TEMPERATURE AND PRECIPITATION FORECAST VERIFICATION TRENDS AT THE ALBANY FORECAST OFFICE: FORECASTERS CONTINUE TO SHOW IMPROVEMENT ON MOS GUIDANCE – PART II

George J. Maglaras

NOAA/National Weather Service Forecast Office
Albany, New York

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 that, for certain significant weather events, field forecasters at the Albany, New York, forecast office have the ability to substantially improve upon computer-generated forecasts of both the temperature and the probability of precipitation. In addition, for all forecasts combined, this paper will show that the decrease in forecaster ability to improve upon computer-generated forecasts that occurred when a new generation of guidance was implemented during the summer of 1993, may have been temporary. This verification study will also examine individual forecaster trends and the differences between veteran and novice forecasters. This paper is a follow-up study to Maglaras (1998). Maglaras (1998) described the results of a study which showed a significant correlation between abnormal temperature patterns and the ability of Albany, New York, forecasters to improve upon computer-generated forecasts of both the temperature and the probability of precipitation.

1. Introduction

The NOAA/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 Deaven 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 AFOS-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.

For a time since 1993, verification results indicated, overall, that the skill of the local forecasts for probability of precipitation (PoP) was about the same as the skill of the FWC forecasts, while the local 12-h maximum/minimum temperature (TEMP) forecasts were a little better than the FWC forecasts (Dagostaro and Dallavalle 1997). The verification results appeared to suggest that local forecasters added 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. The results also supported the hypothesis that the convergence of human and machine (MOS) forecast skill appears to be inevitable (Roebber and Bosart 1996). However, a verification study carried out at the Albany, New York, forecast office (Maglaras 1998), showed that local forecasters at Albany significantly improved upon the FWC guidance when large temperature anomalies occurred. 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. Past experience shows web page "hit" counts of around 500 to the Internet home page of the Albany forecast office on a typical weather day, but when a major storm or unusual weather event approaches, the "hit" count rises to between 1000 and 2000.

This study will expand on the results of Maglaras (1998), and examine the performance of NWS forecasters at Albany when significant weather events occur. Specifically, forecaster performance will be examined for those cases when record temperatures occurred, when there were large temperature changes from one day to the next, and for significant precipitation events. This study will also examine individual and station verification trends. Specifically, the long-term station verification trend will be discussed, with the emphasis on how Albany local forecasters were affected by the switch from the LFM-based FPC guidance to the NGM-based FWC guidance, and how they have recently done an increasingly better job of improving on the FWC guidance. For individual forecasters, the emphasis will be on the differences between veteran and novice forecasters. Veteran forecasters are defined as those forecasters who had previous forecast experience and who also were familiar with forecasting for the Albany forecast area. Novice forecasters are defined as those forecasters who had no forecast experience when their verification scores were first included in the study, or they had previous forecast experience, but not at Albany.

The results of the individual and station verification trends will also be contrasted to Roebber and Bosart (1996), who found that forecaster skill is largely determined by experience. They also found a steady erosion of human forecast skill relative to MOS forecasts and speculated on the future role of human forecasters in the forecast process.

2. Definitions

PoP and TEMP forecasts were examined for Albany, New York, and for Burlington, Vermont, by using AEV data for the period of April 1990 through March 1999. 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. Due to the different approaches used to examine the verification data and because individual verification was involved, the specific sample periods may vary. However, all data were within the April 1990 through March 1999 period. In addition, some of the verification results were based on seasonal calculations. The seasons were defined as cool (October-March) and warm (April-September), respectively. In some cases, the verification results were calculated on a significant weather event basis. For example, verification results were calculated for all significant precipitation events lumped together. As the results from each aspect of this study are presented, the sample period and the method of calculation are defined.

For each season or significant weather event type, the Frequently and Effectively Departs Significantly (FEDS) score (Maglaras 1991) was used to determine the local forecast improvement over MOS for TEMP, PoP, or 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 data sample, the FEDS score is calculated by multiplying the frequency (in percent) of significant changes (**F**), by the percent improvement over MOS (**I**) when significant changes are made, and then dividing by ten. To this total, the overall percent improvement over MOS (**OI**) is then added. Hence:

$$\mathbf{FEDS} = ((\ \mathbf{F} \times \mathbf{I}\ ) / 10) + \mathbf{OI}$$

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 (or, equivalently, the mean square error for PoP forecasts). Forecasters who frequently deviate substantially 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 season in the sample, the TEMP and PoP FEDS scores were calculated for all forecast periods and both forecast cycles combined. For each significant weather event category, the TEMP and PoP FEDS scores were calculated for all forecasts that fell within the definition of the significant weather event. The combined TEMP/PoP FEDS score for each season or significant weather event category was calculated by adding the corresponding TEMP and PoP FEDS scores. When seasonal calculations were completed, each season included about 3200 individual PoP forecasts and 4300 TEMP forecasts. In general, for seasonal calculations, local forecasters made significant changes to MOS TEMP (PoP) forecasts 15% to 35% (10% to 20%) of the time.

Two methods were used to verify forecasts when record temperatures occurred. The first method **(RECORDS ONLY)** verified only the forecasts made for the specific time that a record maximum, record minimum, record low maximum or record high minimum temperature occurred at ALB (Albany, New York) or BTV (Burlington, Vermont). For example, if ALB had a record maximum temperature occur between 1200 UTC 23 March and 0000 UTC 24 March, only four TEMP forecasts would be verified (PoP forecasts were not verified for the RECORDS ONLY category). These four forecasts would be the first period maximum TEMP forecast for ALB from 0000 UTC 23 March, the second period forecast from 1200 UTC 22 March, the third period forecast from 0000 UTC 22 March, and the fourth period forecast from 1200 UTC 21 March. For the 51-month significant weather sample period of July 1993 through September 1997, record temperatures occurred 82 times at ALB and/or BTV.

