CORRESPONDENCE

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

KIRIEN WHAN AND MAURICE SCHMEITS

R&D Weather and Climate Modeling, Royal Netherlands Meteorological Institute, De Bilt, Netherlands

(Manuscript received 21 June 2019, in final form 2 July 2019)

1. Introduction

We appreciate Dr. Glahn's comment (Glahn 2019, hereafter G19) on our recent paper (Whan and Schmeits 2018, hereafter WS18) as it gives us the opportunity to clarify the methodology and highlight an important feature of extended logistic regression (ELR).

We apologize for the confusion over the number of predictors selected in ELR. As shown in Figs. 3 and 4, and described in the text of appendix A of WS18, we use a total of four predictors. This means that we use the threshold predictor and three other selected predictors from the list of potential predictors in Table 3 of WS18. WS18 and a previous study using ELR have found equal or greater skill with multiple predictors, compared to a single predictor (Lugt 2013). For example, Fig. 6 in WS18 demonstrates the equal or greater skill in ELR models with additional predictors compared to the model using only H-A precipitation as a predictor. Another way to incorporate the dependence on a threshold in ELR has been introduced by Ben Bouallègue (2013).

We disagree with G19 that the ELR results are rather poor. For example, in the afternoon period they are skillful until around 20 mm h^{-1} and then converge to no skill with respect to climatology (Fig. 8a of WS18). While 10 or 20 mm h^{-1} may sound like a "lower threshold" (G19), these numbers are above the 95th percentile (Table 2 of WS18), and so skillful forecasts of these amounts is nontrivial.

However, we agree that the behavior of the ELR model in WS18 warrants further discussion. We in no way wished to discourage use of ELR, as it has been shown to be skillful in this and many previous applications (e.g., Wilks 2009; Schmeits and Kok 2010; Ruiz et al. 2012; Messner et al. 2014a,b). We transformed the predictor, HARMONIE-AROME (HA) precipitation, by taking the cube root, and keep the response variable (calibrated radar precipitation) and the threshold predictor (in mm h^{-1}), as mentioned in the last paragraph of section 2a (WS18). In WS18 and the discussion below, we transform the threshold predictor when we transform the response variable. We tested various combinations and noted that "the highest skill is generally achieved for the parametric methods (ELR and ZAGA) when either both the response (observed precipitation) and predictor variables (HA precipitation) are transformed, or when only the predictor variable is transformed" (WS18). We reached this conclusion after testing all four combinations of transformations; that is, by transforming 1) the predictor, response and the predictor threshold (Fig. 1a), 2) neither (Fig. 1b), 3) only the predictor variable (Fig. 1c), or 4) only the response variable (Fig. 1d). Scatterplots show the best linear relationship between the predictor and response when both are transformed (Fig. 1). In WS18 we tested these various combinations on a limited dataset (we used a threefold cross validation that used two-thirds of two years for the training test, and the remaining one-third of two years of data as the testing set) by then examining the skill of forecasts, according the BSS, for precipitation amounts up to 20 mm h^{-1}, using a single set of precipitation values for the threshold predictor. We noted in WS18 that the skill of ELR depended on the choice of transformations. In addition, ELR depends heavily on the choice of thresholds used for the threshold predictor in training

