9.4 OBSERVING VERIFICATION TRENDS AND APPLYING A METHODOLOGY TO PROBABILISTIC PRECIPITATION FORECASTS AT A NATIONAL WEATHER SERVICE FORECAST OFFICE

Samuel L. Lashley*, L. Fisher, B.J. Simpson, J. Taylor, S. Weisser, and J.A. Logsdon
NOAA/NWS Northern Indiana Weather Forecast Office

A.M. Lammers
NOAA/NWS Louisville Weather Forecast Office

1. Introduction

Verification of weather forecasts has become a major focus and hot topic in recent years as National Weather Service (NWS) forecast offices look to make improvements against objective Model Output Statistics (MOS)(Glahn and Lowry 1972). Recent research and literature have shown that MOS output has improved over the years since its inception and its verification scores are now comparable or even better than local forecasts prepared by humans, especially at longer forecast projection times (Dallavalle and Dagostaro 2004; Baars and Mass 2005).

When discussing verification for Probability of Precipitation (PoP), the Brier score (Brier 1950) is a commonly accepted verification method for determining accuracy of PoP forecasts. The Office of Climate, Water and Weather Services (OCWWS) Performance Branch of the NWS uses one half the original score defined by Brier (1950) to compute Brier Score (BS) for select forecast points (NWS Directive 10-1601). Brier score is computed from 12-h PoP forecasts issued by NWS offices using the Point Forecast Matrix (PFM). Brier score is also calculated from computer model output, using 12-h MOS PoPs from the Global Forecast System (GFS) model (MEX MOS). Brier score is calculated for each 12-h forecast period contained in a typical 7-day forecast, which is issued routinely twice per 24-h period by each local NWS office. Comparisons of NWS and MOS PoP forecasts are then made for all forecast periods using the Brier score to determine the perceived accuracy and overall improvement of PoP forecasts made by humans over MOS.

The Brier score is a useful method for comparing and verifying a large number of PoP forecasts. This method is often used to compute Brier scores for forecast periods, calculated over rather long periods of time, such as months, seasons or even years. Numerous papers have been written on verification of precipitation forecasts since Brier first introduced the Brier score in 1950. Many of the more recent papers have focused on comparing human forecasts to those of MOS for 12-h projection times out through seven days and have concluded that human forecasts cannot make consistent improvements over MOS, especially at longer projection times (Roebber and Bosart 1996; Mass 2003; Baars and Mass 2005). However, these studies looked only at data within specific forecast projection periods and did not account for how each individual forecast verified in its entirety.

While the conclusions of these studies are considered valid and worthy of consideration, we believe human forecasters can and do add value to forecasts when compared to MOS if continuity and consistency over the course of a 7-day forecast are also considered as part of the verification system. We use the terms continuity and consistency to describe forecasts that have a nearly uninterruptable and coherent flow from one forecast projection period to the next over the course of a complete 7-day forecast cycle.

Few, if any, studies exist that have compared verification and continuity of human and MOS PoP forecasts over complete 7-day forecast cycles. In other words, if a forecast was issued with a 30 percent chance of precipitation in the day 7 projection period (144-168 hrs), how would this PoP change in the successive forecasts issued over the next 7 days leading up to the first period projection time (0-12h)? Does the PoP increase steadily or erratically over a 7-day forecast cycle for an expected precipitation event? Do MOS and human produced forecasts handle PoPs differently throughout a 7-day forecast cycle? Can verification scores for PoP be misleading? These are questions we are attempting to find answers to in this paper.

We believe continuity and consistency from one forecast to the next are as important as accuracy in measuring the quality of weather forecasts, and can add value for customers who routinely monitor daily forecasts. Continuity, consistency and forecast value are difficult concepts to measure objectively, but they have been deemed important in forecasts by many users of NWS local forecasts. A lack of continuity between daily PoP forecasts can imply uncertainty, especially if changes between successive forecasts

* *Corresponding Author Address:*
Samuel L. Lashley, National Weather Service, 7506 E 850 N, Syracuse, IN. 46567

e-mail: sam.lashley@noaa.gov

show relatively large swings in the PoP with no apparent trend. This uncertainty can translate into a lack of confidence in forecasts by customers who are trying to make definitive plans or decisions with as much lead time as possible. A lack of confidence in forecasts can lead to delayed decisions and added frustrations by customers.

Hurricane forecasts from The National Hurricane Center (NHC) are one example of how continuity can be used from one forecast to the next to instill confidence in the outcome of a forecast. The NHC philosophy of trending and gradually adjusting the forecast path of tropical systems from one issuance to the next rather than making dramatic changes has proven successful from the perspective of the public, emergency managers and media. When gradual but consistent changes are made to the forecasts, continuity is preserved and confidence in the forecast grows as a pattern is established. This methodology can allow customers to make earlier decisions as they recognize forecast trends over time and can anticipate with high confidence where the forecast may ultimately end up.

To illustrate our point further, we ask the following questions. Would NWS customers rather see a series of forecasts over a 7-day period that showed improvement over MOS but lacked continuity; or would they prefer a series of forecasts that showed continuity and converged toward the correct solution, but perhaps had a slightly poorer verification score compared to MOS? While those in the profession of meteorology and statistics understand the inherent uncertainty in PoP forecasts, the public generally may not. Therefore, when PoP appears to change inconsistently as projection time decreases (i.e. 50 percent on day 7, 30 percent on day 5, 70 percent on day 3), it can be argued that the public might perceive this as uncertainty and have little confidence in the actual PoP forecast. However, if a similar forecast showed continuity over time (i.e. 30 percent on day 7, 50 percent on day 5, 70 percent on day 3), the public perception might be one of greater certainty and confidence, leading to quicker decisions based on a higher confidence in the expected outcome.

We contend that continuity and consistency in PoP forecasts are important traits that can add value to forecasts, especially when compared to MOS. Continuity can also lead to high customer confidence in forecasts over time, which arguably is as important as verification when compared to MOS. This paper will look at 12-h PoP forecasts, their continuity and verification over 7 day forecast cycles. We will show examples where computed Brier score for 7-day forecasts may be better (lower) for MOS than NWS forecasts, but the continuity of NWS forecasts from one issuance to the next are better than MOS. This will be shown objectively using a relatively new forecast verification method, known as the Ruth-Glahn Forecast Convergence Score (FCS). This score does not measure the accuracy of a forecast based on its binary outcome (e.g. 1 for precipitation and 0 for no precipitation), but it does measure the continuity and convergence of forecasts from one projection time to the next over a 7-day forecast cycle. We will also discuss how forecast accuracy and forecast continuity can be combined into a methodology that can be used to provide more accurate and consistent forecasts to NWS customers while also improving collaboration and consistency among NWS forecast offices in the National Digital Forecast Database (NDFD) era.

