# Postprocessing of GEFS Precipitation Ensemble Reforecasts over the U.S. Mid-Atlantic Region

XINGCHEN YANG, SANJIB SHARMA, AND RIDWAN SIDDIQUE

*Department of Civil and Environmental Engineering, The Pennsylvania State University, University Park, Pennsylvania*

STEVEN J. GREYBUSH

*Department of Meteorology, The Pennsylvania State University, University Park, Pennsylvania*

ALFONSO MEJIA

*Department of Civil and Environmental Engineering, The Pennsylvania State University, University Park, Pennsylvania*

(Manuscript received 29 June 2016, in final form 31 January 2017)

ABSTRACT

The potential of Bayesian model averaging (BMA) and heteroscedastic censored logistic regression (HCLR) to postprocess precipitation ensembles is investigated. For this, outputs from the National Oceanic and Atmospheric Administration's (NOAA's) National Centers for Environmental Prediction (NCEP) 11-member Global Ensemble Forecast System Reforecast, version 2 (GEFSRv2), dataset are used. As part of the experimental setting, 24-h precipitation accumulations and forecast lead times of 24 to 120 h are used, over the mid-Atlantic region (MAR) of the United States. In contrast with previous postprocessing studies, a wider range of forecasting conditions is considered here when evaluating BMA and HCLR. Additionally, BMA and HCLR have not yet been compared against each other under a common and consistent experimental setting. To compare and verify the postprocessors, different metrics are used (e.g., skills scores and reliability diagrams) conditioned upon the forecast lead time, precipitation threshold, and season. Overall, HCLR tends to slightly outperform BMA but the differences among the postprocessors are not as significant. In the future, an alternative approach could be to combine HCLR with BMA to take advantage of their relative strengths.

## 1. Introduction

Numerical weather prediction (NWP) models are used, as part of an ensemble prediction system (EPS), to generate ensemble forecasts of a future weather variable or quantity (Tracton and Kalnay 1993; Toth et al. 2003; Buizza et al. 2005). The ensemble forecasts, in turn, can be used to determine the probability and uncertainty of the weather variable. In the case of precipitation forecasts, however, the magnitude and dispersion of the ensemble forecasts are normally characterized by the presence of biases (Sloughter et al. 2007; Wilks 2009), which makes the determination of forecast probabilities from such ensembles unreliable. To correct the biases and improve the reliability of ensemble forecasts, a number of techniques have been developed and implemented (e.g., Raftery et al. 2005; Wilks 2006b; Bröcker and Smith 2008). These techniques are collectively known as statistical weather postprocessing or calibration.

Postprocessing for ensemble prediction systems has several goals: correct systematic forecast errors or biases, which can be achieved by optimally weighting ensemble members according to past performance, and correct (calibrate) ensemble spread so that it is a useful estimate of forecast uncertainty. Some of the available techniques for postprocessing weather forecasts are regression-based methods (Bremnes 2004; Clark and Hay 2004; Hamill et al. 2004; Friederichs and Hense 2007; Wilks 2009; Roulin and Vannitsem 2012; Messner et al. 2014a,b), Gaussian ensemble dressing (Roulston and Smith 2003; Wang and Bishop 2005), nonparametric methods (Brown and Seo 2010), and Bayesian model averaging (BMA) (Raftery et al. 2005; Sloughter et al. 2007; Schmeits and Kok 2010), among others (e.g., Wu et al. 2011). Many of these techniques share in common the model output statistics (MOS) approach (Glahn and

*Corresponding author e-mail*: Alfonso Mejia, amejia@engr.psu.edu

DOI: 10.1175/MWR-D-16-0251.1

Lowry 1972; Wilks 2006b) since, as part of their methodology, they require the derivation of statistical forecast equations as a function of one or more outputs (predictors) from the NWP model. Additionally, some of the postprocessing techniques allow the complete characterization of the predictive probability density function (pdf) of precipitation forecasts (Sloughter et al. 2007; Wilks 2009; Messner et al. 2014b).

Some of the techniques mentioned have been evaluated for the case of ensemble precipitation forecasts (Sloughter et al. 2007; Wilks and Hamill 2007; Wilks 2009; Brown and Seo 2010; Schmeits and Kok 2010; Messner et al. 2014a,b; Zhu et al. 2015). For instance, Sloughter et al. (2007) extended the BMA approach introduced by Raftery et al. (2005) to the case of ensemble precipitation forecasts. As a statistical weather postprocessor, BMA generates bias-corrected predictive pdfs from the ensemble forecasts (Sloughter et al. 2007; Fraley et al. 2010). Bremnes (2004) employed quantile regression to estimate the conditional quantiles of future precipitation using the forecast precipitation amounts as predictors, alongside other weather-related variables such as the mean relative humidity and wind flow. Wilks (2009) proposed and implemented the extended logistic regression (ELR) approach to include the threshold quantiles of the precipitation forecast as predictor variables, as opposed to relying on the precipitation amounts alone. Messner et al. (2014a) complemented the ELR approach by including the precipitation ensemble spread as a predictor. They termed this approach heteroscedastic extended logistic regression (HELR). They also proposed two additional logistic regression-based approaches for postprocessing precipitation: heteroscedastic ordered logistic regression (HOLR) and heteroscedastic censored logistic regression (HCLR) (Messner et al. 2014b). It is useful to note that HCLR fits the same model as HELR, with the only difference being that the HCLR parameters optimize the continuous predictive pdf, as opposed to the quantile thresholds (Messner et al. 2014b).

A few precipitation postprocessing studies have compared the performance of different postprocessing techniques under a common set of experimental conditions, for example, by using the same geographic region, dataset, and training period to evaluate the postprocessors (Wilks 2006a; Sloughter et al. 2007; Schmeits and Kok 2010; Mendoza et al. 2015; Messner et al. 2014b). The general findings from these studies indicate that the performance of the postprocessors, both relative to sampled climatological conditions and to each other, vary depending on the training strategy (Greybush et al. 2008; Zhu et al. 2015), verification metric considered (Mendoza et al. 2015), forecast lead time (Schmeits and Kok 2010), and bias-correction type (Schmeits and Kok 2010; Erickson et al. 2012), among other factors.

In this study, our primary goal is to assess and verify the potential of BMA and HCLR to postprocess precipitation ensemble reforecasts from the National Oceanic and Atmospheric Administration's (NOAA's) National Centers for Environmental Prediction (NCEP) 11-member Global Ensemble Forecast System Reforecast version 2 (GEFSRv2). We employ GEFSRv2 since its reforecasts, based on a consistent model run, are available over a long time period. This is relevant because forecasts produced by a model whose structure changes in time will produce less statistically consistent forecasts. Although this situation may be unavoidable in operational forecasting, it should be avoided when interest lies in assessing the performance of different postprocessors. We use multisensor precipitation estimates (MPEs) as the observed precipitation when training the postprocessors and verifying the raw and postprocessed ensemble precipitation forecasts. Additionally, we highlight that our evaluation here of BMA is more comprehensive than previous ones since we account for the effect of training period length, spatial pooling strategy, lead time, and seasonality on the BMA postprocessed precipitation forecasts. Moreover, BMA and HCLR have not been compared against each other yet.

We select BMA and HCLR for this study for various reasons. BMA is desirable because it provides an integrated approach for combing ensemble members from a single or multiple NWP models. At the same time, techniques based on logistic regression have been shown to perform as well as or slightly better than BMA in several applications (Sloughter et al. 2007; Schmeits and Kok 2010), while being less computationally demanding. The latter becomes particularly relevant when working with long reforecast datasets. Furthermore, HCLR has recently been shown to outperform and overcome key shortcomings of other logistic regression-based techniques, such as allowing the determination of the full predictive pdf of precipitation forecasts (Messner et al. 2014a).

Key questions that we seek to address with this study are as follows: How does the BMA and HCLR postprocessed forecasts compare against the raw precipitation ensembles? What is the dependence between the performance of the postprocessors and the forecast lead time, training period length, spatial pooling, seasonality, and precipitation threshold? Which postprocessing method is more reliable for the U.S. mid-Atlantic region (MAR)? The remainder of

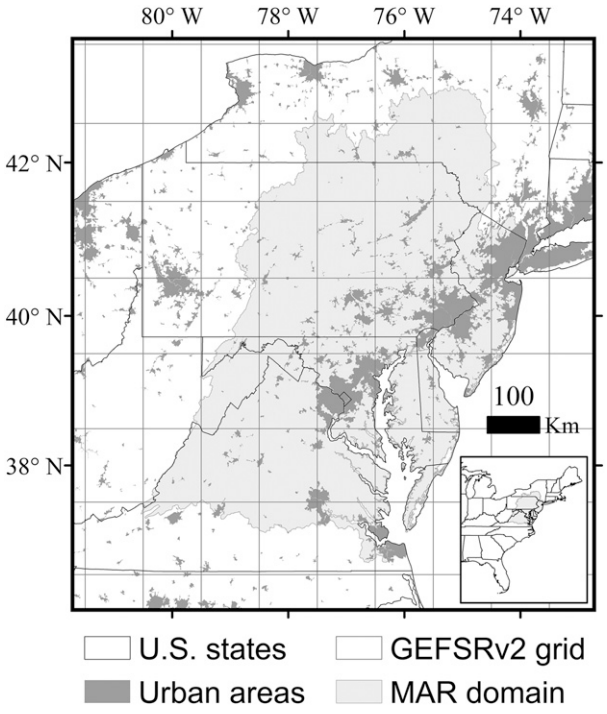


FIG. 1. Map illustrating the geographic domain of the U. S. MAR. The map also shows the major rivers, urban areas, and the GEFSRv2 grid. The inset illustrates the location of the MAR within the eastern portion of the United States.

the article is organized as follows. In section 2, we describe the study area, datasets, and methodology employed. The main results are examined in section 3. Section 4 discusses the results. Last, section 5 summarizes the key findings.