Corresponding author: Kirien Whan, whan@knmi.nl

DOI: 10.1175/MWR-D-19-0210.1

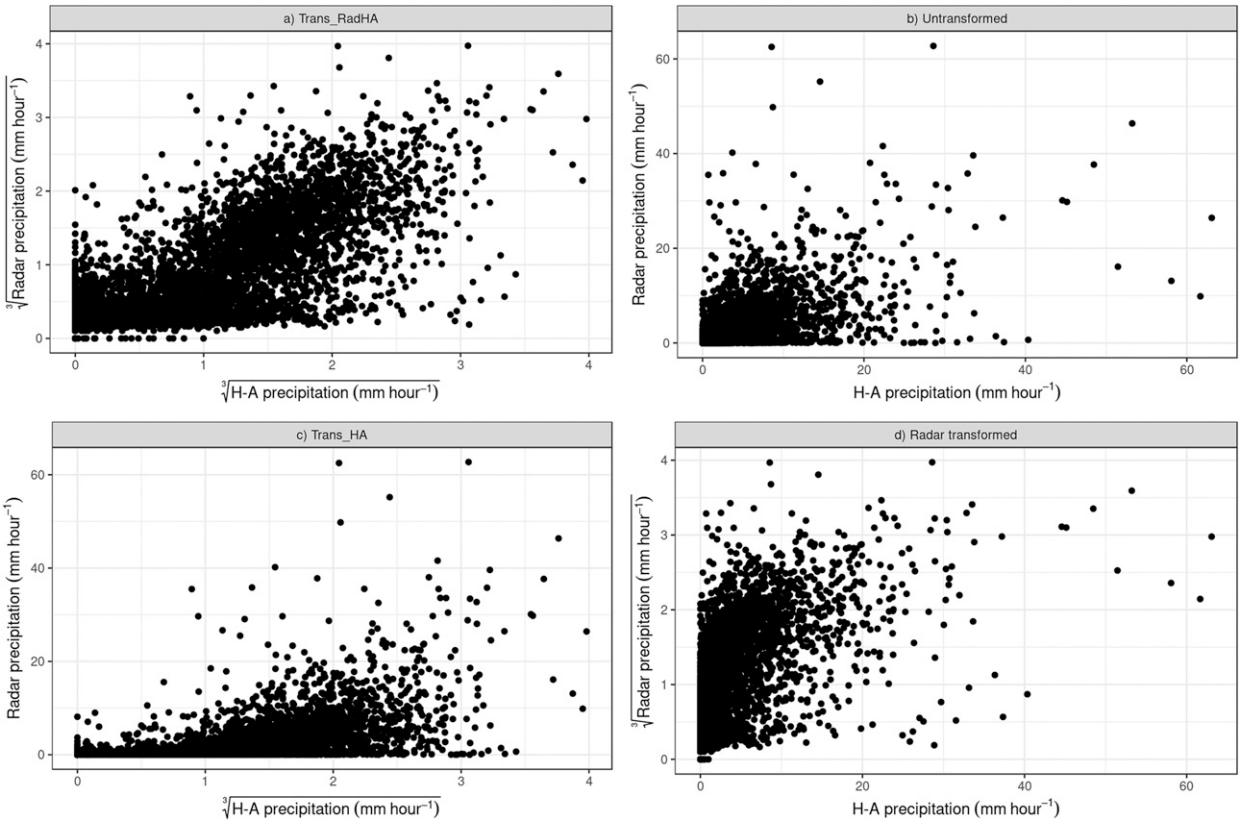

FIG. 1. Scatterplots of the response (radar precipitation) and most important predictor variable (H-A precipitation), when (a) both the predictor and response are transformed, (b) neither variable is transformed, (c) only the predictor variable is transformed, and (d) only the response variable is transformed. Transformations are made by taking the cuberoot.

(Messner et al. 2014a; Lugt 2013). We explore the implications of, and interactions between, these two choices below.

2. Data and methods

We use the same dataset and experimental design as in the final comparison of WS18, except where noted in the following sentences. Shortly, we use a threefold cross validation of three summer half-years (2010, 2011, and 2013), where we train on two years and test on the remaining one year. We focus only on the skill of ELR during the afternoon verification period (1200–1800 UTC) and the shortest lead time (+6–12 h). To shed light on the influence of the threshold predictor, we compare BSS for several sets of precipitation values listed in Table 1. Additionally, we compare the procedure used in WS18, where we only transformed HA precipitation (Trans_HA), with the transformation of radar precipitation, the predictor threshold and HA precipitation (Trans_RadHA). The predictor threshold is transformed in all cases where radar precipitation is transformed. This is similar to the experiment mentioned above and in WS18 but here we use the leave-one-year-out cross validation as in the final comparison of methods in WS18.