2. Background

It can be argued that probability of precipitation (PoP) is one of the most difficult weather elements to forecast and verify. Brier (1950) noted that verification of forecasts has been controversial since the beginning of the 20th century. Brier (1950) clearly states that it is unsatisfactory for forecasters to try and hedge their forecasts or play the system in order to improve their verification. Brier writes, *"Numerous systems have been proposed but one of the greatest arguments raised against forecast verification is that forecasts which may be the "best" according to the accepted system of arbitrary scores may not be the most useful forecasts. In attempting to resolve this difficulty the forecaster may often find himself in the position of choosing to ignore the verification system or to let it do the forecasting for him by "hedging" or "playing the system." This may lead the forecaster to forecast something other than what he thinks will occur, for it is often easier to analyze the effect of different possible forecasts on the verification score than it is to analyze the weather situation. It is generally agreed that this state of affairs is unsatisfactory, as one essential criterion for satisfactory verification is that the verification scheme should influence the forecaster in no undesirable way."*

Brier clearly notes the argument that forecasts which verify best may not necessarily be the most useful. Several other papers have been written in recent years which attempt to show the relationships between accuracy and the value of forecasts (Murphy and Winkler 1987, Murphy and Ehrendorfer 1987, Murphy 1991, 1993, Brooks and Doswell 1996, Murphy and Wilks 1998). Murphy and Ehrendorfer (1987) showed that no single measure of performance can completely describe forecast quality or take into account the complexity or dimensionality of verification. Such a practice might lead to erroneous conclusions with respect to forecast accuracy (Murphy and Ehrendorfer 1987; Murphy 1991). We believe that continuity and consistency in daily forecasts are another dimensionality of verification that needs to be carefully considered.

Forecasters have often noted the lack of continuity in model forecasts from one forecast cycle to the next.

The Global Forecast System (GFS) model has been noted to be particularly inconsistent with its mass fields from one model cycle to the next, especially in the later projection periods and in the winter months. These inconsistencies between model runs are then transposed into MEX MOS, leading to large swings in PoP between successive forecasts for the same forecast period.

The MEX MOS has also been noted to be climatologically biased in its later projection periods (days 6 and 7), which often leads to routine PoP forecasts near climatologically normal values in these periods (15 to 35 percent in much of the Great Lakes region, depending on the time of year). These PoP values can often be misleading to human forecasters and result in an unnecessary addition of precipitation probabilities into public forecasts in the later projection periods. This is especially true when model mass fields may indicate relatively weak synoptic systems which may or may not produce precipitation.

After observing these inconsistencies in model data over the years, forecasters at the National Weather Service Northern Indiana Office (KIWX) began experimenting with different forecast concepts and philosophies in order to reduce the number of times probabilities were unnecessarily added to the forecast, especially in the longer forecast projection times (i.e. 84 to 168 hours). KIWX forecasters have also worked hard to make their PoP forecasts more consistent from one issuance to the next, trying to avoid what has been termed "flip flopping the forecast". The result has been a significant improvement in continuity from one forecast to the next, especially when compared to MOS. This will be shown objectively using the Ruth-Glahn Forecast Convergence Score. However, to follow this philosophy and maintain consistency and continuity over the course of a 7-day forecast, verification via Brier score comparisons to MOS may suffer slightly. We believe this slight decline in verification scores is justified given the remarkable improvement over MOS in the consistency scores and public perception of forecast quality.

Figure 1 is a simplified conceptual model of how we believe continuity in forecasts may affect customer confidence in forecasts (specifically with respect to PoP in this paper). In our model, we expect forecasts that exhibit little consistency (high variability) from one forecast to the next to be perceived by customers as forecasts with greater uncertainty and will result in lower customer confidence. In contrast, a forecast that is more linear over time (low variability) and trends toward a correct outcome will be perceived as having greater certainty and therefore higher customer confidence. While this conceptual model currently lacks significant scientific validation, its premise has been derived from discussions with regular users of NWS local forecasts and warrants consideration and future studies to prove or disprove this theory.

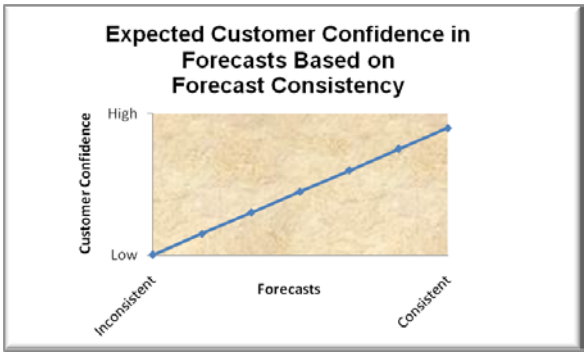

Figure 1. Simple conceptual model showing perceived relationship between forecast consistency and customer confidence in forecasts.

3. Methods and Data

In order to document the perceived inconsistencies and to compare MOS to NWS PoP forecasts, it was decided to locally collect and record MEX MOS and NWS PFM PoP forecasts for two locations within the KIWX Forecast Area (FA). The data could then be analyzed to determine if there was validity in the forecaster's perceptions of model and MOS inconsistencies and to also determine if there was merit to their forecast methodology of showing continuity in the forecast, even if it meant deviating from MEX PoP and possibly suffering some loss with respect to verification scores.

Complete 7-day 12-h PoP forecasts were collected from 1 January 2007 through 31 December 2007 for two locations, Fort Wayne Indiana (FWA) and South Bend Indiana (SBN). These sites were selected because they represent two major cities within the KIWX forecast area and both experience weather situations typical of the Great Lakes region. These sites also have long periods of climatological records and receive routine MEX MOS and PFM issuances. In addition, these two sites have 24-h augmented surface observations which compliment the Automated Surface Observing System (ASOS) observations. This ensures accurate precipitation measurements that will be used for forecast validation in this study.

The 12-h PoP forecasts representative of the objective model output were taken from the 0000 UTC and 1200 UTC MEX MOS, while 12-h PoP for the NWS subjective forecasts were taken from the routine issuance of the morning and afternoon PFM (400 am and 400 pm LST, respectively). This translated into 13 individual forecast projection periods for the 0000 UTC based forecasts; and 14 individual forecast projection periods for the 1200 UTC based forecasts

(the NWS only adds a new day 7 forecast period with the routine 400 pm LST forecasts).

The time periods used for verification in this project vary slightly from those used by the NWS verification branch. The 12 hour periods used locally were from 0000 UTC to 1200 UTC and 1200 UTC to 0000 UTC. The NWS verification branch uses 0600 LST to 1800 LST and 1800 LST to 0600 LST for verification periods (this is to account for different time zones and local diurnal periods across the United States).

In order to track and verify each forecast period's PoP, a spreadsheet was created where PFM and MEX PoP for each forecast period of each forecast issuance was entered manually then tracked automatically by the spreadsheet. A second spreadsheet used by the NWS Northern Indiana office to monitor 6 and 12 hour precipitation amounts at SBN and FWA for inclusion in daily climate products was used as the verification source for 12-h precipitation amounts.

