Eastern Region Technical Attachment
No. 2004-05
October 2004

RELIABILITY TRENDS OF THE GLOBAL FORECAST SYSTEM MODEL OUTPUT STATISTICAL GUIDANCE IN THE NORTHEASTERN US: A STATISTICAL ANALYSIS WITH OPERATIONAL FORECASTING APPLICATIONS

John M. Goff
NOAA/National Weather Service Burlington, VT

ABSTRACT
Global Forecast System (GFS) Model Output Statistical (MOS) Guidance Probability of Precipitation (PoP) bias is examined for the northeastern United States, New England and Burlington, VT. Clear and distinct trends are identified in the data sets, with a mean positive bias noted across lower PoP categories (≤40%), and a mean negative bias across higher PoP (≥60%) categories. This is especially evident in the New England and Burlington, VT data sets. Possible causes of the observed lower PoP category bias are discussed, namely the coarseness in model resolution and the inherent design of the regional regression equations that drive the GFS MOS PoP scheme. Applications of the observed bias to operational forecasting techniques are then presented. It is argued that by adjusting forecast PoP values five to ten percent below GFS guidance across the lower PoP categories during the first three forecast periods, improvement over guidance may be realized in the long run. Due to good observed GFS MOS reliability (low bias) across the higher PoP categories, discreet adjustment of these values in either direction is not recommended.

1. INTRODUCTION

With ever increasing demands on meteorologists to produce highly detailed and more accurate forecasts, verification and performance measures within the National Weather Service (NWS) have gained more importance over the past several years. Forecasters rely on MOS (Model Output Statistics) guidance as a first guess in preparing both gridded and point based forecasts within the IFPS framework. Forecaster and MOS guidance performance is traditionally measured against observed data. Recently, the national verification program has adopted the Global Forecast System (GFS) model output statistics (MOS) as the standard against which all NWS forecasts will be measured (NWS 2003). One of the more difficult and transient of these forecasts is that of probabilistic precipitation forecasting (PoP). Local, regional, and national verification programs have often measured the skill of these PoP forecasts through use of the Brier Score (Brier 1950). An additional, but often useful measure of skill in PoP forecasting is bias, or reliability. Through the use of reliability trends, information can be gleaned which may persuade forecasters to adjust GFS MOS or other model guidance, and thereby improve gridded and point PoP forecasts in their area of responsibility.

A brief discussion of the statistical concepts of the Brier Score and reliability, and their importance in gauging PoP forecast performance is presented. The gathering of GFS reliability data and its treatment methods are then discussed, after which a detailed

examination of GFS reliability trends for the northeastern United States (U.S.) are presented for three forecast periods. GFS reliability performance for the smaller data sets of New England and Burlington, VT (KBTV) are also analyzed. Clear and consistent trends are noted through the use of reliability plots, with a distinct 5-10% positive bias at lower PoP categories, and a slight negative to no observed bias at higher PoP categories. These positive biases are more amplified during the cool season, and for the New England and KBTV data sets. Applications of the observed positive bias to operational forecasting are then presented, which may help lower respective PoP forecast bias across the northeast. This may be difficult to achieve, as discreet adjustment of point based PoP forecasts can be difficult process within the gridded forecast process.

2. GAUGING PoP FORECAST PERFORMANCE

a. The Brier Score

One of the more popular methods of measuring overall skill in PoP forecasting is the Brier Score (Brier 1950). The widely accepted definition of the Brier Score (BS) shown below in Equation (1) retains inherent statistical value in the meteorological community,

$$\mathrm{BS} = \frac{1}{n}\sum_{k=1}^{n}\left(y_k - o_k\right)^2 \qquad (1)$$

where BS is the Brier Score, n denotes the number of events, y is the PoP forecast expressed in decimal fashion from 0 to 1, and o is the observation, where o = 1 if the event occurs and o = 0 if the event does not occur[1].

The BS is analogous to the average squared difference between the forecast and observation pairs (Mean Squared Error) of the probability forecasts, and is thus a measure of **accuracy**. It is bounded by zero and one, and is negatively oriented with more accurate forecasts having lower Brier Score values (Wilkes 1995). Thus a BS = 0 has perfect accuracy and a BS = 1 has no accuracy. This method of measuring PoP forecast skill has several advantages, one of which is that it can be used as a comparative scheme among different forecast models. However, there are several aspects of the BS which could be considered a disadvantage, one of which is that it is does not indicate whether a forecast is inaccurate due to a wet or dry bias (AWS 1978). Thus, a forecaster could over forecast (wet bias) or under forecast (dry bias) the same event by the same margin, and receive the same score. As a result, it remains unclear whether the forecast had a positive or negative bias.

b. Reliability

Reliability is an equally valuable measure of PoP forecast skill in that it is a measure of **bias**. When statistical forecasts retain little or no bias over the entire range of possible forecast values, they are said to be reliable. In other words, it measures the forecaster's ability to accurately assign probability values (AWS 1978). An example of a reliability plot is given in Figure 1. In the example, a mean guidance PoP of 50 percent over a sufficient period should occur 50 percent of the time (i.e., perfect reliability or no bias). Consistent occurrences of perfect

[1]Equation 1 is actually only half of the Brier Score as originally introduced by Brier (1950). Thus the original score would equal twice that of BS above.

reliability in PoP forecasts are not frequently observed, with most forecasts falling either above or below the line of perfect reliability (black line in Fig. 1). The bias is calculated by measuring the vertical or ordinate distance between the line of perfect reliability and the observed value. Forecasts falling above the line would have a negative, or dry bias, while those below the line would have a positive, or wet bias.