The second method **(RECORD PERIODS)** verified all forecasts made during a period of time near the occurrence of the record temperature. If a record temperature occurred at ALB or BTV on a particular day, all TEMP/PoP forecasts made for both stations on that day and on the previous two days were verified. Using the same example from the previous paragraph, if ALB had a record maximum temperature occur between 1200 UTC 23 March and 0000 UTC 24 March, TEMP/PoP forecasts made for all periods and both cycles on 21 March through 23 March for both ALB and BTV would be included. This would result in 48 TEMP forecasts (36 PoP forecasts) being added to this sample due to this one record temperature. The reason this approach was used was to ensure that forecasters not only did well forecasting specifically for the record temperature event, but that the other TEMP and PoP forecasts made during this period were also good. What value would it be to make excellent forecasts of the record temperature if all the other forecasts for ALB and BTV during this period were poor?

The criteria for significant temperature changes **(SIG TEMP CHANGES)** from one day to the next were seasonally adjusted so that a relatively similar number of cases could be included in the sample from all seasons. Specifically, for the months of December through February, a significant temperature change **(SIG TEMP CHANGE)** was defined as a change in the maximum or minimum temperature of 25 °F or more, or a change in the daily mean temperature of 20 °F or more, from one day to the next. For the months of September through November and for March through May, the values for SIG TEMP CHANGES were defined as 20 °F and 15 °F or more, respectively, while for the period of June through August, the values were 15 °F and 10 °F or more, respectively. An approach similar to the one used for temperature RECORD PERIODS was used for SIG TEMP CHANGES. When a SIG TEMP CHANGE occurred at ALB or BTV, all TEMP/PoP forecasts from the two-day period over which the SIG TEMP CHANGE occurred, plus all the forecasts from the preceding day were verified for both stations. For example, if a significant minimum temperature change occurred between 23 and 24 March at ALB, TEMP/PoP forecasts from all periods and both cycles on 22 March through 24 March for ALB and BTV would be included. As a result, 48 TEMP forecasts (36 PoP forecasts) would be added to this sample due to this one SIG TEMP CHANGE. During the 51-month significant weather sample, there were 97 days with SIG TEMP CHANGES. The reasons for using this approach were similar to those for temperature RECORD PERIODS.

Significant precipitation events **(SIG PCPN EVENTS)** were defined as those cases where the liquid equivalent was one inch or more, or the total snowfall was ten inches or more, and the precipitation fell over a period of 48 hours or less. Some subjective judgment was used in order to create the sample of SIG PCPN EVENTS. For example, if two smaller precipitation events occurred over a 48-hour period, but these smaller events clearly were the result of separate synoptic features with a long precipitation free period between the two weather systems, then a significant precipitation event **(SIG PCPN EVENT)** was not defined, even if the total precipitation amount met the criteria.

When a SIG PCPN EVENT occurred at ALB or BTV, all TEMP/PoP forecasts from the one or two days over which the precipitation fell, plus all the forecasts from the preceding two days were verified for both stations. For example, if a SIG PCPN EVENT occurred on 23 and 24 March at ALB, TEMP/PoP forecasts from all periods and both cycles on 21 March through 24 March for ALB and BTV would be included. This would result in 64 TEMP forecasts (48 PoP forecasts) being added to this sample due to this one SIG PCPN EVENT. During the 51-month significant weather sample, there were 60 SIG PCPN EVENTS. Once again, the reasons for using this approach were similar to those for temperature RECORD PERIODS.

3. Verification Results for Significant Weather Events

a. Temperature verification results

Table 1 shows the temperature verification results, for all forecast periods and cycles combined, for those cases when record temperatures, SIG TEMP CHANGES, and SIG PCPN EVENTS occurred during the NGM MOS guidance period from July 1993 through September 1997. For record temperatures, verification results are shown for both methods used, namely, for RECORDS ONLY and for RECORD PERIODS. Table 1 includes the TEMP FEDS score, the overall percent improvement over guidance, and the number of forecasts made for each of these categories. Also included in Table 1 are the temperature verification results for all forecasts **(ALL)** made during the 51-month sample period.

Table 1. TEMP FEDS score and overall percent improvement over guidance for the ALL, RECORD PERIODS, RECORDS ONLY, SIG TEMP CHANGES, and SIG PCPN EVENTS categories for all periods and cycles combined. The scores were calculated for the period from July 1993 through September 1997. Also shown is the number of forecasts in each category.

	FEDS Score	% Improv Over MOS Guidance	# of Fcsts
All	34.9	4.5	24320
Record Periods	61.6	7.8	2504
Records Only	116.7	13.7	322
All	34.9	4.5	24320
SIG Temp Changes	35.8	4.7	4630
All	34.9	4.5	24320
SIG PCPN Events	56.5	7.0	3040

As shown in Table 1, for ALL forecasts, the overall percent improvement over MOS TEMP guidance by Albany local forecasters was 4.5%, and the FEDS score was 34.9. When record temperatures occurred, the improvement by Albany local forecasters over MOS TEMP forecasts for RECORD PERIODS was considerably higher. The TEMP FEDS score was 61.6 and the overall percent improvement over guidance was 7.8%, almost double the scores for ALL forecasts. The 2504 forecasts included in the RECORD PERIODS sample was 10.3% of ALL temperature forecasts.

For RECORDS ONLY, Table 1 shows that Albany local forecasters greatly improved on MOS TEMP guidance. The TEMP FEDS score for RECORDS ONLY was 116.7, and the overall percent improvement over guidance was 13.7%. Both scores were more than three times higher than the scores for ALL forecasts. The 322 forecasts included in the RECORDS ONLY sample was 1.3% of ALL temperature forecasts.