## 2. Data and methodology

### a. Study area

The U.S. MAR is selected as the study area. The geographic location and boundary of the MAR is illustrated in Fig. 1. The MAR comprises the state of Delaware and the District of Columbia, along with parts of the states of Maryland, New York, New Jersey, Pennsylvania, Virginia, and West Virginia (Polsky et al. 2000; Greene et al. 2005). It only occupies approximately 5% of the total landmass of the United States, but it contains approximately 10% of its population (~41 million people) (Siddique et al. 2015). Some of the largest metropolitan areas in the United States are located in the MAR (e.g., Baltimore, Philadelphia, and Washington, D.C.) Additionally, the MAR comprises several major U.S. river basins including the Delaware, Susquehanna, Potomac, and James Rivers. The climate in the MAR is relatively humid. The average annual temperature is approximately 11°C and the mean annual precipitation is approximately 900–1200 mm (Polsky et al. 2000).

### b. GEFSRv2

For the precipitation ensemble forecasts, we use outputs from the GEFSRv2 dataset. GEFSRv2 are the retrospective forecasts produced using the 2012 operational version (version 9.0.1) of the NCEP's Global Ensemble Forecast System (Hamill et al. 2013). The model runs for the GEFSRv2 were initiated once a day at 0000 coordinated universal time (UTC) (Hamill et al. 2013). Initial conditions were perturbed using the ensemble transform technique with rescaling (Wei et al. 2008). The forecast lead times extend from 1 to 16 days and each forecast cycle consists of forecasts valid for 3-hourly accumulations from day 1 to day 3 and 6-hourly accumulations from day 4 to day 16. We use here for the evaluation of the postprocessors 24-h accumulations from days 1 to 5. The native resolution of the reforecasts is ~0.5° on a Gaussian grid for forecasts in the first week and ~0.67° for forecasts in the second week. The GEFSRv2 data are also available at the ~1° resolution for the entire range of lead times (days 1–16). We use here the 1° resolution dataset to facilitate coverage of the entire MAR. Further details about the GEFSRv2 dataset or information on how to access it are provided elsewhere (Hamill et al. 2013, 2016). An important aspect of the GEFS reforecasts are that they use a single model and single set of physics packages, perturbing only initial conditions, with small stochastic perturbations during the forecast phase; this is in contrast to systems like the Short-Range Ensemble Forecast (SREF) that employ different dynamical cores and physics options within an ensemble.

### c. MPEs

We use multisensor precipitation estimates (MPEs) to train the postprocessors and verify the raw and postprocessed ensemble precipitation forecasts. The MPEs were provided by the NOAA's Mid-Atlantic River Forecast Center (MARFC) (Lawrence et al. 2003). Similar to the NCEP stage-IV MPEs (Prat and Nelson 2015), the MARFC MPE product combines radar estimated precipitation with in situ gauge measurements over the MAR and represents a continuous time series of hourly, high-resolution gridded precipitation observations at $4 \times 4\,\text{km}^2$ cells. We aggregated the MPEs to the temporal (24 h) and spatial scale (1°) of the GEFSRv2 data over the period 2002–07. Note that MPEs are subject to errors, such as radar artifacts, but they are also one of the best high-resolution gridded

precipitation datasets available (Prat and Nelson 2015) and therefore appropriate for this study.

### d. Postprocessing techniques

#### 1) BMA

A brief overview of the BMA technique as used for the postprocessing of ensemble precipitation forecasts is provided here since a detailed description is provided elsewhere (Sloughter et al. 2007). As a statistical weather postprocessor, BMA generates bias-corrected predictive pdfs from the ensemble forecasts (Sloughter et al. 2007; Fraley et al. 2010). Specifically, the BMA predictive pdf is a weighted average of pdfs centered on the individual bias-corrected precipitation forecasts. The weights reflect the predictive skill of the individual ensemble members over a selected training period.

The BMA predictive pdf, $P(y\,|\,f_1,\ldots,f_k)$, for the cube root of precipitation accumulation $y$, given the forecast members $f_1,\ldots,f_k$ at a particular lead time, is given by

$$P(y\,|\,f_1,\ldots,f_k)=\sum_{k=1}^{K} w_k\{P(y=0\,|\,f_k)I[y=0]+P(y>0\,|\,f_k)g_k(y\,|\,f_k)I[y>0]\}. \quad (1)$$

The weight $w_k$ is the posterior probability of ensemble member $k$ being the best one, provided that $\sum_{k=1}^{K} w_k=1$. Here $K$ is the total number of ensemble members; $K=11$ for the GEFSRv2 data. The weights are specified according to the relative performance of each ensemble member during the training period employed for parameter estimation. The precipitation forecasts are transformed via the cube root since this transformation has been found to improve the modeling of $P(y>0\,|\,f_k)$ (Sloughter et al. 2007), which is normally represented by a gamma pdf as further explained in the next paragraphs. Note that we tried other transformations (square and fourth root) but the cube root performed better. The term $P(y=0\,|\,f_k)$ is the probability of the cube root of precipitation being equal to zero given the forecast member $f_k$ and assuming that $f_k$ is the best forecast member. The term $I[.]$ is the indicator function that is equal to 1 if the term inside the brackets holds true and 0 otherwise. The term $P(y>0\,|\,f_k)$ is the probability of the cube root of precipitation being greater than 0 given the forecast member $f_k$ and assuming that $f_k$ is the best forecast member.

The term $P(y=0\,|\,f_k)$ is determined as

$$P(y=0\,|\,f_k)=\Lambda[\gamma(f_k)]=\frac{\exp[\gamma(f_k)]}{1+\exp[\gamma(f_k)]}, \quad (2)$$

where $\Lambda(.)$ denotes the cumulative distribution function (cdf) of the standard logistic distribution and $\gamma(f_k)$ is defined as

$$\gamma(f_k)=a_{0,k}+a_{1,k}f_k^{1/3}+a_{2,k}\varepsilon_k. \quad (3)$$

Equation (3) is a logistic regression with parameters $a_{i,k}$ $(i=1,2,3)$ that need to be estimated for each ensemble member $k$. The predictor $\varepsilon_k$ is equal to 1 if $f_k=0$ and 0 otherwise. The parameters in Eq. (3) are determined directly from the ensemble forecast and observed data, using logistic regression with precipitation/no precipitation as the dependent variable, and $f_k^{1/3}$ and $\varepsilon_k$ as the two predictor variables.

The term $P(y>0\,|\,f_k)$ is equal to $1-P(y=0\,|\,f_k)$ while $g(y\,|\,f_k)$ is defined as

$$g_k(y\,|\,f_k)=\frac{1}{\beta_k^{\alpha_k}\Gamma(\alpha_k)}y^{\alpha_k-1}\exp(-y/\beta_k) \quad (4)$$

for $y>0$, and $g(y)=0$ for $y=0$. Equation (4) is a gamma pdf with shape parameter $\alpha_k=\mu_k^2/\sigma_k^2$ and scale parameter $\beta_k=\sigma_k^2/\mu_k$. The mean $\mu_k$ and variance $\sigma_k^2$ of this distribution depend on $f_k$ as follows:

$$\mu_k=b_{0,k}+b_{1,k}f_k^{1/3} \quad (5)$$

and

$$\sigma_k^2=c_0+c_1f_k. \quad (6)$$

The parameters $b_{i,k}$ $(i=0,1)$ in Eq. (5) are member specific. They are determined separately for each ensemble member using linear regression with the cube root of the observed precipitation amount as the dependent variable and $f_k^{1/3}$ as the predictor variable.

Last, using the training data, the parameters $c_0$ and $c_1$ in Eq. (6), as well as the terms $w_k$ $(k=1,\ldots,K)$ in Eq. (1) are estimated by maximum likelihood, as in Sloughter et al. (2007). The approach of Sloughter et al. (2007) maximizes the log-likelihood function numerically using the expectation-maximization algorithm (Dempster et al. 1977; McLachlan and Krishnan 1997, 361–369).

To implement the BMA postprocessor, we use 24-h precipitation accumulations from the GEFSRv2 for lead times from 24 to 120 h. To train the BMA, we use the sliding time window approach of Sloughter et al. (2007). In this approach, a sliding time window comprising the $L$ training days preceding the forecast day is used. The

window moves with the forecast day (i.e., the day the forecast is issued) and, typically, it comprises the preceding 20–40 days prior to the forecast day. We use this same approach here with one important modification. We select training days from the 4 years preceding the forecast day using the same calendar days in each year, as opposed to just using training days from a single year. For example, for a GEFSRv2 reforecast issue on 31 March 2005, we select as the training data the days from 1 to 30 March (assuming a 30-day training window) in the years 2002–05; thus, we use in this example a total of 120 training days [i.e., (30 days) × (4 years)]. We select the size of the training window empirically by testing different window sizes.

Additionally, when training the BMA algorithm, it is common to rely on spatial pooling to increase the sample size of the training dataset. However, very few studies have assessed the effect of spatial pooling on the performance of BMA, particularly in the context of precipitation (Kleiber et al. 2011). Thus, we evaluate here the effect of spatial pooling on the BMA algorithm by varying the number of GEFSRv2 cells that are used for training. In this study, we select a total of 20 GEFSRv2 cells since they cover the majority of the geographic domain of the MAR. To test different training scenarios, we use 1, 5, 10, and 20 neighboring cells to train the BMA algorithm. The 1 cell scenario means that each cell is trained individually without pooling data from the other cells. In contrast, the 5 cells scenario means that the 20 GEFSRv2 cells that encompass the MAR are divided into 4 groups of neighboring cells with 5 cells in each group; each group is then trained separately by pooling the data from its 5 cells. For example, for the case of 5 cells and a 30-day sliding window, we use 6600 reforecasts to train the BMA algorithm at a given forecast day [i.e., (30 days) × (5 cells) × (4 years) × (11 members)].