Mathematical functions were performed on the forecast data within the spreadsheet to compute the change in PoP from one forecast period to the next over the course of each 7-day forecast issued in 2007. This computation allowed for the tracking of a PoP over the course of an entire forecast to see how much change actually occurred in both objective and subjective forecasts from one issuance to the next. As an example, a PoP issued for the projection period of 36 to 48 hours with the 0000 UTC forecast issuance (i.e. 40 percent) is subtracted from the PoP issued with the next 1200 UTC forecast for the projection period of 24 to 36 hours (i.e. 20 percent). This shows a decrease of 20 percent for the same forecast period from one forecast issuance to the next. Since we are only interested in the magnitude of change for this study, the absolute value of all changes between forecast periods will be shown.

Brier score and FCS were also computed for each 7-day forecast in 2007, which equated to 730 forecasts (2 forecasts per day, 365 days), and 9,855 forecast periods (365 forecasts with 13 periods, 365 forecasts with 14 periods) for each site in the study. By computing the Brier score and FCS for each 7-day forecast, a comparison could easily be made with regards to how each forecast verified with respect to accuracy and consistency.

4. PoP Analysis and Trends

Analysis and trends in these data for both SBN and FWA were very similar and revealed interesting traits. For brevity, references to these data will be with specific regards to FWA unless otherwise noted. Also, because we are focused on how a PoP forecast behaves throughout an entire forecast cycle, we will begin discussion with the later projection periods and work our way toward the shorter projection time periods.

In this paper, a PoP greater than 14 percent has been used as a threshold for when precipitation chances are considered significant and worthy of inclusion in the forecast. A PoP of 14 percent or less is considered a "dry" forecast. This is mainly due to NWS directives which require local NWS forecasts to include a "non-null" weather element for NDFD when any 12 hour PoP is greater than 14 percent (NWS Directive 10-506). Therefore 14 percent becomes an important threshold with respect to when precipitation chances and the associated weather type are added to local NWS forecasts. For the purpose of this paper, we will refer to any PoP greater than 14 percent as a "measurable PoP", as it indicates the chance for measurable precipitation.

a. Overall Trends and Analysis

Data collected during 2007 clearly showed two different methods by which MEX MOS and PFM PoP forecasts converged toward a solution over the course of 7-day forecasts. Measurable precipitation (0.01 inches or greater) occurred in just under 25 percent of the 730 possible 12-h day-7 periods. This is very close to climatology for the Great Lakes region. Over the course of 2007, MEX MOS showed a distinct bias toward climatology in its later projection periods, forecasting a PoP greater than 14 percent in just over 90 percent of its day-7 (period 13) forecasts. In contrast, PFM PoP forecasts showed a distinct conservative trend, forecasting a measurable PoP in just over 25 percent of the day-7 forecasts, much more representative of the actual number of precipitation events that occurred during the year.

Figure 2 shows the actual number of day-7 PoP forecasts greater than 14 percent for MEX MOS and PFM during each month of 2007. The actual number of occurrences of measurable precipitation (0.01 inches or greater) for each month is also shown. This figure clearly shows the distinct trend of MEX MOS to routinely forecast a PoP near climatology for a large

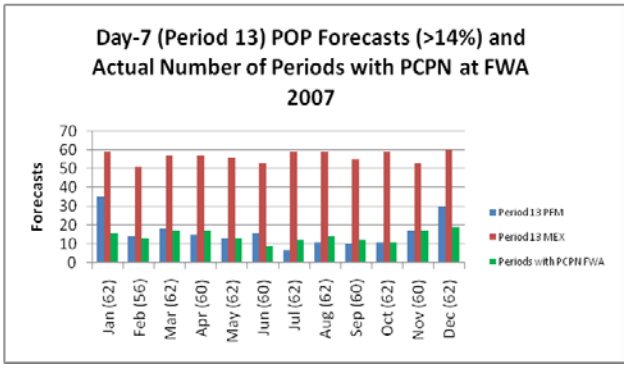

Figure 2. Number of day-7 forecasts by month with PoP greater than 14 percent. Numbers in parenthesis represent the total number of possible forecasts for each month.

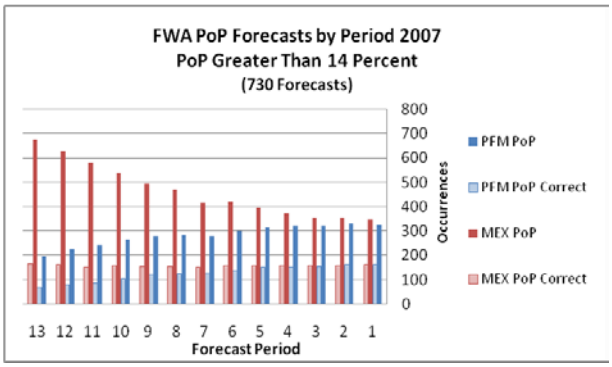

Figure 3a. PoP forecasts greater than 14 percent for each forecast period at FWA in 2007. The number of correct forecasts for each period are also shown (a correct forecast represents a PoP forecast that verified with measurable precipitation).

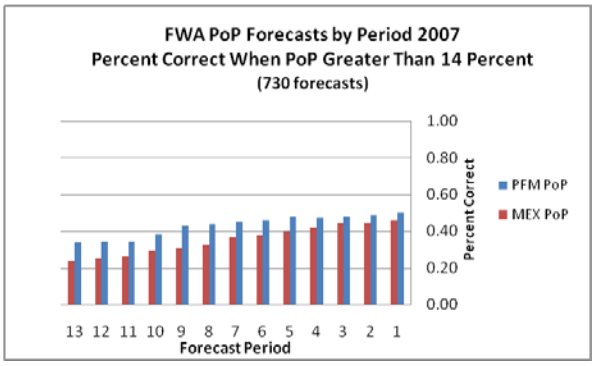

Figure 3b. Similar to figure 3a but shows the percent of correct PoP forecasts by period.

majority of its day-7 forecasts throughout the entire year. This is likely a function of the MOS equations picking up fewer of the direct model fields and adding a few of the geoclimate type predictors (sin and cos of the day of the year, elevation) and relative frequencies when available (personal email communication with MOS developer).

The human produced PFM forecasts appeared to add reasoning and heuristics over MOS forecasts in the later projection periods. Heuristics is a subjective method used to make forecast decisions in the presence of uncertainty and is explained in more detail in a paper by Doswell (2004). Forecasters recognized the MOS bias and were much more conservative with the introduction of measurable probabilities into the day-7 forecast period. The result can clearly be seen in nearly every month where the actual number of periods with measurable precipitation is very close to the actual number of day-7 forecasts where NWS forecasters introduced a measurable PoP.