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. Thus, a FEDS score of 116.7 for TEMP forecasting only, is exceptional.

In order to help the reader compare the magnitude of the FEDS score to more traditional verification scores, the TEMP FEDS score for RECORDS ONLY will be calculated. Overall, for all forecasts combined in this category, the MAE of MOS TEMP forecasts was 7.7 °F, while for local forecasters the MAE was 6.6 °F. This gave an overall percent improvement over MOS (**OI**) of **13.7**. For RECORDS ONLY, forecasters deviated from MOS TEMP forecasts by 3 °F or more, 32.0% of the time. Thus, the frequency of significant changes in percent (**F**) was **32.0**. When local forecasters made significant changes, the MAE of those forecasts for MOS was 9.1 °F, while local forecasters had a MAE of 6.2 °F, which gave a percent improvement over MOS when significant changes were made (**I**) of **32.2**. Thus, the TEMP FEDS score for RECORDS ONLY was **116.7**.

When SIG TEMP CHANGES occur, Albany local forecasters improve on MOS TEMP forecasts, but no more than they do for ALL forecasts. As shown by Table 1, the TEMP FEDS score and the overall percent improvement over guidance when SIG TEMP CHANGES occur are nearly the same as for ALL forecasts. The 4630 forecasts included in the SIG TEMP CHANGES category was 19.0% of ALL temperature forecasts.

Table 1 shows that when SIG PCPN EVENTS occur, Albany local forecasters made a considerable improvement over MOS TEMP guidance. The TEMP FEDS score was 56.5 and the overall percent improvement over guidance was 7.0%. Both of these scores were more than one and a half times higher than the scores for ALL forecasts. The 3040 forecasts included in the SIG PCPN EVENTS category was 12.5% of ALL temperature forecasts.

b. Temperature verification results by forecast period

Table 2 shows the TEMP FEDS score and the overall percent improvement over guidance by forecast period for the ALL, RECORD PERIODS, RECORDS ONLY, SIG TEMP CHANGES, and SIG PCPN EVENTS categories. Table 2 reveals that, for the RECORDS ONLY category, Albany local forecasters made extremely large improvements over guidance through the second period (36 hours), with smaller, but still substantial, improvements thereafter. Specifically, Table 2 shows that the TEMP FEDS score for the first, second, third, and fourth periods, respectively, were 167.8, 161.8, 78.2, and 61.6. The overall percent improvements over guidance were 21.9%, 18.6%, 9.0%, and 7.9%, respectively.

For the ALL category, Table 2 reveals that there is a decreasing trend in the improvement over MOS TEMP guidance by Albany local forecasters from the first to the third period, but the improvement over guidance increases for the fourth period. This trend is also evident in the RECORD PERIODS category. It is generally expected that there will be a decreasing trend in local forecaster improvement over guidance from the first to the third period. Many of the forecast tools (i.e., radar, satellite, and surface data) used by forecasters to make improvements to MOS TEMP forecasts in the first period, become increasingly ineffective for the second and third periods. On the other hand, the increase in local forecaster improvement over guidance from the third to the fourth period likely reflects the fact that the forecast tools (i.e., NGM forecasts) used by the FWC guidance are limited. The reason for this is that the FWC guidance is trying to forecast temperatures in the 48-60 h range with model output that ends at 48 hours.

Table 2 shows that when SIG TEMP CHANGES occur, Albany local forecasters are able to make substantial improvements to guidance in the first and second periods. FEDS scores for the first and second periods, respectively, are 68.9 and 62.7, and the overall percent improvements over guidance are 7.2% and 8.2%. Even though Table 1 showed that for all four periods combined, the local forecaster improvement over guidance for the SIG TEMP CHANGES and the ALL categories was about the same, Table 2 reveals that first and second period improvements over guidance for the SIG TEMP CHANGES category are substantially higher than for the ALL category. After the second period, the ability of local forecasters to improve on MOS TEMP guidance when SIG TEMP CHANGES occur drops off rapidly. The TEMP FEDS scores are 14.1 and 1.7, respectively, for the third and fourth periods, and the overall percent improvements over guidance are 2.7% and 1.3%. Even though this still is a slight improvement over guidance, it

Table 2. Same as Table 1, except by forecast period.

	FEDS Score	% Improv Over MOS Guidance	# of Fcsts
ALL			
First Period	50.0	6.2	6082
Second Period	42.9	5.4	6080
Third Period	19.2	2.8	6080
Fourth Period	29.1	3.9	6078
RECORD PERIODS			
First Period	91.9	11.3	624
Second Period	60.1	6.5	624
Third Period	41.5	6.0	628
Fourth Period	55.2	8.0	628
RECORDS ONLY			
First Period	167.8	21.9	82
Second Period	161.8	18.6	81
Third Period	78.2	9.0	80
Fourth Period	61.6	7.9	79
SIG TEMP CHANGES			
First Period	68.9	7.2	1155
Second Period	62.7	8.2	1157
Third Period	14.1	2.7	1159
Fourth Period	1.7	1.3	1159
SIG PCPN EVENTS			
First Period	94.1	12.2	760
Second Period	60.5	6.9	760
Third Period	42.7	5.2	760
Fourth Period	32.7	4.7	760

is apparent that there is a substantial decrease in the ability of local forecasters, or reluctance on their part, to forecast SIG TEMP CHANGES beyond 36 hours.