Our previous description of BMA assumes that the ensemble members are individually distinguishable where distinct weights may have a physical interpretation. In our BMA postprocessing experiment, however, all the ensemble members come from the same NWP model, which means that the members lack individually distinguishable physical features. In this situation, the ensemble members are exchangeable, which means that the BMA weights can be assumed to be equal (Fraley et al. 2010; Schmeits and Kok 2010) [i.e., $w_k$ in Eq. (1) is equal to $1/K$]. Additionally, the exchangeability condition makes other parameter constraints possible. Specifically, the parameters $a_{i,k}$ ($i = 1, 2, 3$) in Eq. (3) and $b_{i,k}$ ($i = 0, 1$) in Eq. (5) are the same for all the exchangeable members that come from the same NWP model so that $a_{i,k} = a_i$ ($i = 1, 2, 3$) and $b_{i,k} = b_i$ ($i = 0, 1$) (Fraley et al. 2010; Schmeits and Kok 2010). Hereafter, we use the term BMA to indicate the implementation of BMA with exchangeable members. Furthermore, we tested both approaches as part of this study (i.e., BMA with exchangeable and nonexchangeable members), and found that in this case both approaches yield very similar results. Hence, our focus here is going to be on the implementation of BMA with exchangeable members.

### 2) HCLR

HCLR is based on the logistic regression model initially proposed by Hamill et al. (2004) to postprocess precipitation ensembles. In essence, HCLR fits a logistic distribution to the transformed, in this case the cube root of the ensemble mean, and bias-corrected precipitation ensembles (Messner et al. 2014b). Note that the same cube root transformation is used for both HCLR and BMA. Additionally, HCLR uses the ensemble spread as a predictor, which allows HCLR to consider uncertainty information contained in the ensembles. We describe next the HCLR postprocessor as it evolved from the logistic regression model of Hamill et al. (2004) and the extended version of Wilks (2009).

The logistic regression model of Hamill et al. (2004) is given by

$$P(y \le q \,|\, x) = \Lambda[\delta(x)], \tag{7}$$

where $\Lambda(.)$ denotes the cdf of the standard logistic distribution, $y$ is the transformed precipitation, $q$ is a specified threshold, $x$ is a predictor variable that depends on the forecast members, and $\delta(x)$ is a linear function of the predictor variable $x$. Note that the variable $x$ could be replaced by a vector of predictor variables but here we use a single predictor as described in the next few paragraphs.

One limitation with Eq. (7) is that separate logistic regressions with different linear functions $\delta(x)$ need to be fitted to each threshold of interest (Wilks 2009). This results in logistic regressions that can cross each other that, in turn, implies the occurrence of nonsense negative probabilities. To overcome this limitation, Wilks (2009) extended the logistic regression model by adding another predictor variable for the threshold $q$ such that

$$P(y \le q \,|\, x) = \Lambda[\omega(q) - \delta(x)], \tag{8}$$

where the transformation $\omega(.)$ is a monotone nondecreasing function. In addition to avoiding negative probabilities, Eq. (8) has the advantage that fewer parameters need to be estimated; instead of having a linear function $\delta(x)$ for each threshold, $\delta(x)$ is now the same for

all the thresholds. This can be particularly relevant when dealing with small training datasets.

Furthermore, to appropriately utilize the uncertainty information in the ensemble spread, Messner et al. (2014a) proposed the HELR postprocessor. HELR uses an additional predictor vector $\varphi$ to control the dispersion of the logistic predictive distribution,

$$P(y \le q \mid x) = \Lambda\left\{\frac{\omega(q) - \delta(x)}{\exp[\eta(\varphi)]}\right\}, \tag{9}$$

where $\eta(.)$ is another linear function of the predictor variable $\varphi$. Note that $\varphi$ could be replaced by a vector of predictor variables. The exponential function in the denominator of Eq. (9) is used as a simple method to ensure positive values (Messner et al. 2014a).

In HELR, the function $\delta(x)$ is defined as

$$\delta(x) = d_0 + d_1 x, \tag{10}$$

where $d_0$ and $d_1$ are parameters that need to be estimated, and $x = 1/K\sum_{k=1}^{K} f_k^{1/3}$, that is, the predictor variable $x$ is the mean of the transformed, via the cube root, ensemble forecasts. $\eta(\varphi)$ is defined as

$$\eta(\varphi) = e_0 + e_1 \varphi, \tag{11}$$

where $e_0$ and $e_1$ are parameters that need to be estimated, and $\varphi$ is the standard deviation of the cube root transformed precipitation ensemble forecasts.

To determine the parameters associated with Eq. (9), maximum likelihood estimation with the log-likelihood function is used (Messner et al. 2014a,b). For this, one needs to determine the predicted probability $\pi_i$ of the $i$th observed outcome. When determining $\pi_i$, one should account for the fact that $y \ge 0$. One variation of the HELR postprocessor that can easily accommodate nonnegative variables that are continuous for positive values and have a natural threshold at zero, such as precipitation amounts, is censored regression or, as termed by Messner et al. (2014b), HCLR. For HCLR, $\pi_i$ can be expressed as (Messner et al. 2014b)

$$\pi_i = \begin{cases} \Lambda\left\{\dfrac{\omega(0) - \delta(x)}{\exp[\eta(\varphi)]}\right\} & y_i = 0 \\ \lambda\left\{\dfrac{\omega(y_i) - \delta(x)}{\exp[\eta(\varphi)]}\right\} & y_i > 0, \end{cases} \tag{12}$$

where $\lambda\{.\}$ denotes the likelihood function of the standard logistic function. In essence, HCLR fits a logistic error distribution with point mass at zero to the transformed predictand. Such an error distribution appears reasonable for dealing with the transformed precipitation amounts (Schefzik et al. 2013; Scheuerer 2014). As was the case with BMA, to implement the HCLR postprocessor, we use 24-h precipitation accumulations from the GEFSRv2 for lead times from 24 to 120 h. To train the HCLR postprocessor, the same sliding time window approach as in BMA is used.

### e. Verification strategy

To verify the raw and postprocessed ensemble precipitation forecasts, we use the Ensemble Verification System (EVS) (Brown et al. 2010). Also, different metrics are used for the verification analysis, including the Brier skill score (BSS), continuous ranked probability skill score (CRPSS), and reliability diagram. The decomposed components of the CRPS are also examined. The definition of each of these metrics is provided in the appendix. Additional details about the verification metrics can be found elsewhere (e.g., Wilks 2010; Jolliffe and Stephenson 2012). To assess the uncertainty of the verification metrics, the 90% error bars are computed using the block bootstrapping technique (Politis and Romano 1994).

For the verification analysis, we use two years of data, 2006 and 2007; the remaining years, 2002–05, are used to train the postprocessors. The verification is done conditionally upon the season, lead time, and precipitation threshold. The summer (June–August) and fall (September–November) months are the two seasons considered for lead times from 24 to 120 h. For the precipitation threshold, a low (precipitation $> 0$ mm) and high (precipitation $> 10$ mm) precipitation threshold are used. For the low and high precipitation threshold, precipitation amounts greater than that implied by a nonexceedance probability, in the sampled climatological probability distribution, of 0.3 ($\sim$0 mm) and 0.9 ($\sim$10 mm) are selected, respectively. Additionally, we assess the effect on the postprocessed forecasts of spatially pooling data to train the postprocessors.

## 3. Results

### a. Selection of the training length for BMA

An initial step in implementing the BMA postprocessor is to determine the appropriate training length for the sliding time window approach of BMA (Fraley et al. 2010; Sloughter et al. 2007). If the length of the training window is too short or too long, the performance of BMA can become suboptimal or less skillful. To assess the effect of the training length on the performance of BMA, we plot the BSS (relative to sampled climatology) against the training length for the low precipitation threshold ($>$0 mm) in the summer (Figs. 2a