In addition, a histogram showing the population in each PoP category bin often accompanies reliability plots (not shown). Thus bins with larger relative populations have inherently higher statistical value. This study addresses trends most evident across the lower PoP category bins ($0 < PoP \leq 40$), which all have much larger population sizes than those across higher PoP categories ($60 \leq PoP < 100$). This is to be expected as dry forecasts typically far outnumber wet forecasts. For example, in the two-year period of this study the total number of forecasts ranged from 10 to 12 thousand across the lower 10 and 20 PoP categories, to between 2 and 3 thousand across the 80 and 90 PoP categories.

The value in assessing PoP reliability trends in everyday NWS operations lies in the identification of statistical categories where trends of positive or negative bias are consistently observed in forecast PoP values. By noting these biases, meteorologists have a better measure of gauging their skill. A past study has indicated that a reasonable overall PoP bias lies within ± 5% of the forecast probability value (AWS 1978). By plotting reliability curves of relative observed frequency as a function of forecast probability, these trends are more readily seen. Statistical guidance may then be offered which may help forecasters lower their respective biases and corresponding BS. In turn, a discreet increase in forecaster PoP forecast accuracy may be observed over available MOS output.

3. DATA AND METHODOLOGY

a. Calculation of Reliability Data

GFS reliability data for available northeastern U.S. sites are compiled to identify possible trends that may persuade forecasters to re-examine PoP forecasting methodology across this region. The data set consists of 20 sites, which lie generally north and east of a line from Cleveland, OH to Washington, D.C., and are co-located with existing Automated Surface Observation System (ASOS) instrumentation (Fig. 2). All numerical forecast and observed PoP data is obtained from the internal NWS verification website, and is divided into three data sets for analysis: the northeastern U.S. (entire data set), New England, and Burlington, VT (KBTV)[2]. The PoP data is divided into categories, ranging from 0 to 100 percent at 10 percent intervals. The data is recorded for the first three 12 hour forecast periods (not shown). The total number of forecast and observed events for each 12 hour period are summed for each category, and then three period averages calculated, respectively. Reliability scores are then plotted for the two year period from October 2000 to September 2002 for each data set. Combined two year cool season (October through March 2000/01 and 2001/02) and warm season (April through September 2001 and 2002) plots during the same period are also analyzed using the above methodology in an effort to identify any seasonal trends inherent in the data.

b. Applicability to the Brier Score Scheme

[2]The New England data set consists of six sites: Boston, MA, Burlington, VT, Caribou, ME, Concord, NH, Portland, ME, and Providence, RI (see Fig. 2).

Mean GFS reliability trends identified above are then applied to the Brier scheme. Statistical PoP forecast techniques are then offered which may help forecasters lower respective biases and increase accuracy over available GFS MOS guidance. This is done through the use of the Brier Score nomogram (Fig. 3). The nomogram lists potential forecast BS points. A point is defined as the difference between forecast and model BS (multiplied by 100). Each box in the table contains two point values. The number on the left is the potential points gained over guidance, while the right number is the potential points lost. Boxes that contain an x occur where forecast and model PoP guidance are identical, thus no difference will be observed in either scenario. For example, if the guidance PoP for a given forecast period is 40 percent, and the forecast PoP is 50 percent, then the forecaster stands to gain 9 points if the event occurs, but lose 11 points if the event does not occur.

4. RESULTS

Analysis of the GFS reliability plots during the October 2000 through September 2002 period indicates two similar and noteworthy trends observed for all three data sets. These are a distinct and consistent positive bias (over forecasting) of lower PoPs ($0\% < PoP \leq 40\%$), and a slight negative (under forecasting) to no bias for higher PoPs ($60\% \leq PoP < 100\%$). It will be shown that these trends are also clearly evident during the subsequent two year cool and warm season plots.

a. October 2000 - September 2002 Data Set

Two year plots for the northeastern U.S. data set exhibit positive low PoP biases ranging from +3.7 to +6.6%, with a mean bias of +5.0% (Fig. 4). Slight negative biases are observed within the higher PoP ranges, with values between -3.5 and -5.6% observed, with mean bias values of -5.0% (Fig. 4).