Finally, for SIG PCPN EVENTS, Table 2 shows that the FEDS scores and overall percent improvements over guidance for the first, second, and third periods are similar to the scores for the RECORD PERIODS category. However, for the fourth period, the scores for the SIG PCPN EVENTS category did not increase as they did for the RECORD PERIODS category. Apparently, Table 2 suggests that for the first, second, and third forecast periods, local forecasters are able to make substantial improvements to TEMP guidance when SIG PCPN EVENTS occur, as readily as they do for those periods when record temperatures occur (RECORD PERIODS category).

c. *PoP verification results*

Table 3 is the same as Table 1, except it shows the PoP FEDS score and the overall percent improvement over MOS PoP guidance (the RECORDS ONLY category was not verified for PoP). Table 3 reveals that, for ALL PoP forecasts, the improvement over MOS PoP guidance by Albany local forecasters was small. The PoP FEDS score was only 6.4 and the overall percent improvement over guidance was 1.6%. Similarly, the improvement over MOS PoP guidance when SIG TEMP CHANGES and SIG PCPN EVENTS occurred were also small, and nearly identical to the results for ALL PoP forecasts.

Table 3. Same as Table 1, except for the PoP forecasts. The RECORDS ONLY category was not verified for PoP forecasts.

	FEDS Score	% Improv Over MOS Guidance	# of Fcsts
All	6.4	1.6	18360
Record Periods	20.1	3.3	1820
All	6.4	1.6	18360
SIG Temp Changes	5.7	2.1	3475
All	6.4	1.6	18360
SIG PCPN Events	7.3	1.4	2280

When record temperatures occurred, Table 3 indicates that Albany local forecasters were able to make a larger improvement over MOS PoP guidance. The PoP FEDS score for the RECORD PERIODS category was 20.1 and the overall percent improvement over guidance for this category was 3.3%. These scores are more than double the scores for ALL PoP forecasts, but still represent a relatively modest improvement.

d. *PoP verification results by forecast period*

Table 4 shows the PoP FEDS score and the overall percent improvement over guidance by forecast period for the ALL, RECORD PERIODS, SIG TEMP CHANGES, and SIG PCPN EVENTS categories. For ALL PoP forecasts, Table 4 indicates a decreasing trend in forecaster improvement over PoP guidance from the first to the third period, with most of the decrease occurring between the first and second periods. For the RECORD PERIODS and SIG TEMP CHANGES categories, the greatest FEDS score improvement over MOS PoP guidance by local forecasters occurs in the second period, and the second period improvement over guidance for the RECORD PERIODS category is substantial. For the first and third periods, three of the four improvements over guidance are slightly negative.

Table 4. Same as Table 2, except for PoP forecasts.

	FEDS Score	% Improv Over MOS Guidance	# of Fcsts
ALL			
First Period	15.6	3.1	6126
Second Period	4.1	1.2	6126
Third Period	3.4	0.7	6126
RECORD PERIODS			
First Period	11.9	3.6	610
Second Period	36.0	5.6	610
Third Period	-2.4	2.1	612
SIG TEMP CHANGES			
First Period	-0.3	0.4	1157
Second Period	18.1	1.6	1155
Third Period	-5.4	2.0	1157
SIG PCPN EVENTS			
First Period	29.9	6.8	760
Second Period	-9.9	-0.7	760
Third Period	-4.0	-0.3	760

When SIG PCPN EVENTS occur, Table 4 shows that local forecasters were able to make a substantial improvement to MOS PoP guidance for the first period, but did a little worse than guidance for the second and third periods.

4. Station Verification Trends

Figure 1 shows the seasonal combined TEMP/PoP FEDS score trend for the Albany forecast office from the 1990 warm season through the 1998-99 cool season. From the 1990 warm season through the first third of the 1993 warm season, the LFM-based FPC MOS guidance was used as the standard for comparison. During this time, the seasonal TEMP/PoP FEDS scores were relatively stable and generally between 60 and 75, with a sharp rise to over 130 for the 1992-93 cool season. For the last four months of the 1993 warm season, and for all seasons thereafter, the NGM-based FWC MOS guidance was the standard for comparison. Beginning with the 1993 warm season and the use of the FWC guidance, there was a substantial drop in Albany local forecast improvement over MOS guidance. This substantial decrease in local forecast improvement over guidance persisted through the 1995 warm season, and appeared to support the conclusion of Roebber and Bosart (1996) that there has been a continuous erosion of human forecast skill relative to MOS forecasts over the years, and that the convergence of human and machine forecast

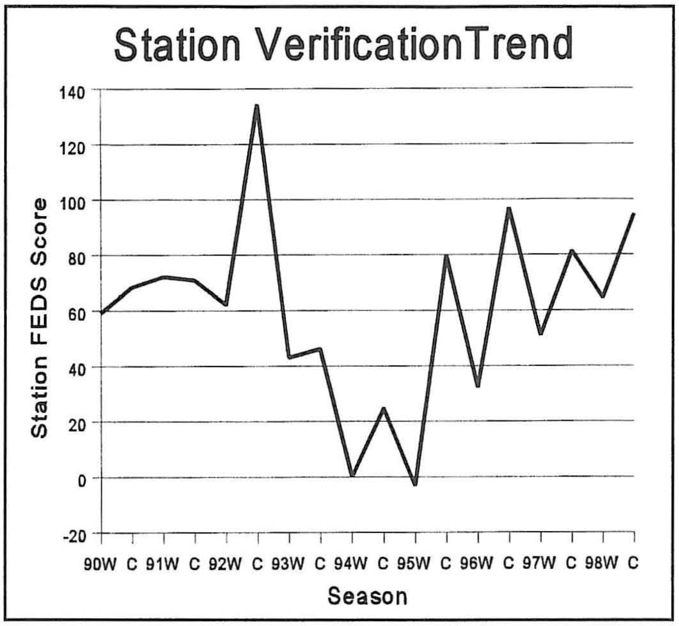

Fig. 1. The seasonal combined TEMP/PoP FEDS score trend for the Albany forecast office from the 1990 warm (W) season through the 1998-99 cool (C) season.

skill appeared to be inevitable. However, after the 1995 warm season, Fig. 1 shows that there was an increase in the ability of local forecasters to improve on guidance. In fact, the TEMP/PoP FEDS scores during the last seven seasons appear to be about as good, on average, as the scores during the FPC period. Also of note in Fig. 1, is the large seasonal dependence of local forecaster improvement over MOS since the FWC guidance began. Local forecasters make larger improvements over guidance for the cool season than for the warm season.

