EASTERN REGION TECHNICAL ATTACHMENT
NO. 99-1
FEBRUARY 1999

A Probability of Precipitation Equation for Columbia, South Carolina Derived from Logistic Regression

Harry Gerapetritis
NOAA/National Weather Service Forecast Office
Columbia, South Carolina

Editor's Note: Authors current affiliation is Greenville-Spartanburg, SC

1. INTRODUCTION

While probability of precipitation (PoP) forecasts are generated from the aviation (AVN) and nested grid (NGM) models for operational use, centralized PoP guidance from the Eta model is not available. Interest thus arose in developing a local PoP equation by using Eta model data as predictors. We also developed PoP equations by utilizing predictor variables taken from both the Eta and NGM models to examine the merits of combining information from different models in a statistical forecast system. The goal in deriving local PoP equations was to produce statistical forecast guidance competitive with the centrally produced, Model Output Statistics (MOS) guidance that would also be easy for forecasters to use in an operational setting.

2. DATA

The limited model output data found in the FOUS messages FRH and FRHT are generally available to forecast offices in a timely fashion and contain data which are appropriate candidates for inclusion as predictor variables in a PoP equation. Thus, seven variables were selected from the FRH and FRHT bulletins as potential predictors. These included precipitation amounts (PTT); mean-layer relative humidities R1, R2, and R3; 700 mb vertical velocity (VVV); lifted index (LI); and sea-level pressure (PS). Data were collected only for the first forecast period relative to the model run time, i.e., the 12-h through 24-h forecast period. Model precipitation amount forecasts represented accumulated amounts for the entire 12-h through 24-h period. All other data were instantaneous values valid at the midpoint of the forecast period, i.e., the 18-h forecast projection. These variables were collected from both the NGM and Eta model for Columbia, SC (CAE) for the local cool season running from October through March. The developmental data set contained 302 cases taken from the 0000 and 1200 UTC model runs from October 16, 1996 through March 31, 1997. A binary representation of the occurrence of measurable precipitation at the CAE Automated Surface Observing System (ASOS) was the dependent variable. Independent data from the 0000 and 1200 UTC model runs of the Eta and NGM for the period October 1, 1997 to March 31, 1998

were reserved for verification. There were 352 cases included in the verification data set.

3. EQUATION DEVELOPMENT

Four separate regression equations were developed based on:

1. Potential predictor variables obtained from the FRH bulletin. This would result in a PoP based solely on the Eta model.
2. Potential predictors obtained from the FRHT bulletin. This would result in a PoP based solely on the NGM model.
3. Potential predictors obtained from both the FRH and FRHT bulletins. This could result in an Eta/NGM mixed-model PoP.
4. Potential predictors obtained by averaging the FRH and FRHT values of each field. This would always result in a mixed-model PoP.

A fifth PoP value was also calculated from the consensus or average of the forecasts from these four types of equations.

The statistical method most suited for a probabilistic predictand, which is bounded between 0 and 100% for all data, is logistic regression (Wilks 1995). This technique uses binary predictands of 0 (for no measurable precipitation) and 1 (for measurable precipitation) and fits regression parameters to yield an equation of the form:

$$P = 1 - \{1 / [1 + \exp(X)]\} \qquad (1)$$

where P = Probability of Precipitation, and X can be expanded as:

$$X = a_0 + a_1x_1 + a_2x_2 + ... + a_nx_n \qquad (2)$$

with

a_0= Constant
a_1 through a_n= Coefficients
x_1 through x_n= Independent Variables

A commercially available statistical software package was employed to carry out this computationally intensive iterative technique. The likelihood ratio chi-square statistic was used to select independent variables for inclusion. Screening regression was carried out via forward selection until the addition of further predictor variables no longer contributed to an appreciable reduction in the likelihood ratio chi-square statistic. The stopping criterion used to avoid “over fitting" the data was a p-value greater than .05 for any remaining variables. The p-value associated with the likelihood ratio chi-square statistic represents the probability of achieving a given reduction by chance. Low p-values, therefore, are associated with variables of high predictive value. Using the rather strict criterion of .05 or less for variable inclusion thus assured that statistically well correlated variables would enter into the equations (Delisi 1998).