While MEX MOS and PFM PoP forecasts differed in the day-7 period, both did show a distinct trend to converge toward similar solutions over the course of 7-day forecasts. Figure 3a shows the number of PFM and MEX MOS PoP forecasts which were greater than 14 percent for each forecast period in 2007 while figure 3b shows the percent of correct forecasts for each period. The day-7 differences discussed previously are clearly present in figure 3a, but a distinct trend can be seen as projection time decreases. The number of cases where MEX MOS is greater than14 percent gradually decreases over the course of 7-day forecasts while the number of PFM forecasts above this threshold increases slightly. Beginning near period 5 (forecast day-3) and continuing through period 1 (forecast day-1), MEX MOS and PFM PoP forecasts become very similar. The MEX trend is a result of it keying in more on the model output fields and less on the geoclimate predictors. Meanwhile, PFM forecasts in day-7 tend to favor the higher probability outcomes of no

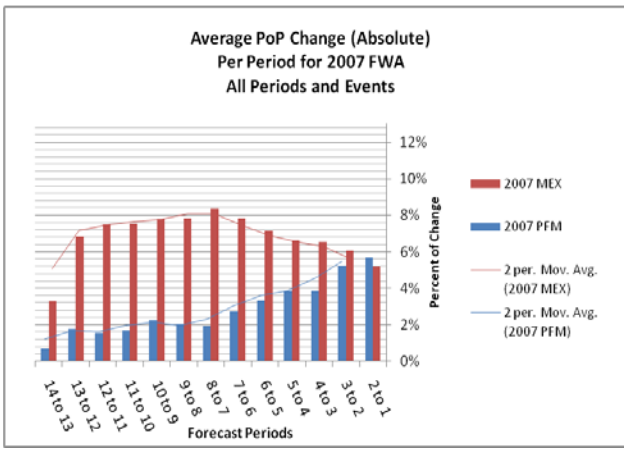

Figure 4a. Represents the average change in PoP (absolute) from one forecast period to the next over the course of 7-day forecasts in 2007.

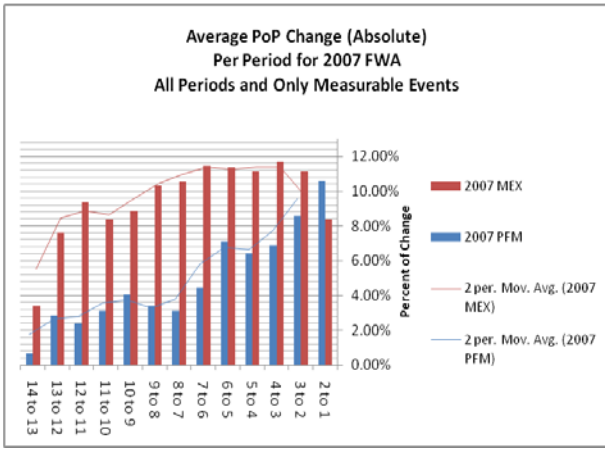

Figure 4b. Similar to figure 4a but represents average change in PoP when measurable precipitation occurred.

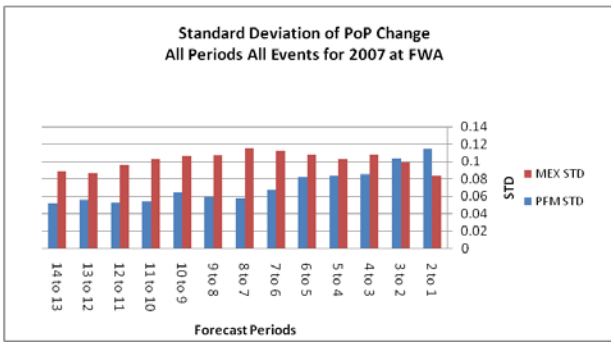

Figure 4c. Standard Deviation of PoP change for all forecast periods.

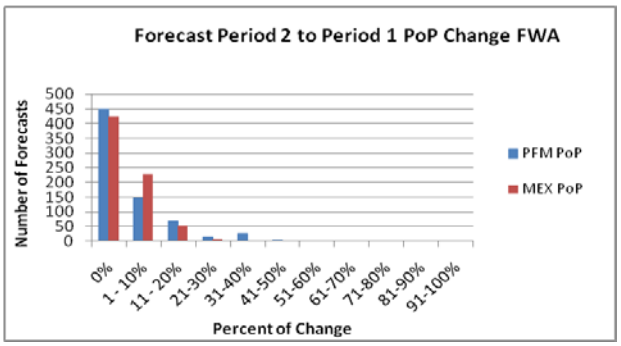

Figure 5. Amount of change in PoP from forecast Period 2 to period 1 for all 2007 forecasts at FWA.

measurable precipitation, and begin keying in on increasing model agreement and consistency, MOS trends, and remote sensing tools, which leads to more refined and accurate PoPs as projection time decreases and confidence increases.

In addition to showing the general trend of PoP forecasts by period, the amount of change in a PoP from one forecast period to the next was tracked over the course of all 7-day forecasts in 2007. Figures 4a and 4b show a comparison of absolute PoP change from one forecast period to the next between successive forecast issuances. Figure 4a represents all forecasts while figure 4b shows changes only when measurable precipitation occurred. The variability of MEX MOS PoP over the course of a forecast can easily be seen. The MEX MOS averaged a change of between 6 and 8 percent for all forecasts (figure 4a) and changed between 8 and 12 percent when precipitation actually occurred (figure 4b). This is significant because it shows a lack of consistency and continuity between forecast issuances. One can argue that changes greater than 10 percent could be perceived by customers as significant, especially when they occur over several successive forecasts and do not trend toward a common outcome. This scenario has been termed as a “flip flop” in the forecast, which creates greater uncertainty than what the actual PoP of one forecast may imply.

PFM PoPs show a distinct trend in both figures 4a and 4b that represents consistency and continuity over the course of 7-day forecasts. On average, only gradual changes are made to the PoP between successive forecasts in the later projection periods. This is followed by a continual and gradual change to the PoP as the forecast projection time decreases. This reflects the human forecaster’s uncertainty at longer lead times, but increasing certainty and confidence in the forecast outcome as the forecast time period nears. On average for all forecast issuances, PFM forecasts changed by less than 6 percent. For precipitation events, PFM forecasts generally changed by less than 6 percent from forecast day-7 (period 13) through day-3 (period 7), but then showed an average increase of between 6 and 12 percent from day-4 (period 7) through day 1 (period 1) .

Figure 4c shows the standard deviation of the percent change for PoP from one forecast to the next. MEX standard deviations are nearly twice those of PFM from near day-6 (period 12) through day-3 (period 6), indicating the higher variability of MEX MOS compared to PFM. The PFM PoP change does show a slightly higher standard deviation in day-1 (period 2 to period 1), which is attributed to forecasters making final refinements of PoP toward zero or 100 percent. Figure 5 shows how much PoP forecasts actually changed from period 2 to period 1 at FWA. The majority of forecasts only showed a change in PoP of 15 percent or less, and only a few forecasts had a change in PoP by more than 25 percent from the second to first forecast period.

b. Reliability Statistics

The early forecast projection periods (those closest in time to the actual observation) are the most critical periods to forecast correctly. It has been proven using reliability statistics that human forecasters and models show the most forecast skill in the early periods (Baars and Mass 2005). Humans are especially better in early periods likely due to superior graphical interpretation and physical understanding as well as the ability to communicate with various user communities all lending toward better meteorological analysis and forecasting (Baars and Mass 2005).