The FWC guidance, based on superior NGM output, was clearly better than the FPC guidance once there was a sufficient amount of data available for equation development. The lack of familiarity with the FWC guidance resulted in an immediate drop in local forecaster accuracy relative to MOS guidance. This was especially true for PoP forecasting. However, after about two years, Albany local forecasters began to increase their improvements over guidance and are now improving on the FWC guidance by as much as they did with the FPC guidance. Apparently, Albany forecasters have become familiar with the FWC guidance, its strengths and weaknesses. Second, newer and better models have become available since 1993, and forecasters are using these newer models to adjust the MOS forecasts accordingly. Third, the introduction of gridded model data fields into the forecasting process during the past few years has provided forecasters with the ability to examine model output in far greater detail than ever before, and to make adjustments to MOS guidance accordingly. These gridded data fields are used not only to examine the NGM output in great detail, but to examine and compare the output from the newer models as well. Finally, the increased emphasis on NWS forecaster training and development may have also contributed to the improvement.

5. Individual Verification Trends

The FEDS-based verification system began at the Albany forecast office during the 1990 warm season. At that time, there were seven forecasters at the office who had previous forecast experience and were familiar with forecasting for the Albany forecast area. These seven forecasters (six of the seven still are at the Albany forecast office) will comprise the group of forecasters who will be referred to as VETERAN forecasters. For each season in the sample, the average TEMP/PoP FEDS score of this group was calculated. The individuals who began forecasting at the Albany forecast office during or after the 1990 warm season, will be referred to as NOVICE forecasters. NOVICE forecasters may have had no previous forecast experience when they began forecasting at Albany, or they may have had previous forecast experience, but were not familiar with the Albany forecast area.

For NOVICE forecasters, the idea of an average seasonal score, as defined in this paper, has a different meaning than for VETERAN forecasters. Thus, the average seasonal TEMP/PoP FEDS score for NOVICE forecasters was calculated in a different manner than for VETERAN forecasters. The sample of NOVICE forecasters is made up of eleven individuals. Each of these forecasters began forecasting at Albany at a different time and the number of seasons each has forecast, varies. As a result, the average TEMP/PoP FEDS score for the first season of the NOVICE forecaster group was calculated based on the TEMP/PoP FEDS score each NOVICE forecaster had in his/her first season forecasting at Albany. The average

Table 5. The departure of the average VETERAN and NOVICE TEMP/PoP FEDS scores from the station TEMP/PoP FEDS score, for each season from the 1990 warm season to the 1997-98 cool season. Also shown is the smoothed 4-season running average of the departures. (The seasonal average scores for VETERAN and NOVICE forecasters were not calculated in the same manner, please see SECTION 5 of the text for further details). The number of forecasters included in the calculation of the average NOVICE departure for each season is shown in parenthesis. (Note: beyond season eight, there were only one or two NOVICE forecasters per season).

	Veteran		Novice	
Season	Departure	4-Sea Avg	Departure	4-Sea Avg
1	+3.8	+7.0	-42.0 (11)	-29.8
2	-0.4	+5.9	+6.2 (10)	-30.1
3	+34.4	+6.3	-67.4 (8)	-25.3
4	-9.8	-1.6	-16.0 (7)	-8.4
5	-0.5	+6.6	-43.3 (6)	+2.9
6	+1.2	+8.5	+25.4 (6)	+13.6
7	+2.8	+11.2	+0.2 (5)	+7.9
8	+23.0	+13.6	+29.3 (3)	+14.7
9	+7.0	+8.4	-0.6 (2)	-10.9
10	+12.1	+6.7	+2.5 (2)	-0.9
11	+12.4	+3.7	+27.5 (2)	+6.8
12	+2.2	-8.9	-73.0 (2)	-4.7
13	+0.2	-8.7	+39.5 (2)	+10.6
14	-0.2		+33.0 (2)	
15	-37.7		-18.3 (1)	
16	+2.8		-12.0 (1)	

score for the second season is based on each NOVICE forecaster's second season at Albany, etc. In addition, the number of NOVICE forecasters available to calculate the NOVICE TEMP/PoP FEDS score tends to decrease as the number of seasons increases. After eight seasons, there were only one or two NOVICE forecasters per season.

Table 5 shows the departure of the average VETERAN and NOVICE TEMP/PoP FEDS score from the station TEMP/PoP FEDS score, for each season from the 1990 warm season to the 1997-98 cool season. The departure from the seasonal station FEDS score was used so that seasonal variations in forecast difficulty could be accounted for. Thus, no matter how difficult (low station FEDS score) or easy (high station FEDS score) it was to improve on MOS guidance for any given season, the deviation from the station score by VETERAN forecasters could be expected to remain about the same over the long run. In order to smooth out the large fluctuations in verification scores that occur from season to season, the departures were smoothed by using a four-season running average and these results are also shown in Table 5.