Reliability curves for New England and KBTV echo the patterns noted in the larger northeast U.S. data subset, with a marked and amplified positive low category PoP bias, and only a slight overall negative bias across the higher category PoP ranges (Fig. 4). Observed New England low PoP biases ranged from +5.5 to +11.2%, with an observed mean bias of +8.6%. Errors for higher PoPs were similar to the northeastern U.S. average, showing a slight negative bias of between -2.3 and -5.3% and a mean of -3.8%. Similarly, KBTV observed low category PoP biases ranged between +5.3 to +11.8%, with a mean bias of +9.2%. Errors for the higher category PoP values where observed to be more variable than the other data sets, but showed less overall bias with values ranging from -3.0 to +3.6 % and a mean of -2.1%. There was also possible evidence of small sampling error for the 80 PoP category in the KBTV data, with a small discontinuity from negative to positive bias and curve smoothness noted at this value. However, upon examination of the KBTV data, the number of observed events at the 80 PoP category was similar in number to those of the other data sets at that value. Thus, the GFS PoP scheme appears to do better over the higher PoP ranges at KBTV than the overall New England or northeastern U.S. averages.

b. Cool Season Data Set

Analysis of the combined cool season reliability plots of 2000/01 and 2001/02 exhibit similar data set signatures as those for the two year combined plots, with distinct positive bias for lower PoP ranges and only a

slight negative bias for the higher category PoP ranges (Fig.5). In fact, of the three data set plots (two year average, two year cool season, and two year warm season), the cool season curves exhibit the greatest overall positive bias for the lower PoP ranges.

Examination of two year cool season plots for the northeastern U.S. indicated a positive low PoP bias ranging from +5.4 to +8.8%, with an observed mean of +7.6% (Fig. 5). Continuing earlier trends, only slight negative higher PoP biases were noted, with values between -2.8 and -5.2%. The mean bias was accordingly lower with a value of -4.0% observed.

As observed in the overall two-year data set, both the New England and KBTV GFS two year cool season reliability plots show clear positive bias for the lower PoP ranges, and only a slight negative bias for higher categories (Fig. 5). The low positive bias and corresponding mean for the New England plots exhibited the largest errors for any site grouping or time period, with values ranging from +7.1 to +16.2%, and a mean bias of +12.2% for the combined cool seasons. For the higher category PoP categories, these errors were less substantial and lower biases were observed. The overall bias ranged from -0.7 to -6.15%, with a mean bias of -3.4%. Similarly, the KBTV GFS data exhibit an overall positive bias for the lower PoP categories, ranging from +5.9 and 13.4% with an observed mean of +10.0%. The KBTV curve showed substantially lower bias for the higher PoP categories. Noted biases ranged from -4.5 to +0.7%, with a mean bias of only -1.9%, respectively.

c. Warm Season Plots

Analysis of the combined warm season plots of 2001 and 2002 continued the trends of the other data sets, with a positive bias for low PoP categories, and a slightly negative bias for high PoP categories (Fig. 6). The curves were also less amplified than those of the corresponding cool season (compare Fig. 5 and Fig. 6).

The curves for the data subsets of the northeastern U.S. and New England were very similar, with low positive biases ranging from 0% to +6.1% (for combined group), and mean values at +2.9% and +4.5%, respectively. At the higher PoP ranges, combined group negative biases ranged from -0.1% to -7.9%, with mean values of -5.6% and -3.6%, respectively.

Slightly larger amplification and bias were observed in the KBTV plots for the period, mainly across the lower PoP categories. Bias at these lower categories ranged from +4.8% to +9.8% with a mean of +8.2%. Though some variability existed in the curves at the higher PoP categories, especially at the 80 percent PoP value, overall reliability was observed to be good, with errors ranging from -2.8% to +9.4%, with a mean of -3.8%.

5. DISCUSSION

From the GFS reliability plots presented, it appears that the consistent and distinct trends noted are independent of temporal and/or spatial constraints. For example, GFS PoP biases for individual 12 hour forecast periods (not shown) were similar to those of the three period averages discussed above. At the higher PoP categories, the GFS MOS PoP scores are quite good, showing relatively low mean bias. The lower PoP categories exhibit more variation in bias compared with higher PoP ranges, with an overall higher mean bias. This may be a result of model resolution and/or inherent design of the GFS MOS PoP regional regression equations. Additional considerations include site location, quality control measures, and the

underestimation of frozen precipitation with existing ASOS instrumentation (Butler and McKee 1998).

a. Model Resolution

One problem that has been noted with global spectral models like the GFS is a lack of horizontal resolution on the regional scale when compared to other higher resolution models such as the MesoEta. This results in a lower topographical resolution in the GFS, which may result in inaccurate model quantitative precipitation forecasts (QPF). For example, the GFS model has a tendency to produce an overabundance of light precipitation amounts across a much larger regional domain than actually occurs. These events are often observed during the winter months across northern areas of the U.S. when numerous weak systems produce only trace(0.00" ≤ trace ≤ 0.01") to light precipitation (≤ 0.10") amounts (Hughes 1980). Given an example where both the MesoEta and GFS are correct in their respective placement of the QPF pattern, the GFS and its' more coarse topographical resolution will often produce light QPF across both mountain and valley locales. However, the MesoEta will more accurately represent the mesoscale detail of the QPF by resolving the adiabatic descent off the higher terrain, thus keeping valley locales dryer and correctly placing the higher QPF across the mountains. This is often observed in winter, particularly during northwest upslope snow events across Vermont. During these events, the GFS MOS PoP guidance for valley sites such as KBTV is often inflated across all PoP categories leading to a positive, or wet bias. Correspondingly, verification scores are less accurate, with higher Brier Scores noted. A step towards a possible solution to this problem would be to increase the topographical (horizontal) and vertical resolution within the GFS model to a level similar to operational models with higher resolution such as the MesoEta model.

b. GFS MOS PoP Regression Equations

Another possible explanation for the reliability trends observed in the data is the inherent design of the GFS MOS PoP regional regression equations. Within the GFS model, the country is divided into separate regions, with each region assigned a unique set of both warm and cool season PoP equations (Figs. 7 and 8). These equations are fixed by region, and season (warm or cool), and were developed for sites exhibiting similar climatology.