Finally, we fit ordinary logistic regression (LR) models for several precipitation thresholds (0.3, 5, 10, 15, 20, 25, 30, 35, and 40 mm h^{-1}), as suggested by G19. We used

TABLE 1. The untransformed precipitation values (mm) used for the threshold predictor in ELR. WS18 corresponds to the 'High' set of quantiles listed in Table A1 of WS18. The precipitation values in WS18 correspond to the following quantiles in the afternoon period: 0.5, 0.7, 0.8, 0.85, 0.9, 0.925, 0.95, 0.99, 0.995, and 0.999.

Name of threshold set	Precipitation thresholds (mm)
WS18	0.12, 0.96, 2.75, 4.12, 6.05, 7.57, 9.68, 20.88, 27.29, 39.87
Set B	0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 15, 20, 25, 30, 35, 40
Set C	0.3, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47, 48, 49, 50
Set D	1, 5, 10, 15, 20, 25, 30, 35, 40

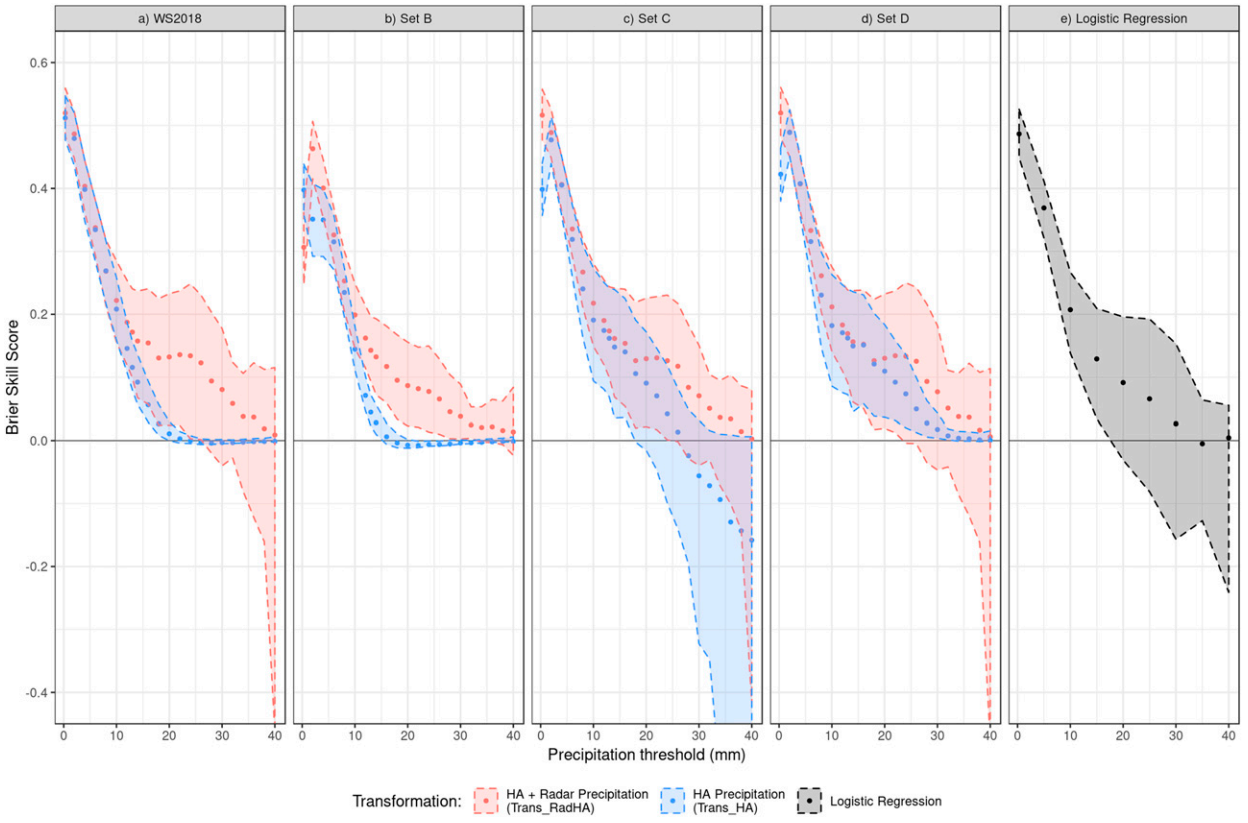

FIG. 2. The BSS for the +6–12-h precipitation forecasts for thresholds between 0 and 40 mm h^{-1} in the afternoon (1200–1800 UTC) verification period. (a)–(d) The BSS from models fit using ELR (shading and dashed lines indicate the 95% confidence intervals of the BSS calculated from block bootstrapping). (e) The BSS from separate logistic regression models. We present the information in (e) continuously for easy comparison with (a)–(d), but it should be noted that the separate logistic regression equations from (e) cannot produce forecasts for all precipitation thresholds. All statistical models use stepwise selection with HA precipitation and all atmospheric indices as potential predictors (see Table 3 in WS18).