The trend of the smoothed four-season running average of the departures for VETERAN and NOVICE forecasters is shown in Fig. 2. Figure 2 and Table 5 clearly reflect the so-called "learning curve" associated with NOVICE forecasters, and they also reveal that VETERAN forecasters show little overall trend during the 16-season period. Based on Fig. 2 and Table 5, NOVICE forecasters, on average, require between two and three years (or between four and six seasons) of forecasting experience before achieving their personal level of forecast accuracy, which will remain fairly constant thereafter. As shown in Fig. 2, the NOVICE TEMP/PoP FEDS score trend becomes similar to the VETERAN trend once enough forecasting experience has been gained. Even though there are only one or two NOVICE forecasters after season eight, and even though these two forecasters should be considered VETERAN forecasters after the sixth season, their scores were left in the NOVICE category so that the long-term general similarity of their scores to the VETERAN scores after six seasons could be evident.

Roebber and Bosart (1996) found that low experience forecasters required between 70 and 100 forecasts before their forecast errors decreased significantly relative to a consensus forecast. Taking into account the shift rotation and other forecast duties at the Albany forecast office, a time period of one to one and a half years would be required before a NOVICE forecaster made 70 to 100 forecasts. In Roebber and Bosart (1996), even after the 70 to 100 forecast period, the mean errors of the low experience forecasters generally were not yet equal to that of the high experience forecasters. Thus, the results of this study, which showed that a period of two to three years is required before NOVICE forecasters achieve the **same** level of forecast ability as VETERAN forecasters, appear to be similar to the results in Roebber and Bosart (1996).

Figure 3 shows the smoothed four-season running average of the TEMP/PoP FEDS score departures for individual VETERAN forecasters one, two, three, and four. Figure 4 shows the same thing, but for NOVICE forecasters one and two. Once again, the learning curve for NOVICE forecasters one and two is clearly evident in Fig. 4. For VETERAN forecasters, most of the time, there is little or no overall trend in their verification scores. For example, Fig. 3 shows that VETERAN forecasters one, two, and three each has approximately the same departure from the station TEMP/PoP FEDS score at the beginning of the sample as they do at the end of the sam-

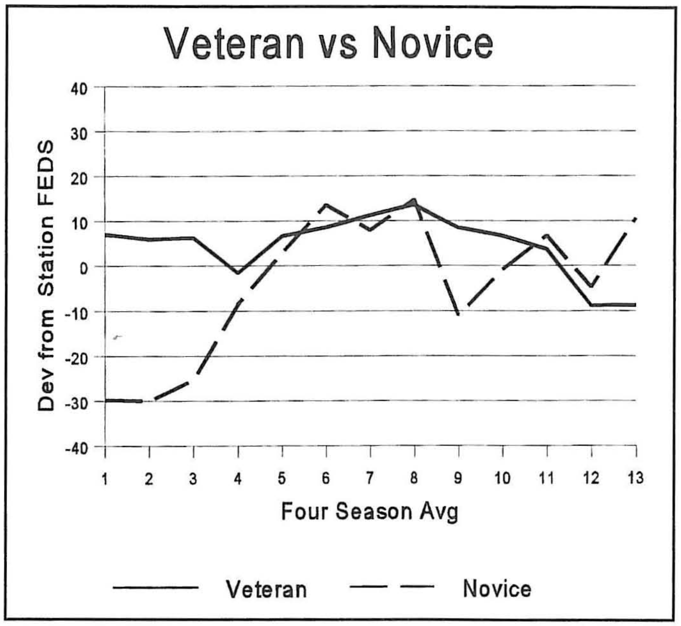

Fig. 2. The smoothed four-season running average of the VETERAN and NOVICE TEMP/PoP FEDS score departures from the station TEMP/PoP FEDS score for the 13 seasons listed in Table 5. (Please note: the departures for VETERAN and NOVICE forecasters were not calculated in the same manner, see SECTION 5 of the text for further details. In addition, beyond season eight, there were only one or two NOVICE forecasters per season).

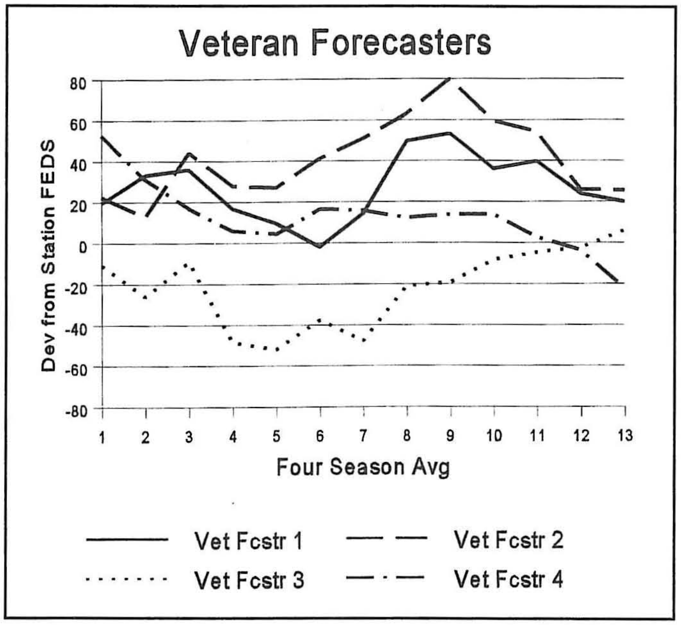

Fig. 3. The smoothed four-season running average of the departures from the station TEMP/PoP FEDS score for VETERAN forecasters one, two, three, and four.