However, the regions for which the regression equations were developed are quite large, and the overall synoptic weather pattern in either season does not always represent the whole region. Thus the assimilation of multiple sites within one large region, and assigning a unique set of equations to govern those sites may be a contributing factor to the biases noted (Antolik 2000). For example, during the cool season from October through March (Fig. 8), note that portions of southern New York and northern Georgia are in the same region, or that northern Vermont is grouped with the upper peninsula of Michigan during the warm season (April through September; Fig. 7). This latter grouping is also grouped together during the cool season, which may in fact be a primary source of the consistent positive bias across the low PoP categories noted at Burlington, VT. Examination of the cool season region 10 clearly shows more sites grouped in the central to western Great Lakes than upstate NY and northern VT (Fig. 8). Thus, general synoptic and/or mesoscale weather conditions for this region may be substantially influenced

by moisture from the lakes, and may influence the GFS MOS PoP scheme in areas of northern New England which are further removed from this moisture source.

One possible solution to this problem is to adjust these PoP regions so that they are smaller geographically and contain only stations with similar climatologies. By doing so, the consistent low PoP biases observed may be reduced. However, there are limits to making the regions smaller, and doing so may threaten the stability of the regression equations in the GFS MOS PoP scheme (Antolik 2000).

c. Other Error Sources

Finally, other discreet error sources may exist that could explain the observed nature of the reliability plots. Site location could play a significant role due to the fact that most major observation sites, especially inland locations, lie in valleys near population centers as opposed to higher or elevated terrain. The atmospheric and topographical coarseness (low resolution) of the GFS scheme may be unable to resolve the distinct drier climatology that occurs at these sites due to adiabatic descent and other localized effects.

Another, but more obscure source of possible error lies in the accuracy of existing ASOS precipitation measurement techniques. This is particularly true in the cool season, when numerous light frozen precipitation events are known to occur. This may be due to the fact that the instrumentation lacks the sensitivity to record these as measurable events (≥0.01"), or related to other unknown physical limitations of the existing measurement system. There are potentially additional errors with ASOS precipitation measurements. For example, the national verification database is editable, and the quality control program at the Meteorological Development Laboratory (MDL) erases many hand ASOS edits that failed spatial consistency checks during the period of this study. Thus, the ASOS values in these cases were reset to zero. This was particularly true at sites in close proximity to the Great Lakes such as Syracuse, NY, and the adjustment of the ASOS values back to zero may have played a contributing factor to the noted dry bias observed.

6. APPLICATIONS TO OPERATIONAL PoP FORECASTING IN THE NORTHEASTERN U.S.

Using the trends discussed, future PoP forecasts in the northeast can be improved using the results of this study. Due to the consistent and marked nature of the positive bias exhibited by the GFS MOS PoP scheme across lower PoP categories and the findings of AWS (1978), it seems plausible that by using a forecast PoP value 5 to 10 % below the GFS PoP on a consistent basis, lower bias and higher overall accuracy will be observed over the model. This would appear to have the most value at the New England sites and for a majority of the northeastern U.S. sites used in this study (Fig. 2). As noted earlier, GFS PoP biases for the individual 12 hour forecast periods were very similar to the three period averages discussed above. For example, during the two year period from October 2000 to September 2002, the GFS forecasted a 30 PoP during the first 12 hour forecast period **113** times at Albany, NY. Only **11** cases of measurable precipitation were observed, resulting in an event occurrence of 9.7% (a positive, or wet bias of 20.3%). Applying the BS nomogram to this scenario, if the forecasters at Albany had lowered their PoP forecast to 20 % during these cases (10% below the GFS PoP), they would have lost

165 points but gained 510 points, a net gain of 345 points over the model during the two year period. Similar trends were noted for reducing the 30 PoP value at Albany for the second and third forecast periods as well. As stated above, these trends are consistently echoed in the data sets across much of the northeastern U.S., and particularly New England over the October 2000 to September 2002 time period for the first three forecast periods, regardless of season.

