CORRESPONDENCE

Comments on "Comparing Area Probability Forecasts of (Extreme) Local Precipitation Using Parametric and Machine Learning Statistical Postprocessing Methods"

BOB GLAHN

Meteorological Development Laboratory, NOAA/National Weather Service/Office of Science and Technology Integration, Silver Spring, Maryland

(Manuscript received 26 March 2019, in final form 3 July 2019)

Whan and Schmeits (2018, hereafter WS) comprehensively compare three methods of postprocessing: extended logistic regression (ELR); a zero-adjusted gamma distribution; and a machine learning-based method, quantile regression forests. Their predictand is ". . . hourly calibrated radar precipitation in the Netherlands (calibrated against rain gauges" (WS). The potential predictors ". . . are taken from KNMI's high-resolution (2.5-km horizontal grid spacing), non-hydrostatic NWP model HARMONIE-AROME (HA)" (WS). My comments deal only with the ELR method.

Wilks (2009) introduced the "extended" version of the logit model (logistic regression) as a way of not only producing consistent probability forecasts but also of being able to produce probability forecasts for any desired category of the predictand (e.g., amount threshold of precipitation). This is done by introducing into the logistic equation ". . . the predictand threshold itself as one of the regression predictors" (Wilks 2009). The extension is easily demonstrated graphically with one predictor as Wilks (2009) does on the log-odds scale (his Fig. 1) and as Glahn (2014) shows on the probability scale where the functional form of the logit model is apparent (his Fig. 4). In these figures, the probability of precipitation is being predicted and the single predictor (model forecast precipitation) is on the abscissa. In these examples, the binary predictands (the categories) were made from the same element as, and rather well predicted by, the single predictor, precipitation amount. The predictor value, multiplied by its coefficient, predicts the probability of a category (amount of precipitation) after the addition of the constant function (threshold) for that category multiplied by its coefficient.

To my knowledge, the use of ELR has been limited to one or at most two predictors. Graphical depiction of the relationships is difficult with two or more predictors, and I believe has not been done. WS use more than two predictors. Many variables were available as potential predictors (WS's Table 3). It is not clear to me how many were selected, but there seems to be at least four (WS's Figs. 3 and 4). On the other hand, their appendix A states they used three. There is a question as to how well ELR performs with multiple predictors, especially when the predictors may not have a linear relationship to the predictand. This issue was raised by Glahn (2014).

In the results by WS, the logit shows up rather poorly, and merits little discussion by them. In their Figs. 8 and 9, the logit has some skill for the lower thresholds ($\sim$10 mm h^{-1}) and is competitive with the other two methods tested (blue line versus red and green), but the logit skill tails off rather quickly and becomes asymptotic to the zero skill line. The other two methods show results both above and below the zero line, as is many times the case on test data, but ELR never goes below the line, as nearly as can be deduced by the scale of presentation. Also, the error bars for the two techniques other than ELR are generally wide, but the ELR does not show this; the error bars are narrow and never go below zero. For the larger amounts, ELR seems to predict very nearly the same as the (unconditional) climatological frequencies—never worse and never better!

The logit model has been shown to produce good results on many sets of data (e.g., Applequist et al. 2002; Wilks and Hamill 2007). Even though the logit is theoretically more pleasing than linear regression, the logit

Corresponding author: Bob Glahn, harry.glahn@noaa.gov

DOI: 10.1175/MWR-D-19-0089.1

and linear regression show about the same degree of skill after the linear regression results are truncated to zero and one. Linear regression has been used for probabilistic postprocessing by the Meteorological Development Laboratory much more than logistic regression, primarily because linear regression is easier to deal with in an operational environment where millions of equations are developed and used. I am concerned the results for the logit presented in WS's Figs. 8 and 9 may discourage use of the logit model.

It seems that results presented by WS (e.g., their Figs. 8 and 9) for the ELR were obtained by using the cube root of precipitation as the predictand and one of the predictors, but the ''separation function'' [their function $g(q)$] was in its original noncube root form. I believe it is imperative that when the predictand is highly related to a predictor, the separation function be in the same units as, or be a linear function of, the predictand. There is still the question in my mind of the applicability of whether such a separation function would be appropriate for other predictors. It may well be that the choice of the separation function will play a larger role in predictor selection (in a screening framework) than the conditional statistical relationships of the predictors to the predictand. In other words, how does one go about choosing a suitable separation function?

The ''extension'' to the logit is very appealing and has taken the postprocessing community by storm. But no definitive tests have been made concerning its general use where multiple, interrelated predictors that can be anything thought to be related to the predictand are used.

What does one give up if the extension feature is not used? First, the probabilistic prediction of any specific desired amount cannot be made unless that amount was one used in the development (training). Although theoretically very pleasing, in an operational sense, this may be of limited importance. Second, the nonextended logit will sometimes give inconsistent results; this happens within the important range of probabilities if many thresholds are used (they are close together) and with small data samples. While these two benefits of the ELR are real and may be important, skill should not be substantially sacrificed to achieve them. Until proven otherwise, I believe the extension feature of the logit should be used with caution for more than one predictor and should always be checked against the logit without the extension feature. This caution was raised by Wilks (2009) when he states, ''The question from a practical perspective is whether a functional form for $g(q)$ can be specified, for which Equation (5) provides forecasts of competitive quality to those from the traditional single-quantile Equation (1).''

So, the question for WS is, ''Why do the logit model results asymptote to zero with very narrow error bars? Usually, with an inadequate sample, the equations will be unstable, and the results erratic. With ELR, where the equations for all thresholds are developed together, and actually become one equation, the many instances of low precipitation amounts swamp the effect of the higher amounts, so the results depend primarily on the low amounts and the extension function.[1] What would be the results if nonextended logit were used? Reasons for the strikingly dissimilar results for the ELR and the other two methods are needed.

Acknowledgments. The views expressed are those of the author and do not necessarily represent those of any governmental agency.

REFERENCES

Applequist, S., G. E. Gahrs, and R. L. Pfeffer, 2002: Comparison of methodologies for probabilistic quantitative precipitation forecasting. *Wea. Forecasting*, **17**, 783–799, https://doi.org/10.1175/1520-0434(2002)017<0783:COMFPQ>2.0.CO;2.

Glahn, B., 2014: A nonsymmetric logit model and grouped predictand category development. *Mon. Wea. Rev.*, **142**, 2991–3002, https://doi.org/10.1175/MWR-D-13-00300.1.

Whan, K., and M. Schmeits, 2018: Comparing area probability forecasts of (extreme) local precipitation using parametric and machine learning statistical postprocessing methods. *Mon. Wea. Rev.*, **146**, 3651–3673, https://doi.org/10.1175/MWR-D-17-0290.1.

Wilks, D. S., 2009: Extending logistic regression to provide full-probability-distribution MOS forecasts. *Meteor. Appl.*, **16**, 361–368, https://doi.org/10.1002/met.134.

——, and T. M. Hamill, 2007: Comparison of ensemble-MOS methods using GFS reforecasts. *Mon. Wea. Rev.*, **135**, 2379–2390, https://doi.org/10.1175/MWR3402.1.

[1] The forecast for a larger threshold being dependent on data for lower thresholds can be helpful, given a suitable $g(f)$, when the number of cases for the larger threshold is small. Without the extension feature, it might not be possible to develop a stable equation for the larger threshold. You are, though, at the mercy of the $g(f)$.