stepwise selection with a maximum of two predictors. We first tested the skill of the models when the stepwise selection was applied separately to each LR model, but chose to use a common set of predictors to reduce the occurrence of quantile crossings (i.e., the issuing of higher exceedance probabilities for larger precipitation amounts compared to smaller precipitation amounts). We chose two predictors as we found the models with two predictors to be the most consistent in terms of selected predictors. The predictors used in the LR models are the following: transformed HA precipitation and the modified Jefferson index.

3. Results

We first compare the effect of transforming the predictor (Trans_HA) with forecasts made when transforming both the response and predictor (Trans_RadHA), while using the same threshold predictor values as in WS18 (Fig. 2a). The blue line in Fig. 2a corresponds to the blue line in Fig. 8a in WS18, and it shows the concerning behavior noted by G19 where ELR converges to the no skill line around 20 mm h^{-1} using Trans_HA. The red line in Fig. 2a shows the skill of ELR if both the response and predictor variables are transformed (Trans_RadHA). There are few differences in the skill until around 10 mm h^{-1} but after this point the skill of Trans_HA drops quickly to no skill, while the median skill of Trans_RadHA remains high but with much larger uncertainty. The lower confidence bound of Trans_RadHA actually crosses the zero line around 25 mm h^{-1}, while Trans_HA converges to zero around 15 mm h^{-1}. The higher median BSS of Trans_RadHA in Fig. 1a is also preferable to Trans_HA. The behavior of Trans_RadHA is much more similar to that of zero-adjusted gamma distribution (ZAGA) and quantile regression forests (QRF) (from Fig. 8a in WS18).

It is known that the training threshold values chosen for the threshold predictor in ELR are important. When using the untransformed response variable (Trans_HA, as in WS18) the shape of the BSS depends heavily on the