Understanding that the statistical probability of dry weather occurring during any given forecast period is inherently higher than an occurrence of wet weather, it natural to see that there will be a higher number of low PoP forecasts than higher ones. Thus it seems reasonable that adjusting these low PoP values slightly downward offers a higher potential gain in BS than adjusting higher category PoPs for two reasons. First, the GFS reliability scores were quite good at PoPs above 50% throughout the data sets. Due to the smaller sample size and lower overall bias for these data sets, discreet adjustment of these values in either direction is not recommended. Second, as the smaller sample size of wet events suggests, high PoP events are by nature more rare, thus the potential for losing considerable points for just a few random events exists. Table 1 illustrates potential 1st period BS points won/lost over GFS PoP guidance (New England data set for October 2000 to September 2002) when forecast PoP values are lowered by the recommended 5 or 10% from GFS values. The table also shows the mean observed frequency for each PoP category for the data set during the same time period. By using the suggested method on a consistent basis, the table illustrates potential gains actually outweigh potential losses over the long run assuming the identified trends continue.

The methodology of applying the results of this research is may be difficult to achieve within today's gridded forecast framework. The forecaster must initially populate the gridded database with the GFS MOS 12-hour PoPs, and adjust these values down 5-10% to achieve the desired affect. Additionally, with collaboration and consistency issues along the boundary of two or more forecast offices, the above recommendations may be difficult to apply.

7. CONCLUSION

Through the use of reliability data the GFS PoP scheme exhibits inherent and consistent positive bias (over forecast) at lower PoP values across the northeastern U.S. during the two year period from October 2000 through September 2002. Relatively good reliability and the associated negative bias (under forecast) were observed at higher PoP ranges during the same period. Coarseness of the GFS models' topographical resolution may be a contributing factor to the observed wet bias across the lower PoP values. Major observation sites are often located across lower elevations, such as river valleys. Thus the lower topographical resolution of the model may not resolve the drier climatology that occurs at these sites. Another plausible explanation of the observed low PoP wet bias lies in the design of the MOS PoP regional regression equations within the GFS. Regions of the country are defined by similar climatology, and a unique set of regional PoP regression equations. These regions may be too broad for the equations to accurately calculate PoP values on a consistent basis during longer seasonal time periods. Additionally, the accuracy and sensitivity of existing measurement equipment and quality control techniques may also have a negative impact on the reliability data. By applying the noted positive biases to day to day operational

forecasting techniques, recommendations in adjusting GFS MOS PoP guidance may be offered such that mean forecast PoP bias may be lowered, and overall accuracy increased over the GFS MOS PoP scheme. It is shown that this can be achieved by lowering forecast PoP values across the lower PoP categories by 5 to 10% from corresponding GFS PoP values during the first three 12 hour forecast periods. As noted at the end of section 6, this may be difficult within today's IFPS framework. Collaborative and/or consistency efforts may also be affected by discreetly adjusting the GFS MOS PoP values.

Further research is needed to determine whether the trends identified using the above process continue for longer time periods, are echoed across other regions of the country on both seasonal and multi-year time frames, and whether these results can be applied to gridded forecasts. Adjustment of physics parameterization within operational model suites, as well as occasional changes to model MOS equations could potentially make longer term comparisons more difficult. For example, if the time frame of interest encompasses a longer time period (i.e. multi-year), comparison of data of before and after these changes may not yield significant results, especially if the changes do indeed affect the results of the equations employed. However, the human forecasters' capability of improving PoP forecasts over existing GFS MOS and other model guidance justifies that future and more in depth examination of the trends identified in this study is relevant.

Acknowledgments

The author would like to thank Mark Antolik of the NWS Meteorological Development Laboratory for his help in providing guidance and expertise with the GFS MOS PoP scheme, and for maps of the GFS MOS PoP regional regression equation boundaries. Thanks also to Paul Sisson, Science and Operations Officer at the NWS in Burlington, VT for his overall guidance and oversight during this process.

REFERENCES

Antolik, M.S., 2000: An overview of the National Weather Service's centralized statistical quantitative precipitation forecasts. *J. Hydrology*., **239**, 306-337.

Air Weather Service (AWS) Pamphlet 105-51, 1978. Probability Forecasting: A Guide for Forecasters and Staff Weather Officers.

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

Butler, R. D., and T. B. McKee, 1998: ASOS heated tipping bucket performance assessment and impact on precipitation climate continuity. Climatology Report 98-2, Atmospheric Science Paper No. 655, Dept. of Atmos. Sci., CSU, Fort Collins, CO, June, 83 pp.

Hughes, L. A., 1980: Probability forecasting—Reasons, procedures, problems. NOAA Tech. Memo. NWS FCST 24, National Oceanic and Atmospheric Administration, 84 pp.

National Weather Service (NWS), cited 2004: NATIONAL WEATHER SERVICE INSTRUCTION 10-1601 VERIFICATION PROCEDURES. [Available online at http://www.nws.noaa.gov/directives/010/pd01016001a.pdf]

Wilkes, D.S., 1995. *Statistical methods in the atmospheric sciences: An introduction.* Academic Press, 467 pp.