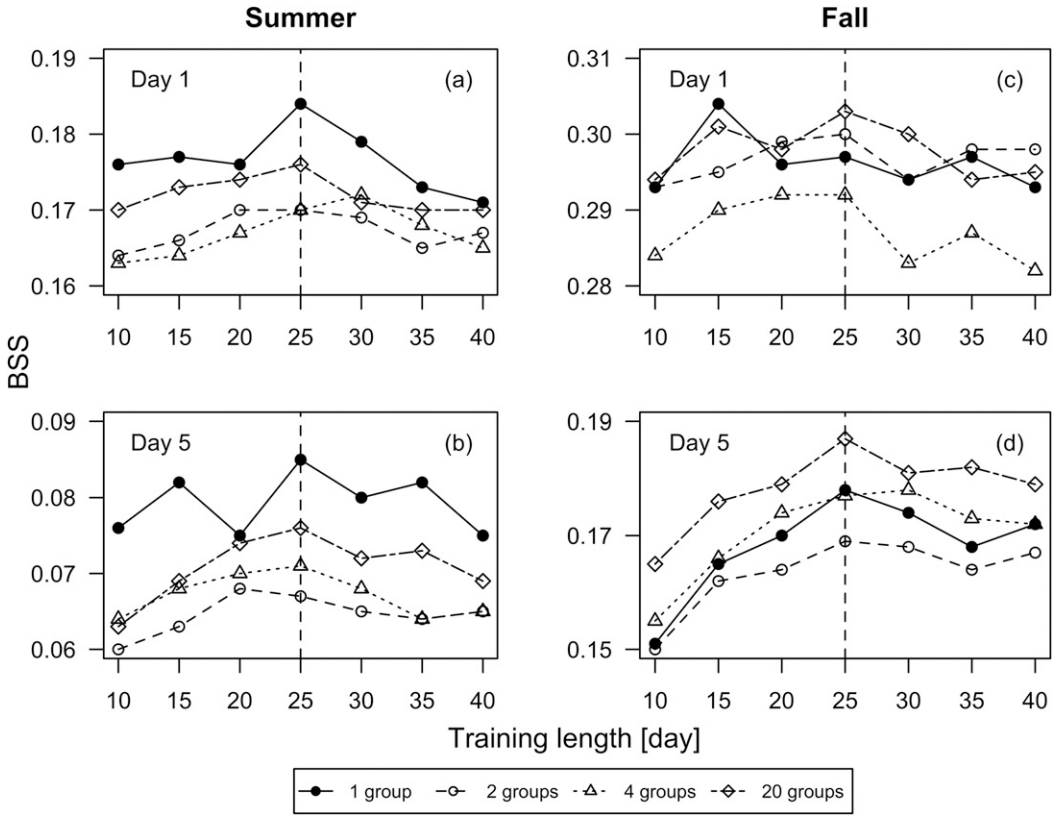


FIG. 2. BSS (relative to sampled climatology) for the low precipitation threshold (>0 mm) vs the BMA training length for forecast lead times of (a) 1 and (b) 5 days during the summer and lead times of (c) 1 and (d) 5 days during the fall. The different BSS curves represent the number of cells used to train the BMA.

and 2b) and fall (Figs. 2c and 2d). We find that the BSS tends to peak or reach a maximum value at a training length of ~25 days [i.e., (25 days) × (4 years) (Fig. 2)]. For the most part, after 25 days the value of BSS declines (Fig. 2). This is the case for both forecast lead times of 1 (Figs. 2a and 2c for the summer and fall, respectively) and 5 days (Figs. 2b and 2d for the summer and fall, respectively). The results are similar independently of the number of GEFSRv2 cells used to train the BMA algorithm (Fig. 2) (i.e., the optimum value of the training length still tends to be ~25 days). For example, in Fig. 2a, when using 20 cells or training each cell separately (1 cell), both curves reach a maximum at 25 days.

Figure 3 shows the same information as Fig. 2 but plots instead the CRPSS (relative to sampled climatology) against the training length. In Fig. 3, the general tendency is as in Fig. 2, the skill of the BMA postprocessed forecasts tends to reach a maximum at ~25 days. We also evaluated (not shown) the effect of the training window length on the HCLR postprocessor and found that it does not have a significant impact on the performance of HCLR. Because of this, and to implement the postprocessors under similar conditions, hereafter the same window length of 25 days is used to train and implement both postprocessors.

### b. *Effect of spatial pooling on the performance of the postprocessors*

To assess the effect of spatial pooling on the performance of the postprocessors, we plot the BSS (relative to sampled climatology) against the number of cells used to train the BMA and HCLR postprocessors (Figs. 4a and 4b for the summer and fall, respectively). Note that the same training window length of 25 days is used for BMA and HCLR. The forecasts from both postprocessors show notable gains in skill relative to the raw ensembles for the summer (Fig. 4a) but the gains seem largely insignificant for the fall (Fig. 4b). The general tendency in Fig. 4, nevertheless, is for the BSS to marginally decline as the number of cells used for training are increased. Additionally, the HCLR seems to perform slightly better than BMA (Fig. 4a).

We also show the CRPSS (relative to sampled climatology) as a function of the number of cells used to train the BMA and HCLR postprocessors for the summer (Fig. 5a) and fall (Fig. 5b). For the summer (Fig. 5a),

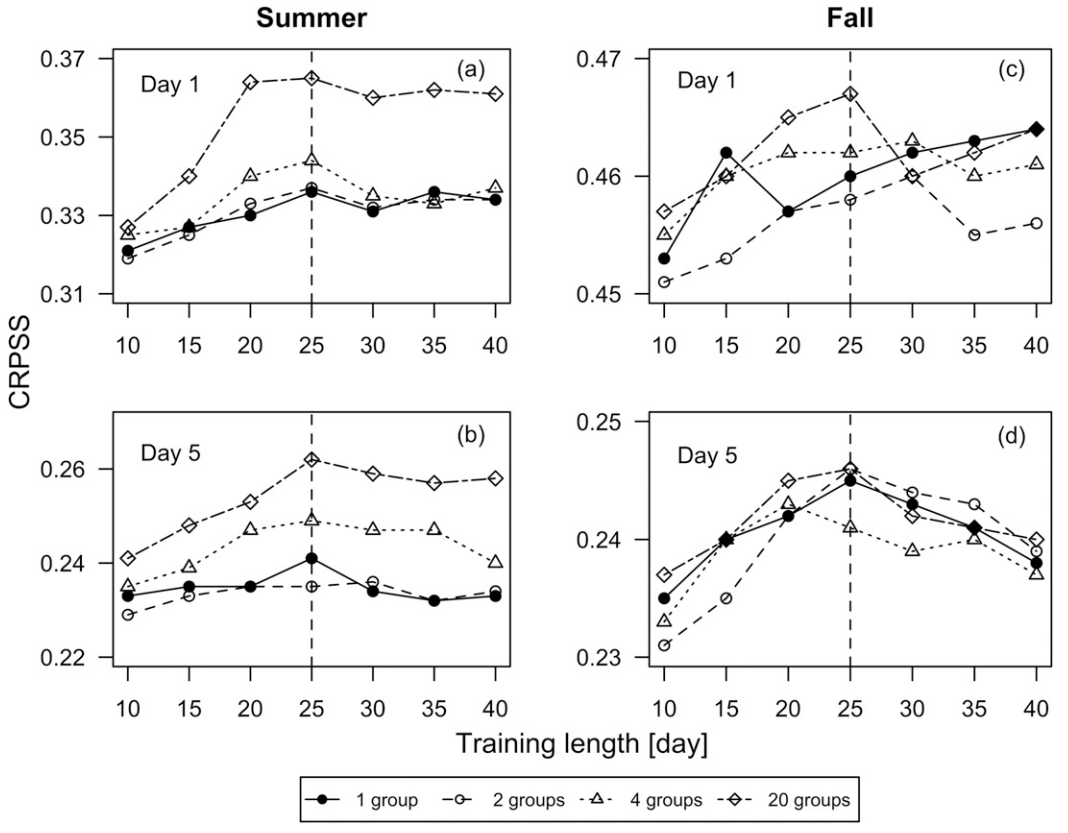


FIG. 3. CRPSS (relative to sampled climatology) vs the BMA training length for forecast lead times of (a) 1 and (b) 5 days during the summer and lead times of (c) 1 and (d) 5 days during the fall. The different CRPSS curves represent the number of cells used to train the BMA.

the postprocessors seem to significantly improve upon the raw ensembles, and the skill declines slightly as additional cells are used to train the postprocessors, as was the case with the BSS (Fig. 4a). For the fall (Fig. 5b), only HCLR seems able to improve upon the raw ensembles. However, overall the differences in skill between the postprocessors appear not as significant, particularly in the summer (Fig. 5a).

According to the results in Figs. 4 and 5, for the remainder of our analysis, we train the postprocessors separately at each GEFSRv2 cell since this approach seems to perform somewhat better than when cells are spatially pooled. Note that this is different from the way BMA is normally implemented (Sloughter et al. 2007; Fraley et al. 2010). Spatial pooling is normally required by BMA to increase the sample size used for training because the typical training window length of 25–30 days is small. We are less constrained here by the length of the training window since we sample data from the previous four years when training the postprocessors. This is feasible in this case because we are working with reforecasts but it may not be as feasible when dealing with outputs from an actual forecasting system. Another reason why the single cell training for BMA performs better here than in previous studies (Sloughter et al. 2007; Fraley et al. 2010) may be due to the fact that our application, based on a global forecasting system, relies on lower-spatial-resolution forecasts than previous ones. Previous applications have tended to rely on higher-resolution regional forecasting systems where neighboring cells may be more similar to each other than in the GEFSRv2.

### c. *Verification of the raw and postprocessed precipitation ensembles*

#### 1) BSS

The BSS (relative to sampled climatology) indicates that generally the skill of the postprocessed ensemble precipitation forecasts is improved relative to the raw ensembles (Fig. 6). The relative improvements in skill are generally greater in the summer (Figs. 6a and 6b) than fall (Figs. 6c and 6d). Additionally, the improvements tend to be greater for the low precipitation threshold (>0 mm) (Fig. 6c) than the high threshold (>10 mm) (Fig. 6d). Overall, the skills gain from