For the PoP weather element, the reliability statistics are based off of the principal that a 20% PoP should verify 20% of the time, a 40% PoP verifies 40% of the time, and so on to achieve “perfect reliability.” While both PFM and MEX MOS forecasts are not “perfect” due to the limited human knowledge of the atmosphere at the current time, early forecast periods have been shown to have reasonably good reliability by recent studies (Baar and Mass 2005). For example, by examining the reliability statistic for the first 4 periods at FWA in Figures 6a-d, one can observe that MEX and PFM have the best reliability in

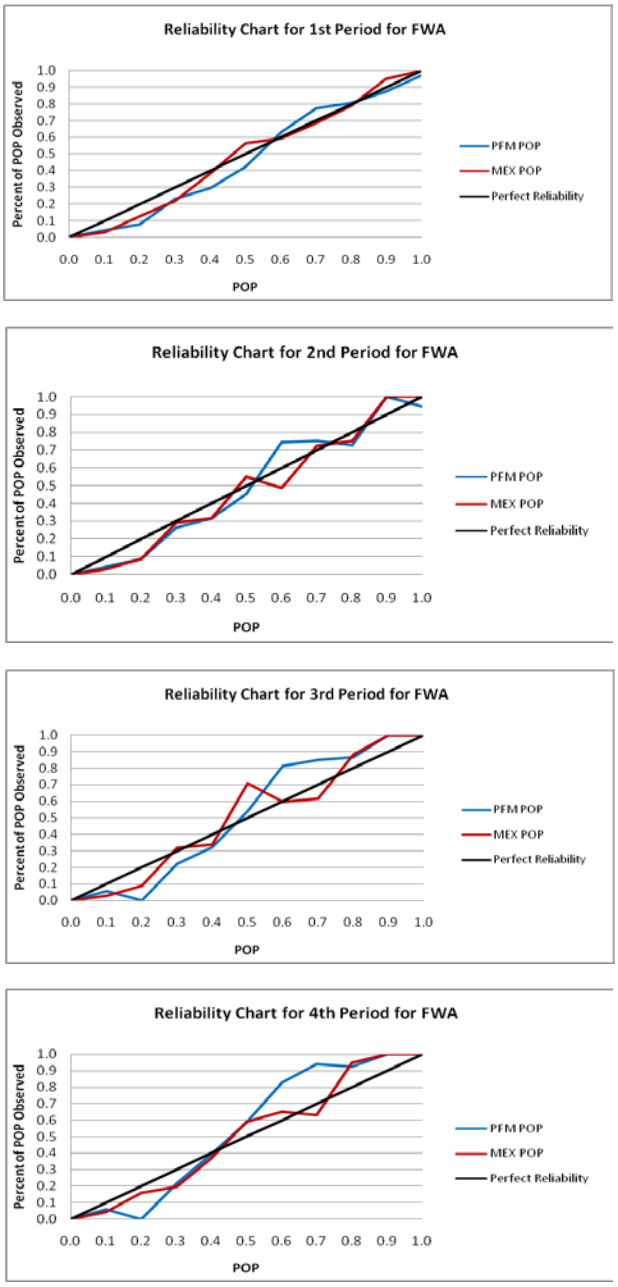

Figure 6a-d. Reliability charts for first 4 forecast periods at FWA.

the first 2 periods by their close proximity to the perfect reliability line. As one goes out further in time toward periods 3 and 4, the reliability becomes more variable, especially for PFM.

Generally it appears that both MEX and PFM forecasts for low PoPs exhibit an over prediction for periods 1 through 4, similar to results found in Baars and Mass (2005). Periods 3 and 4 at FWA show that PFM forecasts were under-predicted for higher PoPs. This factor reveals the conservative forecast approach of the NWS Northern Indiana office, which tends to begin with a low PoP in days 4-7 and gradually increase to a high PoP toward day-1 for precipitation events as forecaster confidence grows. This approach leads to higher customer confidence in the forecast over time, but also leads to less than perfect reliability. SBN reliability charts in this study contained the same aforementioned forecast trends despite a slightly different systematic bias.

5. Specific Forecast Comparisons and Verification

It has been shown that MEX MOS trends toward climatology in its later projection times while PFM trends toward the higher probability events of no precipitation. Both forecasts generally trend over the 7-day forecast cycle toward similar solutions. It has also been shown that MEX MOS PoP changes significantly from one forecast period to the next, showing little consistency, while PFM PoP forecasts tend to systematically make greater changes as forecast projection time decreases. Now we will look at how these factors come into play in a typical 7-day forecast and how verification statistics can be misleading.

Between 0000 UTC and 1200 UTC on 21 August 2007, a record 3.20 inches of rain was measured at site FWA. As a result of this convective thunderstorm event, widespread flooding occurred across a large part of northeast Indiana and northwest Ohio, including the devastating flooding on the Blanchard River in northwest Ohio on 21-23 August 2007. Near record crests occurred in Findley Ohio on 22 August 2007 and in Ottawa Ohio on 23 August 2007. Most of these two towns were inundated with flooding for several days and were the focus of national media attention.

Figures 7a and 7b represent the actual forecasts made over the course of 13 periods for the 0000 UTC to1200 UTC 12-h period of 21 August 2007 for site FWA. Figure 7a represents the KIWX PFM PoP forecasts while figure 7b shows the MEX MOS PoP forecasts from each successive 0000 UTC and 1200 UTC model cycle leading up to this event.

The trend of both forecasts showed increasing probabilities over the 7 day forecast cycle. However, the MEX MOS forecast in figure 7b revealed an inconsistent trend with occasional decreases in PoP. The most dramatic decline occurred from forecast period 5 to period 4 where the MOS PoP dropped from 61 to 39 percent. This was followed in the very next MOS issuance by an increase from 39 to 78 percent from forecast period 4 to period 3. Meanwhile, the PFM PoP remained consistent through these declines, eventually increasing the PoP after several additional forecast issuances. This stair-step approach showed forecasters improving over MOS by using heuristics and their knowledge of model limitations and trends. This approach led to a more

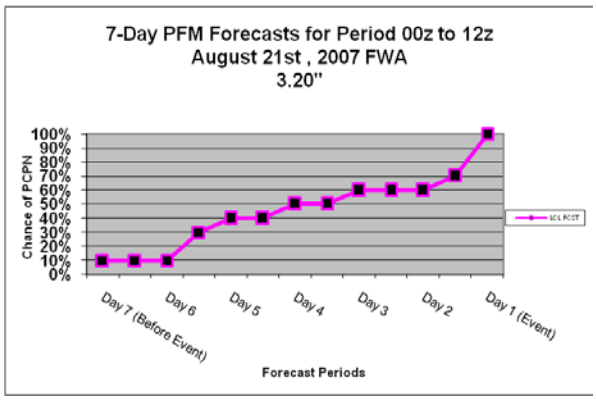

Figure 7a. Actual 7-day PFM PoP forecast for the period 0000 UTC to 1200 UTC 21 August 2007.

