          |
| TEMP/PoP FEDS                  | +0.649          | +0.260          | -0.133          |
The correlation of forecaster improvement over MOS guidance to "abnormal" long-term precipitation patterns, generally, was not conclusive in this study. One reason might be that, unlike temperature, precipitation is not a continuous variable. As a result, on a daily basis, or for a specific forecast period, daily precipitation departures from normal have little meaning and were not calculated. Even on a longer term basis, such as the monthly basis used in this study, the precipitation anomaly measures could produce mixed results. For example, below (above) normal precipitation amounts can occur even when the precipitation frequency is above (below) normal, if most of the precipitation events during the month were light (heavy). Perhaps the use of another precipitation anomaly measure might have produced more conclusive results. However, it is difficult to conceive of any measure that would quantify the true "abnormal" nature of the precipitation pattern on a daily basis, as does the ADTD for the temperature pattern.

4. Discussion

Based on the results of this study, it can be concluded that, overall, local forecasters at the Albany forecast office are successful at making significant changes to, and improving on MOS forecasts of both TEMP and, to a lesser extent, PoP, during periods of "abnormal" temperature conditions. The study also showed a clear trend of increasing local forecaster improvement over guidance with increasing temperature departures from normal. This should come as no surprise since it is a well known fact that MOS guidance has difficulty with rare events or with weather patterns that deviate substantially from climatological normals (Lowry 1980; Murphy and Dallavalle 1984; Maglaras and Carter 1986; Carter et al. 1989; and Dallavalle and Erickson 1993). Conversely, during periods of "normal" temperature conditions, or during the warm season when deviations from normal generally are much less, local forecast improvements over MOS guidance are reduced.

Lowry (1980), Murphy and Dallavalle (1984), Maglaras and Carter (1986), Carter et al. (1989), and Dallavalle and Erickson (1993) indicated that MOS guidance usually performs well within the range of the average conditions which occurred in the developmental sample. The guidance will show a decreasing trend in accuracy as the weather conditions deviate further and further from this "normal range." Also, this decreasing trend will be more pronounced at later forecast periods. (As noted in the previous section, for a specific day or forecast period, the idea of "normal range" for a non-continuous variable, such as precipitation, is not relevant). These characteristics of MOS will not change, even when future MOS developments occur based on more accurate numerical forecast models.

The findings of this study and the inherent characteristics of MOS guidance leave the meteorological community with a dilemma. On the one hand, we have an automated system for forecasting (MOS guidance and computer worded forecasts) which performs very well during periods of near normal temperatures and during much of the warm season. During these periods the local forecasters add little value, overall, to the TEMP and PoP forecasts before they are issued to the general public. On the other hand, local forecasters perform much better during periods when the temperature deviates significantly from normal. At these times, the local forecasters add substantial value to the TEMP and PoP forecasts before they are issued to the public.

Most of the time, MOS guidance and computer worded forecasts will serve the public well. However, for those periods when the temperature is "unusual" or "extreme," local forecasts will provide better information to the public. Since the public's awareness of the forecast is greatly heightened during periods of "unusual" or "extreme" weather, local forecasters add significant value to the forecast during those periods when the public is most interested in the forecast, and most in need of accurate forecast information. In addition, in order to maintain their proficiency at making large improvements over guidance, local forecasters need to produce PoP and temperature forecasts on a daily basis. If the forecasts for routine situations were delegated exclusively to MOS and computer worded forecasts, the likelihood of forecaster improvement over guidance for periods with anomalous temperature regimes would be diminished considerably. Hence, the apparent trend to migrate towards the automatic generation of most products might need to be reexamined and modified in an appropriate manner.

Future verification work at the Albany forecast office will involve trying to quantify local forecaster improvement over guidance for specific significant temperature and precipitation events.

Acknowledgments

The author thanks the staff of the National Weather Service office in Burlington, Vermont, for providing much of the Burlington climatological data used in this paper. Thanks also to Gary Carter of the NWS Eastern Region Headquarters for his many suggestions for improving this paper. Finally, thanks to all the forecasters (past and present) at the Albany NWS Forecast Office who, without complaint, have allowed this author to conduct research related to the accuracy of their forecasts for the past several years.

Author

George Maglaras has been a Lead Forecaster at the NWS Forecast Office at Albany, New York since 1988. He is interested in statistical forecast methods and analysis, and in forecast verification. From 1980 to 1986 he worked at the NWS Techniques Development Laboratory in Silver Spring, Maryland. Between 1986 and 1988 he worked at the NWS Forecast Office in Washington, D.C., where he was part of the Satellite Field Service Station. He earned his B.S. Degree in Meteorology in 1978, and a M.S. Degree in Computer Science in 1981 from the City College of the City University of New York. In addition, between 1984 and 1986 he completed all course work toward a M.S. Degree in Meteorology at the University of Maryland.
References