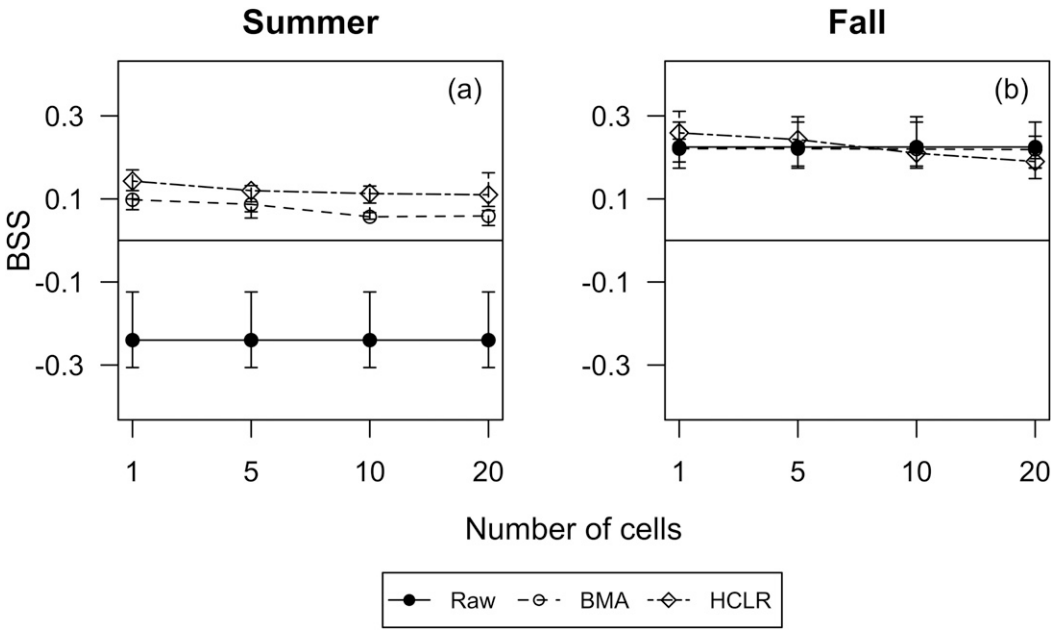


FIG. 4. BSS (relative to sampled climatology) for the high precipitation threshold (>10 mm) vs the number of cells used to train the postprocessors during the (a) summer and (b) fall. The different BSS curves represent the raw and postprocessed precipitation ensembles. The figure is for a forecast lead time of 4 days.

postprocessing decline with increasing lead time. For example, for the high precipitation threshold (>10 mm) in the fall (Fig. 6d), the BSS associated with the postprocessors is slightly better than the BSS of the raw ensembles at a forecast lead time of 1 day; however, the BSS of the postprocessed ensembles becomes slightly less at a lead time of 5 days. Contrasting the postprocessors against each other, it appears that the general tendency is for the postprocessors to perform similarly (Fig. 6). The HCLR, however, tends to show a slight skill

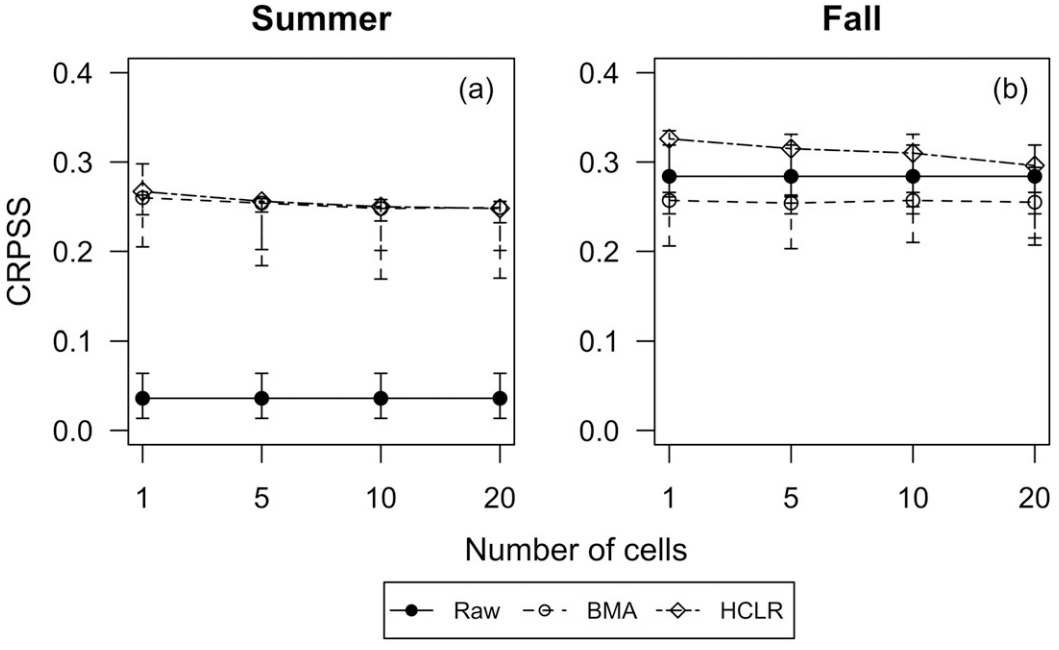


FIG. 5. CRPSS (relative to sampled climatology) vs the number of cells used to train the postprocessors during the (a) summer and (b) fall. The different CRPSS curves represent the raw and postprocessed precipitation ensembles. The figure is for a forecast lead time of 5 days.

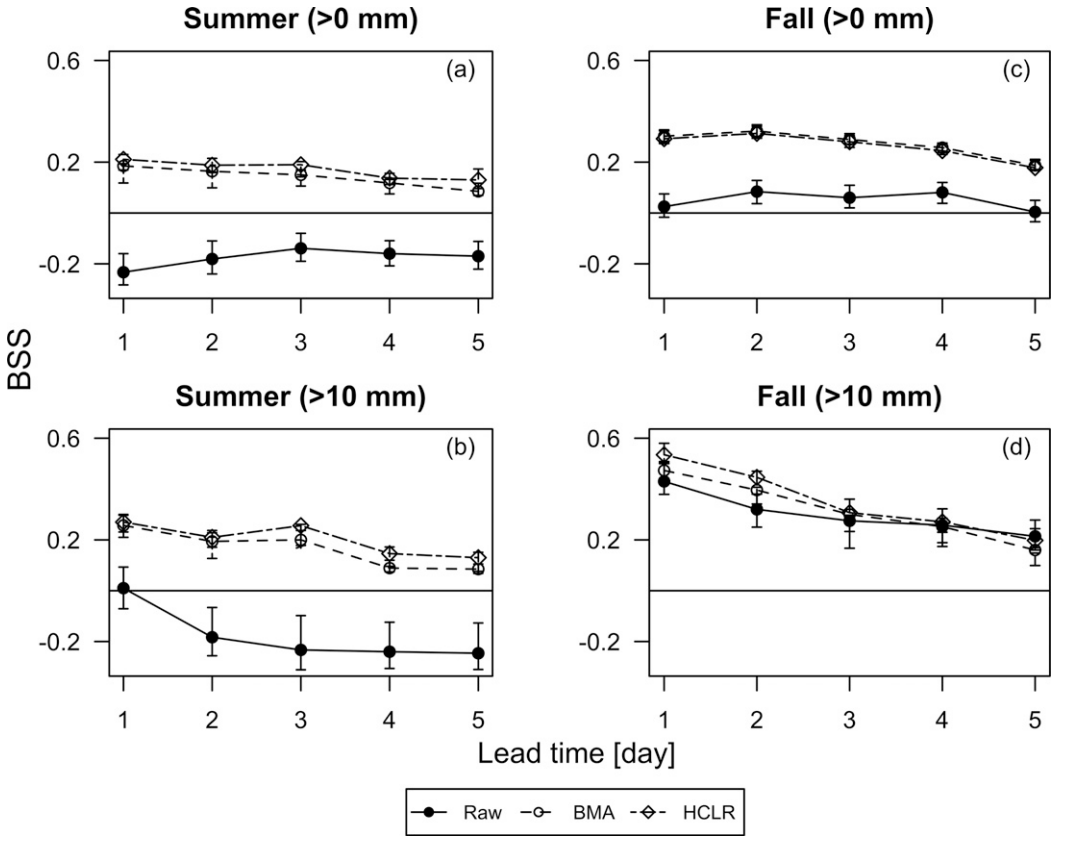


FIG. 6. BSS (relative to sampled climatology) for the (a) low (>0 mm) and (b) high (>10 mm) precipitation threshold during the summer vs the forecast lead time. BSS for the (c) low (>0 mm) and (d) high (>10 mm) precipitation threshold during the fall vs the forecast lead time. The different BSS curves represent the raw and postprocessed precipitation ensembles.

gain over BMA across lead times, precipitation thresholds, and seasons (Fig. 6).

The plot of the BSS (relative to sampled climatology) against the nonexceedance probability associated with different precipitation thresholds (Fig. 7) further confirms the findings from Fig. 6. It demonstrates that for the most part the postprocessors behave similarly with respect to each other. Additionally, the trend in the BSS for the postprocessed forecasts tends to mimic the behavior of the raw ensembles during the fall but not the summer. For example, the BSS values, for both the raw and postprocessed forecasts in the fall, tend to increase with the precipitation threshold (Fig. 7c) while differences in the trend between the raw and postprocessed forecasts are evident in the summer (Fig. 7b). Also, as was the case in Fig. 6, the gains in skill from postprocessing are somewhat more noticeable in the summer (Figs. 7a and 7b) than fall (Figs. 7c and 7d), and generally the gains in skill are reduced for the longer forecast lead times (e.g., day 2 in Fig. 7c and day 5 in Fig. 7d). Indeed, at a lead time of 2 days in the fall (Fig. 7c), the postprocessed ensembles outperform the raw ensembles across probability thresholds. In contrast, at a lead time of 5 days in the fall (Fig. 7d), the postprocessed ensembles outperform the raw ensembles for probability thresholds less than 0.9 but at a probability threshold of 0.9 the raw ensembles exhibit a slightly better skill than the postprocessed ensembles.

### 2) CRPSS

The CRPSS (relative to sampled climatology) shows that the postprocessed precipitation ensembles are overall more skillful than the raw ensembles across lead times and seasons (Fig. 8). As was the case with the BSS (Figs. 6 and 7), the relative gains in skill from postprocessing are greater in the summer (Fig. 8a) than in the fall (Fig. 8b), but the overall skill of the raw as well as postprocessed ensembles is significantly better in the fall than the summer. Furthermore, contrasting the postprocessors against each other, HCLR tends to slightly outperform BMA. Indeed, for the fall, only HCLR shows improvements upon the raw ensembles at a forecast lead time of 5 days.