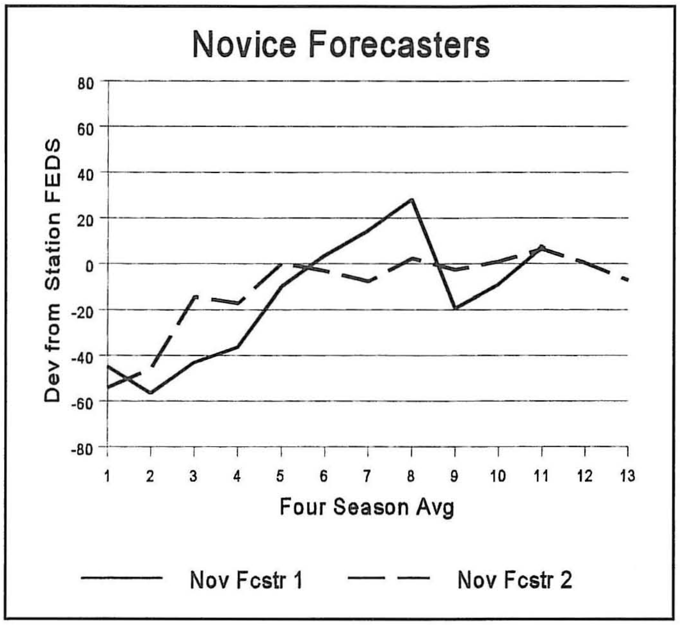

Fig. 4. The same as Fig. 3, except for NOVICE forecasters one and two.

ple, with sizeable fluctuations in between. However, there is no overall trend. VETERAN forecaster four does exhibit a declining trend in Fig. 3, but that trend occurs over a longer period of time and is not as sharp as the trends for NOVICE forecasters.

6. Discussion

Similar to Maglaras (1998), the results of the significant weather event part of this study showed 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. Maglaras (1998) showed a clear trend of increasing local forecaster improvement over guidance with increasing temperature departures from normal. In this study, when record temperatures occurred, local forecasters did progressively better as the verification approach focussed on the specific time interval during which the record event occurred. There was a progressive increase in improvement over TEMP guidance by local forecasters from the ALL forecast category, to the RECORD PERIODS category, to the RECORDS ONLY category. In fact, for RECORDS ONLY, based on nearly a decade of experience with the FEDS score, the improvement over TEMP guidance by local forecasters is exceptional. Even for PoP forecasting, the RECORD PERIODS category was the only category where local forecasters did noticeably better than for the ALL forecast category.

The fact that local forecasters were able to do better than MOS guidance when temperature conditions deviated substantially from normal should come as no surprise. It is well known that MOS guidance has difficulty with rare events (such as record temperatures), 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 (i.e., most of the ALL category), local forecast improvements over MOS guidance are reduced. In general, 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 comprised the developmental sample (as in most of the ALL category). The guidance will show a decreasing trend in accuracy as the weather conditions deviate further and further from this "normal range" (as in the RECORD PERIODS and RECORDS ONLY categories). These characteristics of MOS can be expected, even when future MOS developments occur based on more accurate numerical forecast models.

The improvement over MOS TEMP guidance by Albany local forecasters for SIG PCPN EVENTS was also substantial and nearly equal to the improvement over TEMP guidance for the RECORD PERIODS category. In addition, the improvement over MOS TEMP guidance when SIG PCPN EVENTS occurred was noticeably larger than for the ALL category.

Some of the other findings from the significant weather event section of this study include the following: when SIG TEMP CHANGES occur, in general, local forecaster improvement over TEMP guidance is about the same as it is for the ALL category. However, when broken down by forecast period, the verification scores revealed that local forecasters were successful at making significant changes to, and improving on MOS forecasts of TEMP, in the first and second periods, but in the third and fourth periods, there was a sharp decline in forecaster success. When SIG PCPN EVENTS occurred, the improvement over PoP guidance by local forecasters was small overall, and about equal to the ALL category. However, when broken down by forecast period, local forecasters made substantial improvements to PoP guidance for the first period, but did worse than PoP guidance for the second and third periods.

The station verification trend section of this study revealed the effect of the introduction of a new MOS guidance package on local forecaster ability to make significant changes to, and improve on MOS forecasts of TEMP and PoP. At first, there was a sharp decline in local forecaster improvement over guidance due to the superiority of the numerical model that the new MOS guidance package was based on, and because of local forecaster unfamiliarity with the new guidance package. However, after two years, local forecast improvement over guidance began to increase, and during the past few seasons, local improvement over guidance returned to the same level it was when the old MOS guidance package was the standard for comparison. This appears to contradict the conclusion of Roebber and Bosart (1996) who found that there has been a continuous erosion of human forecast skill relative to MOS over the years. As discussed earlier, the recent increase in forecaster improvement over guidance was the result of forecasters becoming familiar with the new guidance package, the use of newer and better numerical models, the introduction of gridded data fields into the forecast process, and, possibly, the increased attention to NWS forecaster training and development during the past few years. Looking into the future of the NWS, we expect that this cycle is likely to repeat itself. As newer and better models are introduced, MOS guidance will be developed by using the new models, and forecaster improvement over guidance may drop for a period of time when the new package is introduced. However, local forecasters can be expected to make substantial improvements to guidance (after about two years based on this study) as they become familiar with the strengths and weaknesses of the new MOS guidance package, and as they begin to use any new forecast tools that may become available to them after the introduction of the new MOS guidance package.

Even though the convergence of human and machine (MOS) forecast skill is real (Roebber and Bosart 1996), the convergence of forecast skill is due primarily to the superiority of newer MOS guidance packages compared to older packages. Since there usually is less error in MOS forecasts of TEMP and PoP for human forecasters to account for with each new MOS development, the magnitude of the improvement of human forecasters over MOS has decreased over the years and will continue to decrease in the future. However, human forecasters will always be able to use MOS as a base level of forecast skill, and then find methods to make improvements to this base level, especially for significant weather events as have been defined in this study. Thus, theoretically, the true convergence of human and machine forecast skill should not occur until there is **near zero** error in MOS forecasts.