The regression technique for the PoP equation yielded the following constants, coefficients, and predictor variables:

A. From FRH only:

X = 151.27822257 + (9.101395)(PTT)
+ (.06963757)(R2) - (.15503430)(PS)
+ (.09608366)(LI)

B. From FRHT only:

$$X = 87.35621434 + (.07792668)(R2) + (7.13942030)(PTT) - (.09177211)(PS)$$

C. From FRH and FRHT pooled:

$$X = 87.21559839 + (8.29275042)(FRH\ PTT) + (.05792202)(FRHT\ R2) - (.09093334)(FRH\ PS)$$

D. From FRH/FRHT averaged:

$$X = 89.70909324 + (11.60604041)(PTT) + (.05860707)(R2) - (.09344235)(PS)$$

Note that model precipitation (PTT), low to mid tropospheric mean relative humidity (R2), and mean sea-level pressure (PS) all figure prominently in the equations. It is noteworthy to mention that the LI was selected for the Eta equation, but with a sign opposite that which is intuitive. The more positive (i.e., stable) the LI, the higher the PoP for constant values of the other three predictors. While this might at first appear to be due to the prevalence of "cold air damming" episodes during the South Carolina cool season, one cannot so quickly draw this conclusion. Since the predictor variables are inter-correlated, the sign of the LI coefficient may well be simply a compensating factor or correction to balance out one of the other variables.

4. VERIFICATION

Verification was carried out using an independent dataset consisting of the chosen predictors taken from the 0000 and 1200 UTC runs of the Eta and NGM for the period from October 1, 1997 to March 31, 1998, the cool season subsequent to the developmental dataset. Binary predictands representing the occurrence of measurable precipitation were collected from CAE ASOS observations for the same period. The corresponding PoP forecasts based on NGM model data (FWC) and Aviation model data (FAN) were also collected for each case. This verification data set included 352 cases. The Brier Score (Wilks 1995) was computed for the period one forecast PoP (12-h to 24-h) by each local equation as well as for the FWC, FAN, and consensus. The Brier Score is a widely used technique in the comparative verification of two different methods of forecasting the probability of a categorical event. The Brier Score can range between zero and one with lower scores associated with better forecasts. The results are presented in Table 1. Note that local consensus refers to the consensus of the four local equations. Overall consensus refers to the consensus of the four local equations, the FWC PoP, and the FAN PoP.

The verification statistics in Table 1 indicate that the local equation based on predictor variables from both the Eta and NGM outperformed all other equations. In fact, the four most skillful equations used information from more than one model to arrive at a PoP forecast. Consensus forecasting aside, this seems to indicate that there is some utility in adopting a "mixed-model" MOS approach to PoP forecasting, at least for the cool season. One drawback in using the Brier Score for verification is that it is very difficult to determine the significance of differences in the scores.

The best performing single-model equation was that based upon the FRH data. This is likely due to the fact that the Eta model, with its superior resolution and physics, is

generally considered the most skillful among the three models with respect to forecasting precipitation. This higher skill for the predictor variables most likely produced stronger correlations to measurable rainfall. The local equation taken from FRHT data did the poorest, likely due to the fact that the equation was developed by using a smaller sample size and a smaller predictor pool than those used in developing the FWC or FAN equations.

These results are also encouraging in light of the fact that a relatively small developmental data sample yielded regression equations capable of competing successfully with the FWC and FAN MOS. With the frequent updates to the Eta model, it is very beneficial to be able to produce demonstrably competitive PoP guidance from a limited set of developmental data.

Table 1. Brier Score verification of PoP equations.

Source	Brier Score
Local FRH/FRHT Eqn.	.0714
Overall Consensus	.0715
Local FRH/FRHT Avg. Eqn.	.0723
Local Consensus	.0733
Local FRH Eqn.	.0757
FWC	.0791
FAN	.0862
Local FRHT Eqn.	.0931

5. APPLICATION

If a local PoP technique was to be of any use operationally, it would have to be timely and easy to generate. This rationale led to the selection of the FRH/FRHT bulletins as data sources. It further led to the development of a computer program which ingests and decodes the FRH/FRHT data and then calculates the local PoP. This technique is as timely as the later of the two bulletins and requires only two mouse clicks on a graphical user interface.

It is easy to envision a much more ambitious approach to local PoP forecasting. The full suite of gridded binary data could be used to develop considerably more sophisticated equations and the technique could be extended to all observation points within a forecast office's area of responsibility. All the basic gridded data fields along with a variety of derived fields could be gathered as potential predictors. Final equations could be encoded as script files in order to locally produce graphical guidance products. AWIPS era capabilities combined with modern statistical software packages could make this objective achievable at local weather offices. As with any statistical technique, however, care must be used to ensure that guidance equations are applied to data with the same statistical characteristics as the developmental dataset.

ACKNOWLEDGMENTS

The author would like to thank Mark Delisi (NWSFO Philadelphia) for his guidance on statistical theory and Mike Cammarata (SOO, NWSFO Columbia) for his review and suggestions.

REFERENCES

Delisi, M. P., 1998: A local large hail probability equation for Columbia, SC. *Eastern Region Technical Attachment,* No. 98-6, National Weather Service, NOAA, U.S. Dept. of Commerce, 7 pp.

Wilks, D. S., 1995: Statistical Methods in the Atmospheric Sciences. Academic Press, 453 pp.