FIGURES

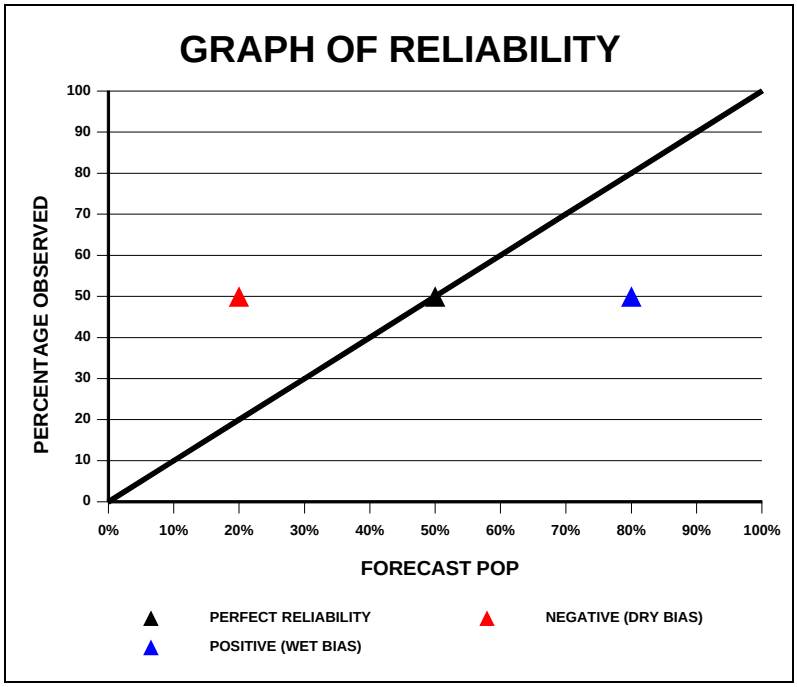

Figure 1. Reliability graph, indicating examples of positive (blue), negative (red), and zero (black) bias.

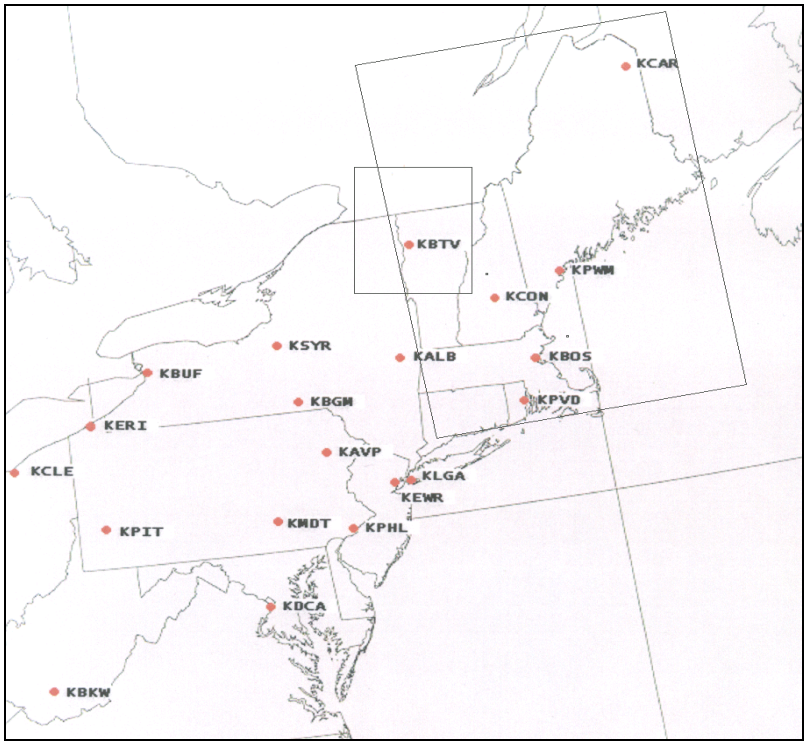

Figure 2. Map of observation sites co-located with ASOS instrumentation used in this study. Map highlights the entire northeast U.S. data set, and the smaller data sets of New England and Burlington, VT.

YOUR POP FORECAST	GUIDANCE POP FORECAST											
	0	5	10	20	30	40	50	60	70	80	90	100
0	X	0/10	1/19	4/36	9/51	16/64	25/75	36/84	49/91	64/96	81/99	+/+
5	10/0	X	1/9	4/26	9/41	16/54	25/65	36/74	49/81	64/86	81/89	+/90
10	19/1	9/1	X	3/17	8/32	15/45	24/56	35/65	48/72	63/77	80/80	99/81
20	36/4	26/4	17/3	X	5/15	12/28	21/39	32/48	45/55	60/60	77/63	96/64
30	51/9	41/9	32/8	15/5	X	7/13	16/24	27/33	40/40	55/45	72/48	91/49
40	64/16	54/16	45/15	28/12	13/7	X	9/11	20/20	33/27	48/32	65/35	84/36
50	75/25	65/25	56/24	39/21	24/16	11/9	X	11/9	24/16	39/21	56/24	75/25
60	84/36	74/36	65/35	48/32	33/27	20/20	9/11	X	13/7	28/12	45/15	64/16
70	91/49	81/49	72/48	55/45	40/40	27/33	16/24	7/13	X	15/5	32/8	51/9
80	96/64	86/64	77/63	60/60	45/55	32/48	21/39	12/28	5/15	X	17/3	36/4
90	99/81	89/81	80/80	63/77	48/72	35/65	24/56	15/45	8/32	3/17	X	19/1
100	+/+	90/+	81/99	64/96	49/91	36/84	25/75	16/64	9/51	4/36	1/19	X

Figure 3. Numerical Brier Score guidance of forecast PoP points gained/lost versus model guidance. (Cell definitions: points gained/points lost; x = no points gained/lost; + = 100).