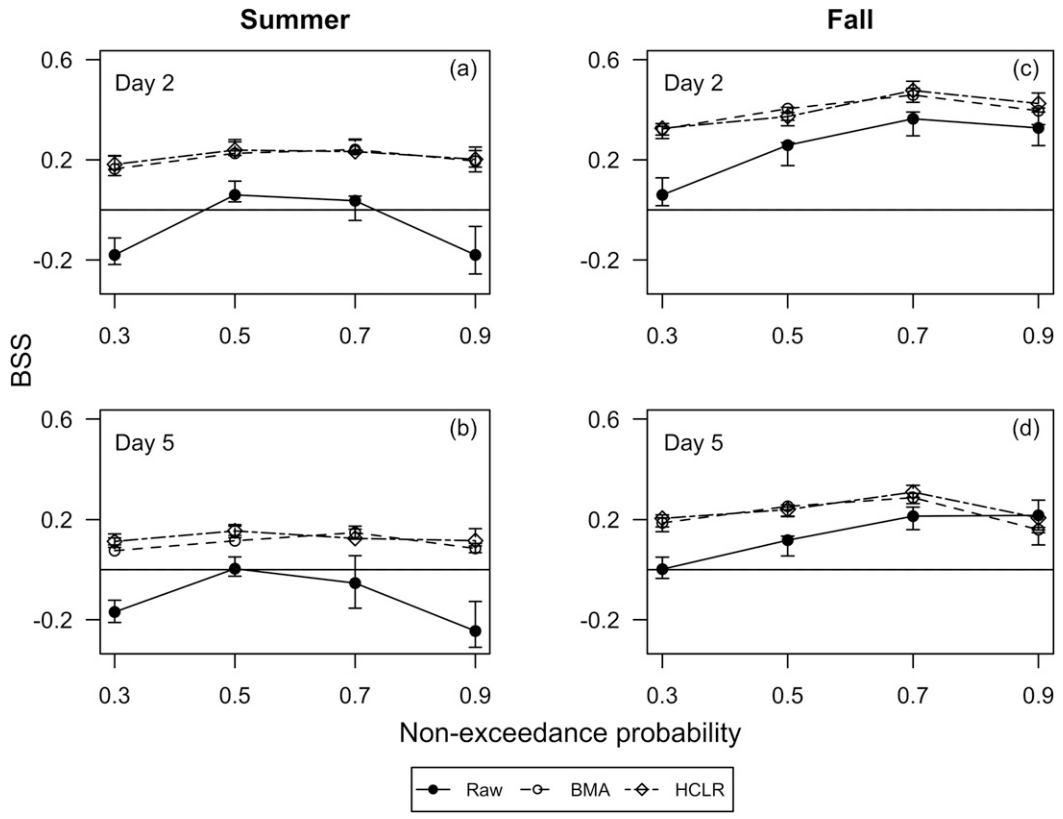


FIG. 7. BSS (relative to sampled climatology) vs the precipitation threshold for forecast lead times of (a) 2 and (b) 5 days during the summer and forecast lead times of (c) 2 and (d) 5 days during the fall. The different BSS curves represent the raw and postprocessed precipitation ensembles.

The CRPS can be decomposed into a reliability ($CRPS_{rel}$) and potential ($CRPS_{pot}$) component (Hersbach 2000). The $CRPS_{rel}$ measures the ability of the precipitation ensembles to generate cumulative distributions that have, on average, the correct or desired statistical properties. While the $CRPS_{pot}$ measures the CRPS that one would obtain for a perfect reliable system. The decomposition of the CRPS shows that the gains in skill from postprocessing are mainly related to improvements in $CRPS_{rel}$ (Fig. 9a). Note that the CRPS, $CRPS_{rel}$, and $CRPS_{pot}$ have a negative orientation (i.e., smaller values are better). The CRPS decomposition reveals that the gains are considerably greater in the summer (Fig. 9a) than fall (Fig. 9b). It also shows that HCLR tends to have similar or even larger $CRPS_{rel}$ (Fig. 9a) than BMA but a smaller $CRPS_{pot}$. The reduction in $CRPS_{pot}$ is the main source of improvement for HCLR over BMA. This means that, in relation to the sampled climatology, the resolution associated with HCLR is likely better than that of BMA. This may be partly due to the fact that HCLR uses the ensemble spread explicitly as a predictor of the dispersion of the predictive pdf (Messner et al. 2014a) and the $CRPS_{pot}$ is sensitive to the spread (Hersbach 2000). The CRPS decomposition also illustrates the fact that BMA can improve the reliability of the forecasts relative to the raw ensembles while at the same time reducing the overall skill of the forecasts. This is observed in Fig. 9b at a forecast lead time of 5 days where BMA has slightly lower $CRPS_{rel}$ than the raw ensembles but much higher $CRPS_{pot}$.

### 3) RELIABILITY

According to the CRPS decomposition (Fig. 9), the postprocessed ensemble precipitation forecasts tend to be more reliable than the raw ensembles. This is further confirmed using reliability diagrams under various forecasting conditions (Fig. 10). In Fig. 10, the reliability of the postprocessed forecasts from BMA and HCLR is improved relative to the raw ensembles across forecast probabilities, lead times, and seasons. There is, however, a tendency to underforecast the small forecast probabilities in the summer (Fig. 10b) and fall (Fig. 10d) (i.e., the postprocessed forecasts tend to be somewhat underconfident). This tendency is significantly more apparent in the raw ensembles than in the postprocessed ones

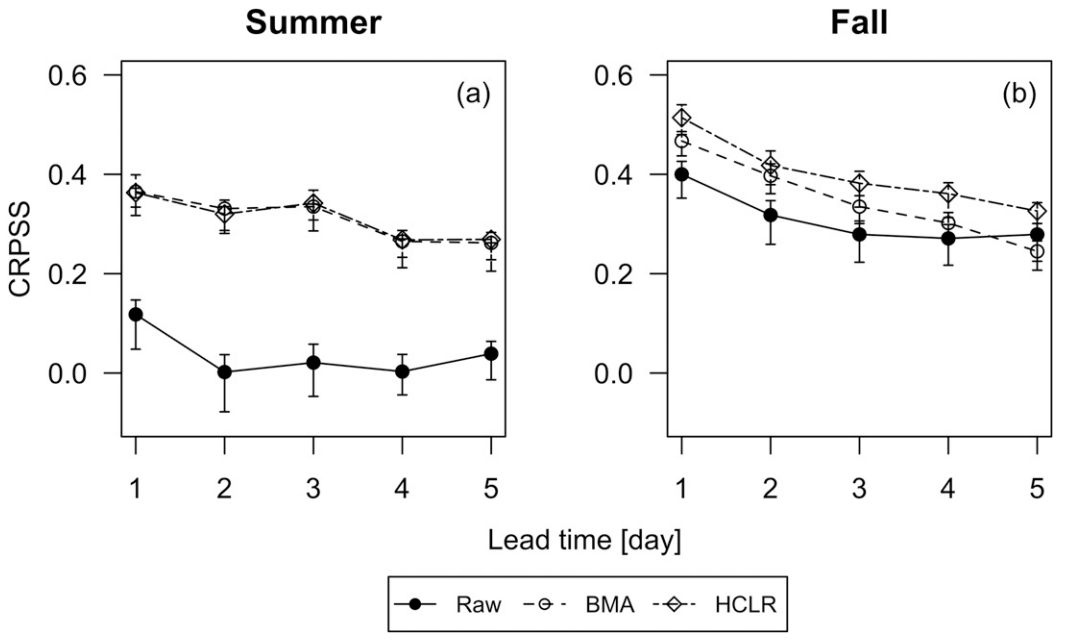


FIG. 8. CRPSS (relative to sampled climatology) for the ensemble precipitation forecasts vs the forecast lead time during the (a) summer and (b) fall. The different CRPSS curves represent the raw and postprocessed precipitation ensembles.

(Fig. 10a). For the larger forecast probabilities, the raw ensembles tend to overforecast the forecast probabilities, that is, the forecasts are overconfident, while the postprocessed ones seem, for the most part, to fix this overforecasting bias (Fig. 10c).

Contrasting BMA and HCLR against each other, they both show similar reliability and sharpness (assessed by examining the insets in Fig. 10). The reliability of the postprocessors does not seem to vary greatly with the season (Figs. 10a and 10c) or forecast lead time (Figs. 10a and 10b). It does vary, however, with the precipitation threshold. The reliability curves associated with each of the postprocessors show more variability for the high precipitation threshold (>10 mm)

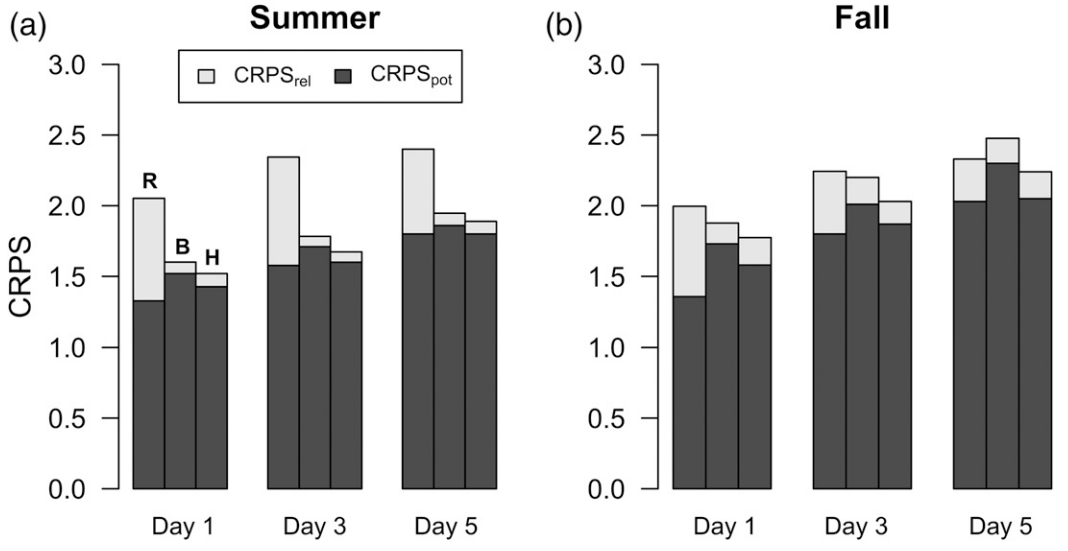


FIG. 9. Decomposition of the CRPS into CRPS reliability ($CRPS_{rel}$) and CRPS potential ($CRPS_{pot}$) for forecasts lead times of 1, 3, and 5 days during the (a) summer and (b) fall. (from left to right) The three columns associated with each forecast lead time represent the raw (R), BMA postprocessed (B), and HCLR postprocessed (H) precipitation ensembles.