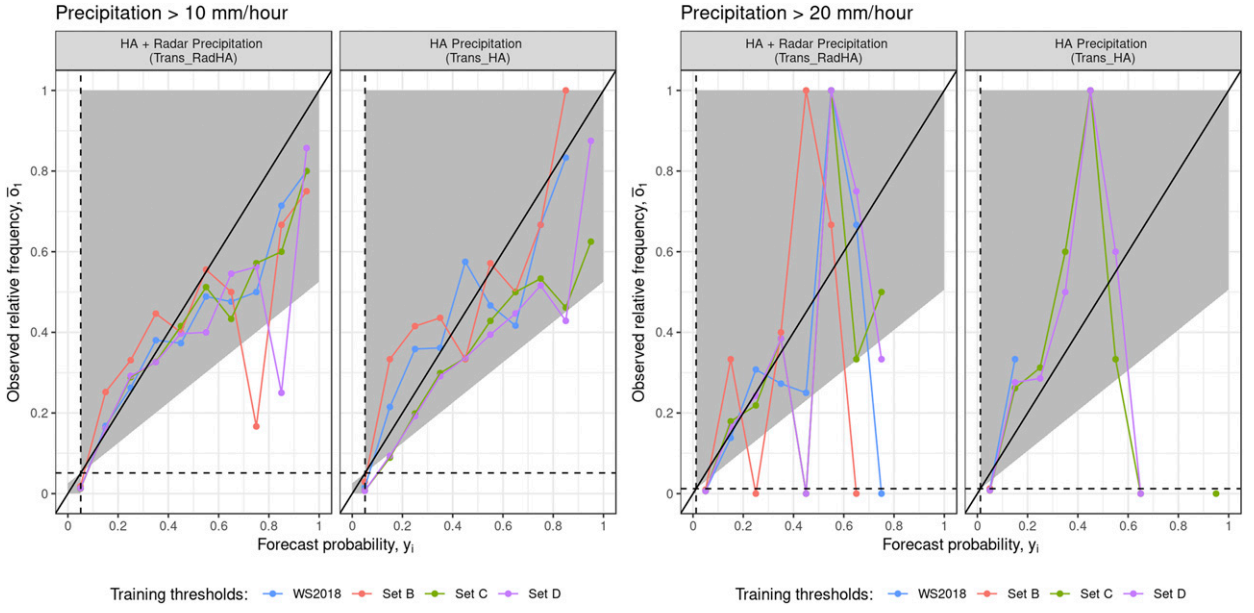

FIG. 3. Reliability diagrams for the probabilities of precipitation >10 and 20 mm h^{-1} predicted by the same models as shown in Figs. 2a–d.

chosen predictor thresholds. In Figs. 2c and 2d the uncertainty around the BSS is larger in Trans_HA, compared to Figs. 2a and 2b, although the skill in Fig. 2d still converges to zero. For the Trans_RadHA models, there is less dependence on the set of chosen predictor thresholds, with all models showing large BSS uncertainty and crossing the no-skill line for precipitation amounts around 25 mm h^{-1} (red lines in Fig. 2).

Reliability diagrams for the probability of precipitation >10 and 20 mm h^{-1} show that all models are reasonably reliable at the lower precipitation threshold (Fig. 3). The differences between Trans_RadHA and Trans_HA for the forecasts of precipitation exceeding 20 mm h^{-1} are noteworthy. It can be seen that when the response variable and predictor threshold are not transformed (Trans_HA), the range of probabilities that are forecast depends heavily on the precipitation thresholds used as the predictor. For example, in Trans_HA high forecast probabilities are only issued when the "Set C" and "Set D" sets of precipitation values are used as the threshold predictor, while in the models using the "WS18" and "Set B" sets of threshold predictors only low forecast probabilities are issued. It is this inability of the latter models to issue high forecast probabilities that results in the very narrow confidence intervals of ELR in WS18 that was noted by G19. This dependence on the threshold predictor is not seen for the Trans_RadHA models, which are all able to issue high forecast probabilities for this extreme precipitation amount.

There appear to be two preferred models from this selection: 1) the model that used transformed radar precipitation and the Set B' set of thresholds as the predictor (Trans_RadHA_SetB; the red line in Fig. 2b), and 2) the model that used untransformed radar precipitation and the Set D set of thresholds as the predictor (Trans_HA_SetD; the blue line in Fig. 2d). These models are preferred given their relatively high median BSS for larger precipitation thresholds and their eventual convergence to the no skill line. This demonstrates that transformation of the predictor threshold is not strictly necessary, as skillful forecasts can be made with an untransformed predictor threshold.