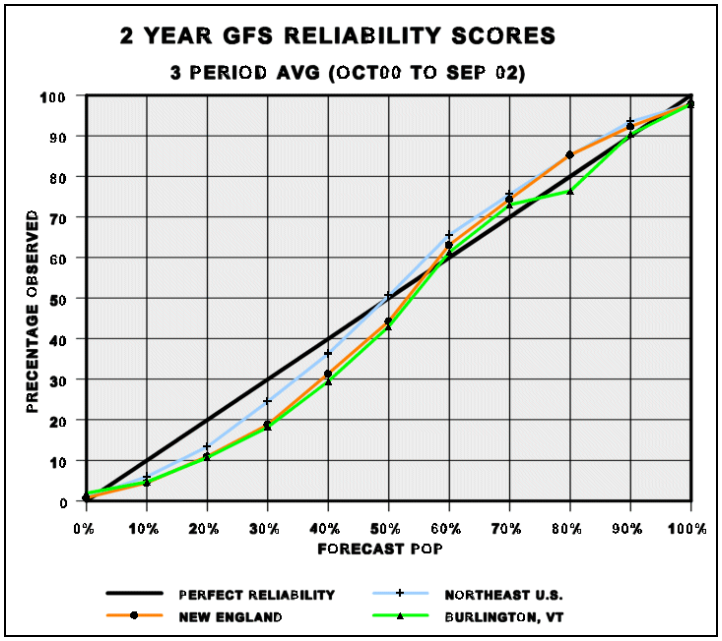

Figure 4. Reliability plots of GFS forecast PoP versus observed PoP occurrence during the two year period from October 2000 to September 2002. Plots are averages for the first three 12 hour forecast periods.

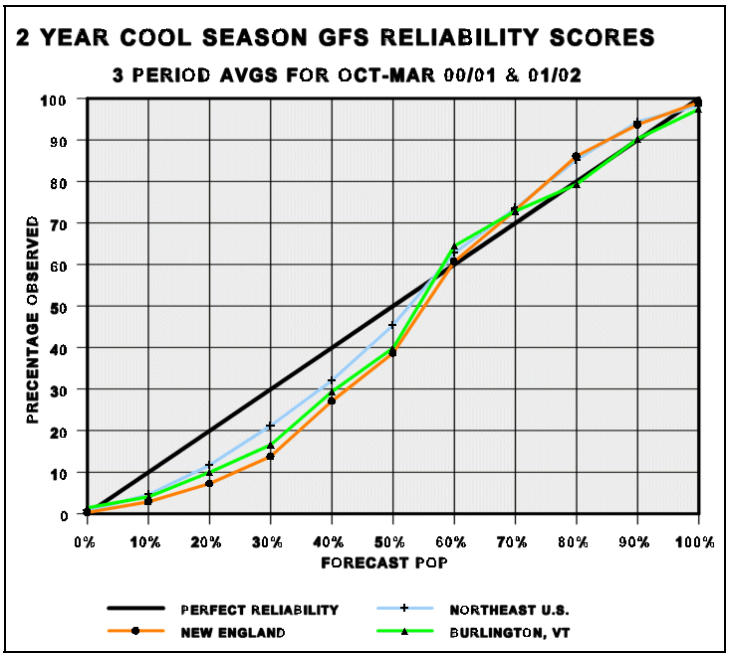

Figure 5. Reliability plots of GFS forecast PoP versus observed PoP occurrence during the combined cool seasons of October to March 2000/01 and 2001/02. Plots are averages for the first three 12 hour forecast periods.

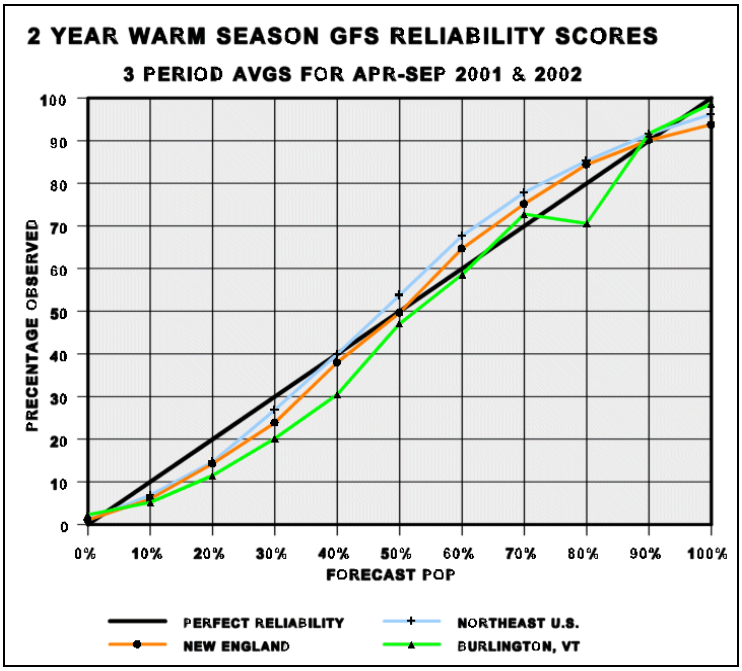

Figure 6. Reliability plots of GFS forecast PoP versus observed PoP occurrence during the combined warm seasons of April to September 2001 and 2002. Plots are averages for the first three 12 hour forecast periods.