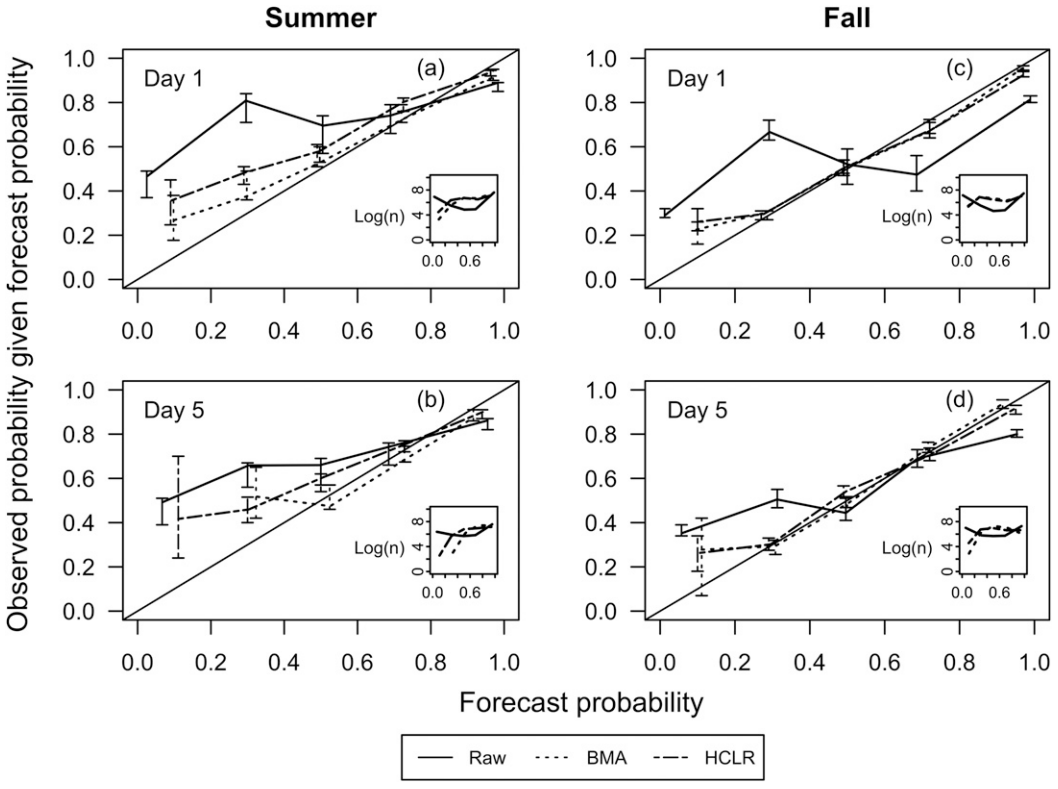


FIG. 10. Reliability diagrams for the low precipitation threshold (>0 mm) in the summer and forecast lead times of (a) 1 and (b) 5 days. Reliability diagrams for the low precipitation threshold (>0 mm) in the fall and forecast lead times of (c) 1 and (d) 5 days. The different reliability curves represent the raw and postprocessed precipitation ensembles. The insets show the sample size in logarithmic (base 10) scale of the different forecast probability bins.

(Fig. 11) than the low one (>0 mm) (Fig. 10), which seems to be due mainly to sample size effects. For the high precipitation threshold, the raw ensembles are strongly overconfident; they overforecast the larger forecast probabilities (Figs. 11a and 11c). The overforecasting is stronger in the summer (Figs. 11a and 11b) than the fall (Figs. 11c and 11d). Nonetheless, the reliability of the postprocessors is overall similar for the high precipitation threshold (Fig. 11). In some cases, forecasts from BMA seem more reliable than forecasts from HCLR (Figs. 10c and 11a) while in other cases HCLR is more reliable (Figs. 10a, 10b, and 11c). Both BMA and HCLR are, overall, able to improve the biases in the raw ensembles to make them more reliable.

## 4. Summary and discussion

Ensemble forecasts can be used to determine the probability and uncertainty of a weather variable. In the case of ensemble precipitation forecasts, the determination of forecast probabilities from ensembles is generally unreliable, because the magnitude and dispersion of the ensemble forecasts are often characterized by the presence of biases (Messner et al. 2014a,b; Sloughter et al. 2007; Wilks 2009). Statistical postprocessing is, therefore, needed to correct the biases and improve the reliability of ensemble precipitation forecasts. In this study, we assessed the potential of BMA (Sloughter et al. 2007) and HCLR (Messner et al. 2014b) to postprocess precipitation ensembles from the 11-member GEFSRv2 dataset (Hamill et al. 2013; Sharma et al. 2017). As part of our experimental setting, we employed 24-h precipitation accumulations for lead times of 24 to 120 h over the U.S. MAR. We used MPEs as the observed precipitation.

To implement BMA, we first selected the length of the sliding time window and the number of cells needed to train the postprocessors. Using the BSS and CRPSS to assess the skill associated with different window lengths, we found that generally the optimum value tended to be ~25 days across lead times and seasons. Similar results have been reported by others (Fraley et al. 2010; Sloughter et al. 2007). We note that the sensitivity of the skill scores to the training window length was not large. Furthermore, since we used

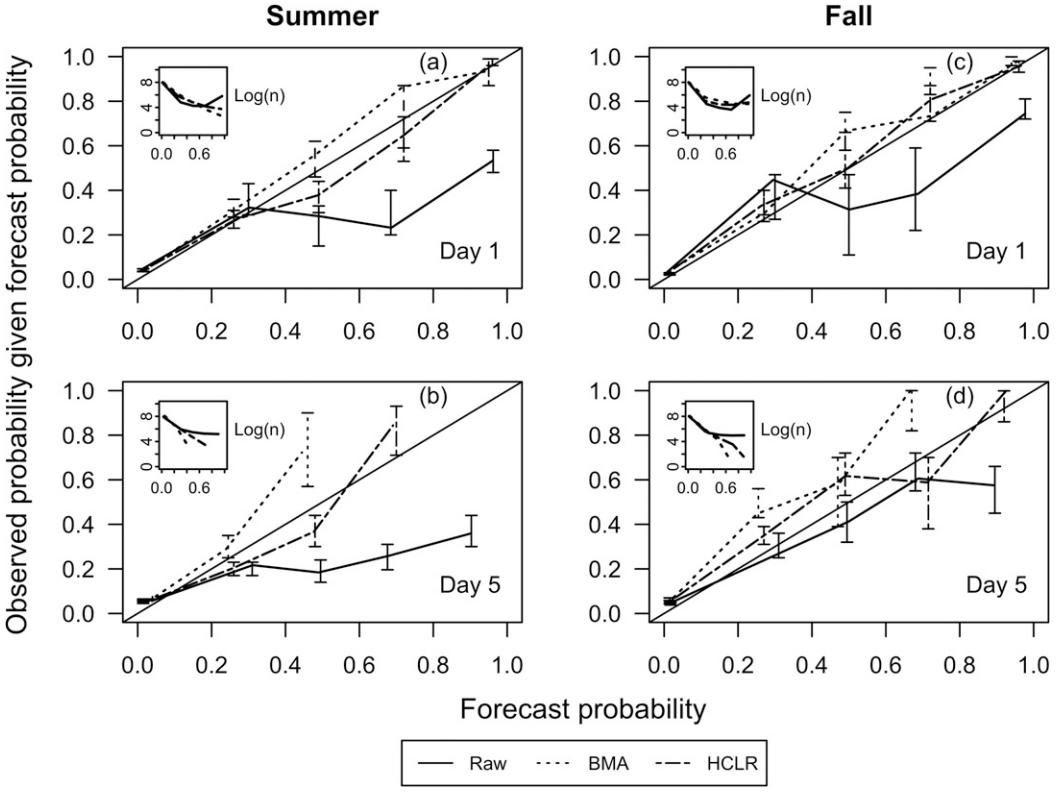


FIG. 11. Reliability diagrams for the high precipitation threshold (>10 mm) in the summer and forecast lead times of (a) 1 and (b) 5 days. Reliability diagrams for the high precipitation threshold (>10 mm) in the fall and forecast lead times of (c) 1 and (d) 5 days. The different reliability curves represent the raw and postprocessed precipitation ensembles. The insets show the sample size in logarithmic scale of the different forecast probability bins.

training data from different years to train the BMA, the effective training length is much greater than 25 days. In terms of the number of cells, we found that training each cell in the GEFSRv2 separately yielded slightly more skillful forecasts than when spatially pooling data from several cells. This may be partly the case here since we sampled data from the previous four years when training the postprocessors, which potentially makes spatial pooling less effective. But relying on past forecasts to train the postprocessors may not always be feasible, particularly when dealing with operational forecasting systems. To implement HCLR, we used the same sliding window approach as in BMA with the same window length of 25 days and training each GEFSRv2 cell separately. We observed that the effect of the window length on the performance of HCLR was not significant. Thus, the same length was used to train both BMA and HCLR.