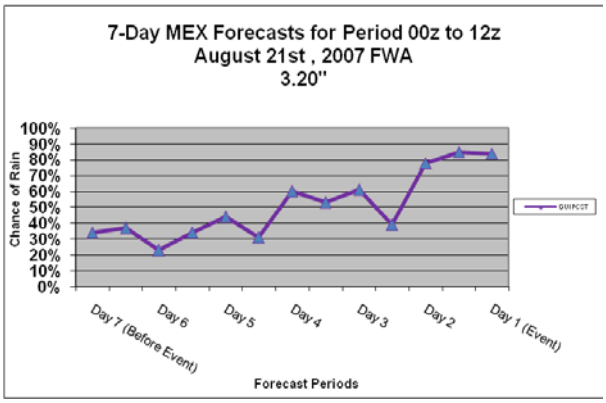

Figure 7b. Actual 7-day MEX PoP forecast for same period as figure 7a.

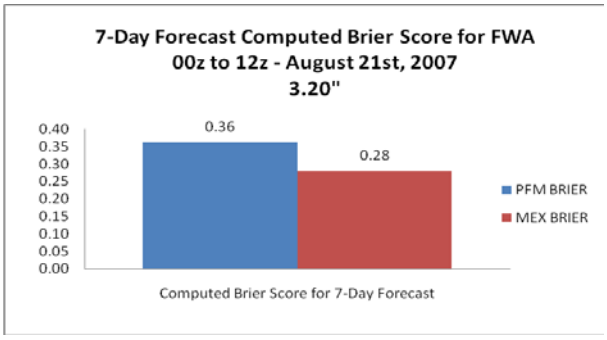

Figure 8. Computed Brier score for PFM and MEX PoP forecasts for the 12-h period 0000 UTC to 1200 UTC on 21 August 2007.

consistent forecast that promoted confidence to customers.

When casual observers look at these two forecast graphs, figure 7a is most often chosen as the forecast they would like to see for precipitation events, despite the initial 10 percent PoP in day 7. However, figure 8 shows the computed Brier score over the length of the 13 forecast periods for the PFM and MEX MOS forecasts for this event. The MEX MOS had a lower Brier score (more accurate) than the PFM by 8 percent, even though the PFM appeared to be a more reliable and consistent forecast. This is one of many 7-day forecasts in our study that showed MOS as having a better verification score compared to NWS PFM, but yet revealed very inconsistent forecasts over the course of the 7-day forecast cycle. These are examples where a single verification method may not necessarily reflect the true value of a forecast (Murphy and Ehrendorfer, 1987). Other methods of assessing forecast value need to be looked at, especially with respect to how our customers are using forecasts. We will now introduce a method by which forecast value and consistency can be measured quantitatively to show value relative to a complete forecast over time.

6. Forecast Convergence Score (FCS)

The Ruth-Glahn Forecast Convergence Score is a new forecast verification score recently developed by David Ruth and Harry Glahn. The FCS was designed to measure forecast convergence toward the end result or observation (D. P. Ruth, personal communication, November 7, 2007). It accomplishes this task by comparing the number and magnitude of "forecast swings" (significant changes to a forecast defined by a user selected threshold) within a series of forecasts. While this score is useful in verifying the continuity of a forecast, it does not directly measure forecast accuracy (D. P. Ruth, personal communication, November 7, 2007). Despite this minor drawback, the FCS looks to be very promising in providing another forecast verification tool and has recently been adopted for use in NDFD to measure forecast consistency among MOS and PFM outputs.

The FCS calculates forecast convergence by examining the number of large swings (changes) over a forecast period. To calculate the FCS, assume N forecasts are made for a series of decreasing projections (F_i where i =1 to N) prior to a single valid time where there is an observed value (Ob). Assume F_1 is the furthest projection and F_N is the nearest to the Ob. Also assume a significance threshold (SigT) which specifies the minimum change necessary to count as a swing.

The FCS is defined by the following formula (D. P. Ruth, personal communication, November 7, 2007):

$$FCS = \frac{T1 + T2}{T3 + T4}$$

Where T1, T2, T3, and T4 are defined as follows:

T1 = the number of forecasts (F_2 through F_N) that changed insignificantly (no more than SigT) from the previous F_{i-1} forecast OR moved closer to the next forecast F_{i+1}. When i = N, the Ob is used for F_{i+1}.

$$T2 = \frac{|F_N - F_1|}{SigT}$$

$$T3 = N - 1$$

$$T4 = \frac{\sum_{i=2}^{N} |F_i - F_{i-1}|}{SigT}$$

The T1 and T3 terms account for the number of swings. The T2 and T4 terms account for the magnitude of the swings.

The FCS ranges in value from 0.0 to 1.0 with a score of 0.0 corresponding to a forecast with many large swings and no convergence toward the observation, and a score of 1.0 corresponding to a forecast with no large swings and all forecasts converging toward the observation. The FCS is to be used on a continuous set of forecasts for a 7 day forecast period (D. P. Ruth, personal communication, November 7, 2007).

The FCS brings a new dimension to forecast verification—forecast consistency. The FCS is used in this study to illustrate that while MEX MOS may improve over PFM statistics using Brier score verification for a particular event, MEX MOS does not necessarily improve over PFM statistics using FCS verification for the same event. For example, figure 9 shows that during the 21 August 2007 case at FWA referenced earlier in this paper, the FCS for the PFM (1.00) was significantly greater than the FCS for the MEX (0.67). This was computed using a significance threshold (Sig T) of 20 percent. In this example, MEX MOS gave a slightly more accurate forecast statistically (MEX Brier Score improved over PFM Brier Score). However, humans gave a much more consistent forecast with continuity and value (PFM FCS improved over MEX FCS). The added value of consistency and continuity combined with a relatively good Brier score leads to a much better forecast than an inconsistent MOS forecast with a slightly better Brier score.

To investigate further, figure 10 shows the average monthly Brier and FCS scores for FWA in 2007. This graphic conveys that while the PFM and MEX had very similar Brier scores on a monthly basis, the PFM FCS consistently improved over MEX FCS during every month in 2007. In fact, the 2007 yearly averages (Table 1) reveal that the PFM and MEX held nearly the same Brier scores at each location while the FCS scores differed by an average value of 0.16 at SBN and 0.15 at FWA. This implies that for the year 2007 as a whole, model and human forecasts

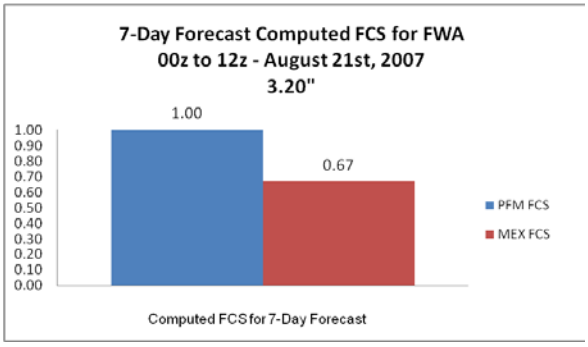

Figure 9. Computed Forecast Convergence Score (FCS) for PFM and MEX PoP forecasts for the 12-h period 0000 UTC to 1200 UTC on 21 August 2007.