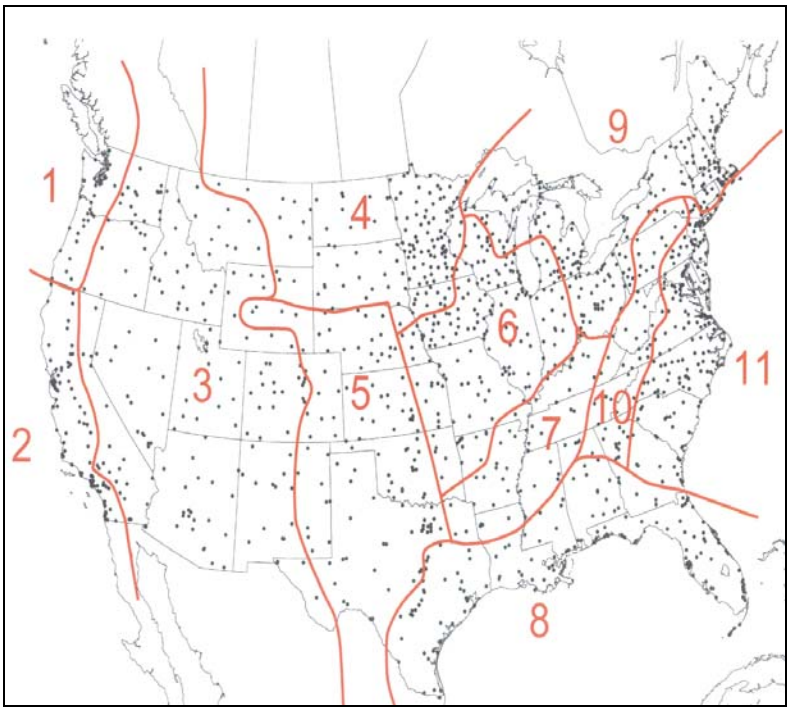

Figure 7. GFS PoP scheme for the US showing warm season regions (red) and dots (black) indicating station locations for which alphanumeric PoP guidance is available.

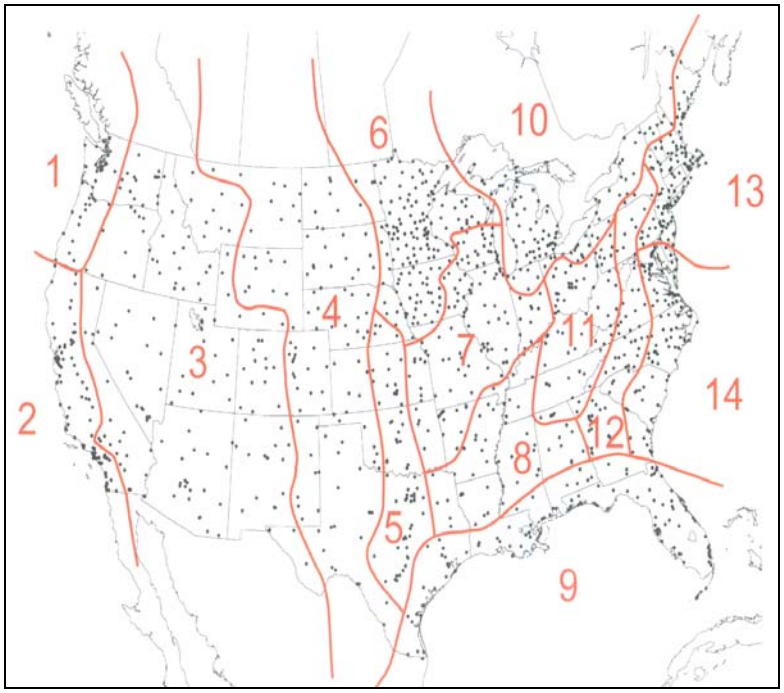

Figure 8. GFS PoP scheme for the US showing cool season regions (red) and dots (black) indicating station locations for which alphanumeric PoP guidance is available.

TABLES

Table 1. Illustration of potential 1st period Brier Score points won/lost over GFS alphanumeric PoP guidance in the two year period from October 2000 to September 2002 for New England sites. Data uses mean subset values obtained from NWS Verification Website.

GFS Guidance PoP (%)	Mean Observed Frequency (%)	Forecast PoP Lowered to (%)	Brier Score Points Gained/Lost over GFS
10	4.7	5	181/81
20	10.0	10	432/272
30	15.3	20	525/285
40	28.4	30	476/351