Finally, we compare the skill of forecasts made using ELR with those made using separate LR models for each precipitation amount. The general pattern is quite similar including large uncertainties in the BSS for higher precipitation amounts forecast with LR (Fig. 2e). This is logical given the smaller dataset with only a few observed cases for the higher thresholds. The median BSS of the LR forecast is larger than the model used in WS18, but lower than all the Trans_RadHA models and Trans_HA fit using Set D for the threshold predictor.

4. Discussion and conclusions

In summary, we showed the BSS and reliability diagrams of several ELR models fit using various transformations of the response (radar precipitation) and the most important predictor variable (HA precipitation), and with various sets of precipitation values for the threshold predictor. We showed that there are more differences in the shape of the BSS curves when the

response variable and threshold predictor is not transformed, and that there are fewer differences when these variables are transformed. The ability of the model to forecast high probabilities for high thresholds depends strongly on the set of precipitation values used as the threshold predictor when only HA precipitation is transformed. This dependence is reduced when both the response and precipitation predictor are transformed. This shows that while transformation of all variables is not strictly necessary, it results in fewer differences between predictions made with various threshold sets. Future work could explore whether these results can be generalized to other settings.

We based our hyperparameter decisions in WS18 on precipitation amounts up to $10\,\mathrm{mm\,h^{-1}}$. This is a reasonably high quantile of the precipitation distribution but it is only after this amount that the largest differences between the models with transformed and untransformed values are seen. It is clear that other choices in the transformation of the response and threshold predictor, and the set of precipitation values for the threshold predictor could have been made (i.e., Trans_RadHA_SetB or Trans_HA_SetD) that result in BSSs that are more skillful.

These results are shown for the final verification dataset using a leave-one-year-out cross validation. This is not the same dataset that we used to make the choices about hyperparameters in WS18. Many methods require careful selection of hyperparameters and this must be done on a different dataset than the final verification set so as not to tune the hyperparameters too closely to the final set. So while it is not strictly fair to compare these results to QRF and ZAGA as shown in WS18, we note that either of the preferred ELR models are now much more competitive to QRF.

REFERENCES

Ben Bouallègue, Z., 2013: Calibrated short-range ensemble precipitation forecasts using extended logistic regression with interaction terms. *Wea. Forecasting*, **28**, 515–524, https://doi.org/10.1175/WAF-D-12-00062.1.

Glahn, B., 2019: Comments on ''Comparing area probability forecasts of (extreme) local precipitation using parametric and machine learning statistical postprocessing methods.'' *Mon. Wea. Rev.*, **147**, 3495–3496, https://doi.org/10.1175/MWR-D-19-0089.1.

Lugt, D., 2013: Improving GLAMEPS wind speed forecasts by statistical postprocessing. Intern Rep. IR-2013-03, KNMI, De Bilt, Netherlands, 18 pp., http://bibliotheek.knmi.nl/knmipubIR/IR2013-03.pdf.

Messner, J. W., G. J. Mayr, D. S. Wilks, and A. Zeileis, 2014a: Extending extended logistic regression: Extended versus separate versus ordered versus censored. *Mon. Wea. Rev.*, **142**, 3003–3014, https://doi.org/10.1175/MWR-D-13-00355.1.

——, ——, A. Zeileis, and D. S. Wilks, 2014b: Heteroscedastic extended logistic regression for postprocessing of ensemble guidance. *Mon. Wea. Rev.*, **142**, 448–456, https://doi.org/10.1175/MWR-D-13-00271.1.

Ruiz, J. J., C. Saulo, and E. Kalnay, 2012: How sensitive are probabilistic precipitation forecasts to the choice of calibration algorithms and the ensemble generation method? Part II: Sensitivity to ensemble generation method. *Meteor. Appl.*, **19**, 314–324, https://doi.org/10.1002/met.262.

Schmeits, M. J., and K. J. Kok, 2010: A comparison between raw ensemble output, (modified) Bayesian model averaging, and extended logistic regression using ECMWF ensemble precipitation reforecasts. *Mon. Wea. Rev.*, **138**, 4199–4211, https://doi.org/10.1175/2010MWR3285.1.

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.