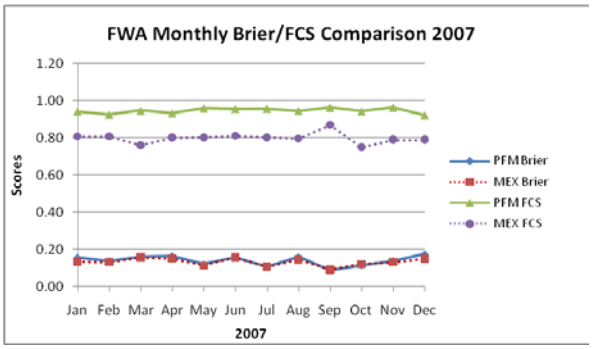

Figure 10. Monthly comparison of Brier score and Forecast Convergence Score (FCS) at site FWA for 2007.

held the same accuracy yet human forecasts were approximately 16% more consistent than model forecasts.

7. PoP Forecast Methodology in NDFD Era

We present a simplified forecast methodology with respect to PoP that incorporates accuracy, consistency, and continuity while also allowing for simplified collaboration and improved consistency among neighboring NWS offices in the NDFD era. This methodology is an attempt to accomplish a balance between accuracy of forecasts (using Brier score); and continuity and convergence between forecasts that trend toward the correct outcome (using FCS). While this methodology may seem to go against conventional probabilistic theory in some regards, it actually embraces the theory.

This methodology requires forecasters to be cognizant of previous forecasts for each forecast period of every forecast and to heed the advice of Brier to not allow the verification system to influence the forecaster. NWS forecasters need to realize that MOS can change dramatically between forecast cycles and therefore there are times when forecasters must remain consistent with the previous forecast until more certainty in the model trends is established. In some cases this may lead to a delayed response in

2007 Average Brier and FCS Scores for SBN and FWA								
	SBN				FWA			
	PFM Brier	MEX Brier	PFM FCS	MEX FCS	PFM Brier	MEX Brier	PFM FCS	MEX FCS
JAN	0.22	0.25	0.95	0.75	0.16	0.13	0.93	0.75
FEB	0.20	0.17	0.91	0.75	0.14	0.13	0.92	0.76
MAR	0.18	0.16	0.93	0.71	0.16	0.16	0.94	0.70
APR	0.14	0.15	0.92	0.75	0.16	0.15	0.92	0.75
MAY	0.12	0.12	0.96	0.76	0.12	0.11	0.95	0.76
JUN	0.11	0.10	0.94	0.75	0.16	0.16	0.95	0.76
JUL	0.12	0.11	0.96	0.77	0.11	0.11	0.94	0.75
AUG	0.18	0.15	0.95	0.73	0.16	0.14	0.94	0.75
SEP	0.11	0.10	0.94	0.84	0.09	0.09	0.96	0.84
OCT	0.12	0.12	0.94	0.67	0.11	0.12	0.94	0.70
NOV	0.20	0.20	0.94	0.78	0.14	0.13	0.96	0.79
DEC	0.16	0.14	0.93	0.80	0.17	0.15	0.92	0.79
YR AVG	**0.16**	**0.15**	**0.95**	**0.79**	**0.14**	**0.13**	**0.95**	**0.80**

Table 1. Actual monthly and yearly statistics comparing Brier Score and FCS.

increasing or decreasing a PoP compared to guidance, but over the long term it is believed this will create greater confidence in NWS forecasts and allow customers to make critical weather decisions at longer projection times more consistently.

Consistency and continuity of PoP forecasts over time must also be considered in the NWS NDFD with respect to neighboring forecast offices. The NWS is striving to provide well collaborated forecasts on a routine basis between borders of NWS forecast offices to provide a consistent, collaborated and seamless forecast across the United States. This can be problematic at times in the later projection periods, especially for potential high impact weather events or when there is little model agreement. Disagreement among offices can be high during these situations, leading to varying PoPs among neighboring offices for a given forecast period, as well as widely changing PoPs over the course of the forecast cycle. This can often lead to the exceedance of the required 20 percent collaboration threshold outlined in NWS Directive 10-506 for 12-h forecast periods.

NWS forecast offices must remember the uncertainty in forecasting precipitation and remember probabilistic theory, which implies a lower probability of an event occurring within a specific forecast time period as projection time increases. Therefore forecasters should be very conservative with the introduction of a “measurable PoP” into the forecast, based on the expected relatively low probability that measurable precipitation will actually occur in any given 12-h period (only about 25 percent of all 12-h periods in this study had measurable precipitation). Since the higher probability event is actually for no precipitation to occur, we recommend forecasters keep this in mind and remain conservative with the addition of a PoP in the later forecast periods where skill and accuracy have been shown to suffer.

We also recommend that once a PoP has been introduced into the forecast for an expected precipitation event, successive forecasts should allow this PoP to continue with either small incremental changes or no changes at all, similar to the figure 7a forecast. As confidence in the expected outcome increases, the PoP can be raised or lowered gradually, with greater changes as the projection time decreases. This would require NWS forecasters to show restraint at times and either not make any changes, or gradually increase (or decrease) PoP by only 10 to 20 percent with each new forecast. This would lead to a more consistent forecast between successive issuances by reducing the variability that is noted in MEX MOS and often translated into human produced forecasts. This would also allow forecasts to remain collaborated through time by reducing the potential for any office to drastically change a PoP based on a specific forecast solution that may be an anomaly.

This philosophy works best for “weak” synoptic systems or weather events that may have greater uncertainty due to model differences and/or lack of run to run continuity. This methodology, as practiced by forecasters at NWS KIWX during 2007, did show merit. Forecasters introduced a “measurable PoP” in only 23 percent of the day-7 forecast periods and precipitation actually occurred in 25 percent of these 12-h periods. It must be noted that forecasters did not always correctly forecast the periods in which precipitation occurred, but they did show a tendency to introduce a PoP into the forecast a correct number of times based on actual precipitation occurrences. In contrast, MEX MOS PoP introduced a PoP greater than 14 percent in nearly 90 percent of the day-7 forecast periods.

Forecasters also need to be cautious with the actual probability that is introduced, keeping in mind the variability and potential error in longer projection times. We recommend that when a PoP is introduced into the forecast in the later projection periods (beyond day 4), that it be within 10 percent of the 12-h climatology PoP for that time of year.

The reasoning behind this methodology is two-fold. First, if forecasters are identifying a weather feature that suggests a PoP is needed, it can be assumed that this feature is representative of a climatologically normal occurrence. Therefore, its probability of occurrence would be in-line with the statistical normal for the time of year that the event is expected to occur. Secondly, model temporal and spatial accuracy in the later projection periods have limited skill compared to the shorter projection periods (Carroll and Maloney III, 2004). Thus any attempt to forecast much higher than climatology may lead to more variability among PoP in successive forecast

issuances, leading to decreased continuity. By starting within 10 percent of climatology beyond day-4 and only being able to make small incremental changes, offices should be able to more easily collaborate12-h PoP forecasts and stay within the designated 20 percent collaboration threshold outlined for NDFD.