The individual verification trend section of this study revealed the learning curve associated with NOVICE forecasters as they acquire forecasting experience and learn to forecast for a new area. On average, two to three years of forecast experience at a given location were required before forecasters reached their personal level of forecast ability, and, in subsequent years, they tended to remain near this level of forecast ability. These findings are similar to the comparison of low and high experience forecasters done by Roebber and Bosart (1996).

7. Conclusion

Hopefully, the findings of this study and those in Maglaras (1998) will be considered in any discussions to determine the future of NWS forecasters. During the past decade, and especially since the introduction of the FWC guidance and the associated decrease in forecaster ability to make improvements over guidance, there has been an apparent desire to migrate towards the automatic generation of most products. Experimental automated systems for forecasting (MOS guidance and computer-worded forecasts) currently perform very well during periods of near normal temperatures, during much of the warm season, and over relatively uniform terrain. However, this study indicates that local forecasters can perform much better during periods when the temperature deviates significantly from normal, and to a lesser degree, when major precipitation events occur. In addition, the lack of local forecaster improvement over guidance that was initially noted after the introduction of the FWC MOS guidance may have been temporary. This study has shown that forecasters at the Albany forecast office have been able to find methods to improve on MOS as regularly and effectively as they did when the FPC guidance was in use.

In the future, in order to find methods they can use to maintain their proficiency at making large improvements over guidance, local forecasters will still need to produce PoP and TEMP 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, or after a new MOS guidance package was introduced, would be diminished considerably, or, using the words of Roebber and Bosart (1996), there is the likelihood that human forecaster skills will atrophy with time unless they are used on a regular basis. Hence, the apparent trend to migrate towards the automatic generation of most products might need to be reexamined and modified in an appropriate manner.

Acknowledgments

I would like to thank Richard Conner, a student from the State University at Albany, for doing much of the research work involved with the significant weather event section of this study. I would also like to thank the staff of the NOAA/National Weather Service Office at Burlington, Vermont, for providing the Burlington climatological data used in this paper. Finally, I would like to thank all the forecasters (past and present) at the Albany forecast office who, without complaint, have allowed me to conduct research related to the accuracy of their forecasts for nearly a decade.

Author

George Maglaras has been a Lead Forecaster at the NOAA/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 NOAA/NWS Techniques Development Laboratory in Silver Spring, Maryland. Between 1986 and 1988 he worked at the NOAA/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

Carter, G. M., J. P. Dallavalle, and H. R. Glahn, 1989: Statistical forecasts based on the National Meteorological Center's numerical weather prediction system. *Wea. Forecasting,* 4, 401-412.

Dagostaro, V. J., 1985: The national AFOS-era verification processing system. *TDL Office Note 85-9*. National Weather Service, NOAA, U.S. Department of Commerce, 47 pp.

_________, and J. P. Dallavalle, 1997: AFOS-era verification of guidance and local aviation/public weather forecasts—No. 23. (October 1994-March 1995). *TDL Office Note 97-3*. National Weather Service, NOAA, U.S. Department of Commerce, 53 pp.

Dallavalle, J. P., and M. C. Erickson 1993: Using the National Weather Service's NGM-based statistical guidance for short-range forecasting. Preprints, *Thirteenth Conference on Weather Analysis and Forecasting*, Vienna, VA., Amer. Meteor. Soc., 44-47.

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.

_________, 1979: Computer worded forecasts. *Bull. Amer. Meteor. Soc.*,60, 4-11

Hoke, J. E., N. A. Phillips, G. J. Dimego, J. J. Tuccillo, and J. G. Sela, 1989: The Regional Analysis and Forecast System of the National Meteorological Center. *Wea. Forecasting*, 4, 323-334.

Jacks, E., J. B. Bower, V. J. Dagostaro, J. P. Dallavalle, M. C. Erickson, and J. C. Su, 1990: New NGM-based MOS guidance for maximum/minimum temperature, probability of precipitation, cloud amount, and surface wind. *Wea. Forecasting*, 5, 128-138.

Lowry, D. A., 1980: How to use and not use MOS guidance. Preprints, *Eighth Conference on Weather Analysis and Forecasting,* Denver, Amer. Meteor. Soc., 11-12.

Maglaras, G. J., and G. M. Carter, 1986: How to use MOS guidance effectively. Preprints, *Eleventh Conference on Weather Analysis and Forecasting*, Kansas City, Amer. Meteor. Soc., 17-22.

_________, 1991: A new verification scheme. *Eastern Region Technical Attachment No. 91-7B*, National Weather Service, NOAA, U.S. Department of Commerce, 5 pp.

_________, 1998: Verification trends at the Albany forecast office continue to show improvement on MOS guidance. *National Weather Digest*, 22:2, 9-14.

Murphy, M. C., and J. P. Dallavalle, 1984: An investigation of MOS minimum temperature errors in North and South Dakota during December 1982. *TDL Office Note 84-16*, National Weather Service, NOAA, U.S. Department of Commerce, 14 pp.

National Weather Service, 1983: The FOUS12 (FO12) bulletin. *NWS Technical Procedures Bulletin No. 325*, NOAA, U.S. Department of Commerce, 12 pp.

National Weather Service, 1992: NGM-based MOS guidance - the FOUS14/FWC message. *NWS Technical Procedures Bulletin No. 408*, NOAA, U.S. Department of Commerce, 8 pp.

Newell, J. E., and D. G. Deaven, 1981: The LFM-II Model-1980. *NOAA Tech Memorandum NWS NMC-66*, NOAA, U.S. Department of Commerce, 20 pp.

Roebber, P.J., and L. F. Bosart, 1996: The contributions of education and experience to forecast skill. *Wea. Forecasting*,11, 21-40.