Therefore if we use a systematic forecast approach that is conservative and consistent between forecast issuances, we can reduce the variability of PoP between forecasts while still having a relatively accurate forecast compared to MOS. This will yield higher customer confidence in forecasts and allow users to anticipate the eventual forecast outcome and make critical decisions at longer projection times.

As a final thought on methodology, there is a current movement within the NWS to populate all periods of the 7-day forecasts with GFS MOS data after each 0000 UTC and 1200 UTC model cycle but prior to NWS forecast issuance time. Forecasters are being encouraged to only make changes to the weather elements in periods where forecaster confidence is high with regards to "beating" the guidance. Otherwise forecasters are encouraged to use the model output as it is loaded. This methodology goes against what Brier was trying to communicate in his paper (Brier 1950).

While it has been shown that this methodology can result in overall improvements in verification scores compared to MEX guidance (Anderson and Zeitler 2007), we believe this methodology is flawed based in part on the data presented in this paper for PoP and the variability seen in MEX MOS PoP. It is beyond the scope of this paper to argue against this grid editing and forecast philosophy. However, a future paper is planned that will address this methodology and show why we believe customer confidence will be lost with such a methodology.

8. Summary and Conclusions

Forecast verification scores can be subjective and misleading when looked at individually and may not reflect the true value of a forecast. Interpretation of a "good" forecast is subjective, but we believe consistency and continuity are needed in forecasts to promote customer confidence. If continuity can be maintained from one issuance to the next, decision makers will be able to anticipate a forecast outcome with confidence and make earlier planning decisions with respect to potential weather events.

Data collected in this study clearly showed a trend for MEX MOS to forecast PoPs greater than 14 percent in the day-7 forecast periods while PFM PoPs were more conservative. While measurable precipitation occurred in just under 25 percent of the day-7 periods, MEX MOS forecast a PoP greater than 14 percent in just over 90 percent of these periods. The PFM forecasts were more representative of the actual observations, with a measurable PoP forecast in just over 25 percent of the day-7 periods. These data also showed the tendency of MEX MOS PoPs to change significantly between successive forecasts while PoP forecasts from the PFM showed a more consistent trend that added confidence to the forecast. It was also shown how the MEX MOS can have an improved verification score compared to the subjective PFM forecasts despite its variability.

The Ruth-Glahn Forecast Convergence Score was introduced as a means by which the variability of PoP could be measured over the course of a 7-day forecast. Incorporating the FCS into verification is one method by which continuity and consistency, which suggests higher confidence forecasts, can be measured and used in addition to traditional verification scores. While model forecasts may be more accurate on a case by case or even monthly time scale, human forecasts tend to be more consistent and have better continuity toward the end result. Rather than base PoP forecast verification solely on a single verification method, such as Brier score, we propose adding the FCS into the verification scheme. This will provide a measure of consistency along with accuracy.

Any method which attempts to use model guidance as a replacement for human intuition, experience and knowledge may be shown to have higher verification scores, but will certainly not be viewed favorably by customers given the inherent fluctuation and variability that may occur between forecasts, especially in complicated weather situations. By following a consistent and continuity based methodology, customers will be able to recognize trends in forecasts with higher certainty and much sooner than with forecasts that are highly variable and drastically change from one issuance to the next.

NWS forecasters should work to improve verification scores but also take into account continuity, consistency and trends when making forecast decisions. Large changes in forecasts from one issuance to the next should be avoided except when it becomes quite obvious such a change is needed. Knowledge of previous forecasts, model bias and anticipation of how the next model cycle will change are crucial to improving overall forecast verification, quality and value.

Acknowledgements

The lead author would like to express sincere gratitude to all the co-authors who put in countless hours collecting and recording the raw data and contributing to this paper by performing multiple tasks and filling shifts. We would like to give special thanks to all the people who provided email correspondence, answered questions, and reviewed this paper. Special

thanks to David Ruth and Valery Dagostaro who provided insightful information into the Ruth-Glahn Forecast Convergence Score. A final thanks to Science and Operations Officer Jeffrey Logsdon for many insightful reviews of this paper.

References

Anderson , G. C. and J. W. Zeitler, 2008: The NWS Southern Region Grid Preparation policy – Making a Difference. *Preprints, 22nd Conference on Weather Analysis and Forecasting/18th Conference on Numerical Weather Prediction*, New Orleans, LA, Amer. Meteor. Soc., P1.34.

Baars, J. A., and C. F. Mass, 2005: Performance of National Weather Service Forecasts Compared to Operation, Consensus, and Weighted Model Output Statistics. *Weather and Forecasting*, **20**, 1034-1047.

Brier, G. W., 1950: Verification of Forecasts Expressed In Terms of Probability. *Monthly Weather Review*, **1**, 1-3.

Brooks, H. E., and C. A. Doswell III, 1996: A Comparison of Measures-Oriented and Distributions-Oriented Approaches to Forecast Verification. *Weather and Forecasting*, **11**, 288-303.

Carroll, K. L. and J. C. Maloney, 2004: Improvements in Extended-Range Temperature and Probability of Precipitation Guidance. *Preprints, 17th Conference on Probability and Statistics in the Atmospheric Sciences*, Seattle, WA, Amer. Meteor. Soc., 1.10.

Dallavalle, J. P., and V. J. Dagostaro, 2004: Objective Interpretation of Numerical Weather Prediction Output – A Perspective Based on Verification of Temperature and Precipitation Guidance. *Preprints, Symposium on the 50th Anniversary of Operational Numerical Weather Prediction*, College Park, MD, University of Maryland, 5.8.

Doswell III, C. A., 2004: Weather Forecasting by Humans-Heuristics and Decision Making. *Weather and Forecasting*, **19**, 1115-1126.

Glahn, H.R. and D. A. Lowry, 1972: The Use of Model Output Statistics (MOS) in Objective Weather Forecasting. *Journal of Applied Meteorology*, **11**, 1203-1211.

Mass, C. F., 2003: IFPS and the Future of the National Weather Service. *Weather and Forecasting*, **18**, 75-79.
-----, 2003: Reply. *Weather and Forecasting*, **18**, 1305-1306.

Murphy, A. H., 1991: Forecast verification: Its Complexity and Dimensionality. *Monthly Weather Review*, **119**, 1590-1601.
-----, 1993: What Is a Good Forecast? An Essay on the Nature of Goodness in Weather Forecasting. *Weather Forecasting*, **8**, 281-293.
-----, and M. Ehrendorfer, 1987: On the Relationship between the Accuracy and Value of Forecasts in the Cost-Loss Ratio Situation. *Weather and Forecasting*, **2**, 243-251.
-----, and R. L. Winkler, 1987: A General Framework for Forecast Verification. *Monthly Weather Review*, **115**, 1330-1338.
-----, and D. S. Wilks, 1998: A Case Study of the Use of Statistical Models in Forecast Verification: Precipitation Probability Forecasts. *Weather and Forecasting*, **13**, 795-810.

Roebber, P. J., and L. F. Bosart, 1996: The Complex Relationship between Forecast Skill and Forecast Value: A Real-World Comparison. *Weather and Forecasting*, **11**, 544–559.