We used the BSS, CRPSS, and reliability diagrams, conditioned upon the lead time, precipitation threshold, and season, to compare against the raw ensembles and each other the BMA, and HCLR postprocessors. From this comparison, we found that overall there is a slight tendency for HCLR to outperform BMA but the differences appear to be not as significant. They become more apparent at the longer forecast lead times (e.g., 5 days) during both the summer and fall.

Although the differences in performance between BMA and HCLR are not as significant, there are other modeling and implementation differences between the two postprocessors that are worth emphasizing. One key difference between BMA and HCLR lies in the predictive pdfs that the postprocessors use. BMA uses a mixed pdf comprising a gamma pdf for the transformed precipitation amounts that are greater than zero and a standard logistic pdf for the point mass at zero. While HCLR uses a censored logistic pdf for all the transformed precipitation amounts. Although the predictive pdfs used by BMA and HCRL have a similar form (point mass at zero with continuous pdf for precipitation values greater than zero), the pdf used by BMA is generally more flexible than that of HCLR, but it also requires the estimation of a greater number of parameters. For the case of exchangeable members, BMA requires 7 parameters plus the member weights (11 in this case); in contrast, HCLR

requires only 4 parameters. The use of a greater number of parameters makes BMA a more complex model than HCLR. Note that here the additional model flexibility and complexity of BMA is not able to improve ensembles beyond the level provided by HCLR (e.g., Figs. 7–9). This suggests that such flexibility and complexity may not be necessary or warranted in this case. Additionally, when the training dataset is small, the use of too many model parameters could lead to overfitting. There are, nonetheless, important advantages to the BMA postprocessor that were not evaluated in our case study, such as the possibility to represent multimodal predictive pdfs and to allow, in a consistent manner, the incorporation of ensembles members from different forecasting systems.

Another difference between the two postprocessors is in the way they consider the ensemble spread. In BMA, the individual ensemble members are dressed with the predictive pdf. Thus, if the raw ensembles are very different from each other (i.e., the spread is wide), the BMA will yield wider predictive distributions, likely characterized by multimodality (Raftery et al. 2005). In the case of HCLR, the ensemble spread is adjusted directly, which may be a desirable trait since this is an ensemble feature that often requires significant manipulation.

## 5. Conclusions

Based on our study results, the following main conclusions are emphasized:

- In terms of the forecast skill (i.e., BSS and CRPSS), the postprocessors show significant gains relative to the raw ensembles in the summer across lead times while gains are less significant in the fall. But overall the raw and postprocessed ensembles are more skillful in the fall than summer. This is probably due to the nature of summertime precipitation as being more convective, and therefore less predictable, whereas fall precipitation tends to be more organized at synoptic (larger) scales.
- The reliability diagrams showed that the postprocessors are able to correct biases in the raw ensembles that ultimately make the postprocessed ensembles more reliable than the raw ones across lead times, precipitation thresholds, and seasons. Both postprocessors result in forecasts with similar reliability.
- By decomposing the CRPS into a reliability ($\mathrm{CRPS_{rel}}$) and potential ($\mathrm{CRPS_{pot}}$) component, we were able to examine more carefully the differences between BMA and HCLR. From this, we observed that the improved performance of HCLR over that of BMA is due to having a lower $\mathrm{CRPS_{pot}}$. Indeed, the $\mathrm{CRPS_{rel}}$ component tends to be slightly lower (better) for BMA than HCLR. We also note that, based on the decomposition of the CRPS, HCLR is the only postprocessor to consistently improve upon the raw ensembles across lead times and seasons.
- In summary, based on our analysis and comparison, we found that generally the postprocessors perform similarly. A future alternative could be to combine the strengths of both BMA and HCLR (e.g., by using HCLR to determine the predictive pdf of each forecasting system and BMA to weight the pdfs). However, this may come at a considerable computational cost, particularly when considering a range of lead times and multiyear reforecasts datasets. Another option that could be explored is to apply the HCLR postprocessor to each individual ensemble member rather than relying on the mean of the ensemble forecasts, this could be used to implement HCLR with different forecasting systems.

*Acknowledgments.* The second, third, and last author gratefully acknowledge the funding support provided by the NOAA's NWS through Award NA14NWS4680012. The authors also acknowledge the computational support provided by the Institute for CyberScience at The Pennsylvania State University.

## APPENDIX

### Verification Metrics

*a. Brier skill score (BSS)*

The Brier score (BS) is analogous to the mean squared error, but where the forecast is a probability and the observation is either a 0 or 1 (Brown et al. 2010). The BS is given by

$$\mathrm{BS} = \frac{1}{n}\sum_{i=1}^{n}[F_{X_i}(q) - F_{Y_i}(q)]^2, \tag{A1}$$

where the probability of $X_i$ to exceed a fixed threshold $q$ is

$$F_{X_i}(q) = P(X_i > q), \tag{A2}$$

$n$ is again the total number of forecast–observation pairs, and

$$F_{Y_i}(q) = \begin{cases} 1, & Y_i > q; \\ 0, & \text{otherwise}. \end{cases} \tag{A3}$$

To compare the skill score of the main forecast system with respect to the reference forecast, it is convenient to define the Brier skill score (BSS):

$$\mathrm{BSS} = 1 - \frac{\mathrm{BS}_{\mathrm{main}}}{\mathrm{BS}_{\mathrm{reference}}}, \tag{A4}$$

where $\mathrm{BS}_{\mathrm{main}}$ and $\mathrm{BS}_{\mathrm{reference}}$ are the BS values for the main forecasting system (i.e., the system to be evaluated) and reference forecasting system, respectively. Any positive values of the BSS, from 0 to 1, indicate that the main forecasting system performed better than the reference forecasting system. Thus, a BSS of 0 indicates no skill and a BSS of 1 indicates perfect skill.

*b. Reliability diagram*

As suggested by Murphy (1973), the BS can be further decomposed into a reliability, resolution and uncertainty component. In this study, instead of using the decomposed BS to quantify the reliability and resolution of the forecasts, we use the so-called reliability diagram. The reliability diagram shows the full joint distribution of forecasts and observations to reveal the reliability of the probability forecasts. For the forecast values portioned into bin $B_k$ and defined by the exceedance of threshold $q$, the average forecast probability can be expressed as

$$\overline{F}_{X_k}(q) = \frac{1}{|I_k|}\sum_{I_k} F_{X_i}(q), \quad \text{where} \quad I_k = \{i: X_i \in B_k\}, \tag{A5}$$

where $I_k$ is the collection of all indices $i$ for which $X_i$ falls into bin $B_k$, and $|I_k|$ denotes the number of elements in $I_k$. The corresponding fraction of observations that fall in the $k$th bin is given by

$$\overline{F}_{Y_k}(q) = \frac{1}{|I_k|}\sum_{I_k} F_{Y_i}(q), \quad \text{where} \quad F_{Y_i}(q) = \begin{cases} 1, & Y_i > q; \\ 0, & \text{otherwise}. \end{cases} \tag{A6}$$

The reliability diagram plots $\overline{F}_{X_k}(q)$ against $\overline{F}_{Y_k}(q)$.

*c. Mean continuous ranked probability skill score (CRPSS)*

The continuous ranked probability score (CRPS), which is less sensitive to sampling uncertainty, is used to measure the integrated square difference between the cumulative distribution function (cdf) of a forecast $F_x(q)$, and the corresponding cdf of the observation $F_y(q)$. The CRPS is given by

$$\mathrm{CRPS} = \int_{-\infty}^{\infty} [F_x(q) - F_y(q)]^2 \, dq. \tag{A7}$$

To evaluate the skill of the main forecasting system relative to the reference forecast system, the associated skill score, the mean continuous ranked probability skill score (CRPSS), is defined as

$$\mathrm{CRPSS} = 1 - \frac{\mathrm{CRPS}_{\mathrm{main}}}{\mathrm{CRPS}_{\mathrm{reference}}}, \tag{A8}$$

where CRPS is averaged across $n$ pairs of forecasts and observations to calculate the mean CRPS of the main forecast system ($\mathrm{CRPS}_{\mathrm{main}}$) and reference forecast system ($\mathrm{CRPS}_{\mathrm{reference}}$). The CRPSS ranges from $-\infty$ to 1, with negative scores indicating that the system to be evaluated has worse CRPS than the reference forecasting system, while positive scores indicate a higher skill for the main forecasting system in comparison to the reference forecasting system, with 1 indicating perfect skill.

Additionally, to further explore the effect of postprocessing on forecast skill, we separate the $\mathrm{CRPS}_{\mathrm{main}}$ into different components according to the procedure developed by Hersbach (2000). Specifically, we consider the CRPS reliability ($\mathrm{CRPS}_{\mathrm{rel}}$) and potential ($\mathrm{CRPS}_{\mathrm{pot}}$) such that

$$\mathrm{CRPS}_{\mathrm{main}} = \mathrm{CRPS}_{\mathrm{rel}} + \mathrm{CRPS}_{\mathrm{pot}}. \tag{A9}$$

The $\mathrm{CRPS}_{\mathrm{rel}}$ measures the ability of the precipitation ensembles to generate cumulative distributions that have, on average, the correct or desired statistical properties. The reliability is closely connected to the rank histogram, which shows whether the frequency that the verifying analysis was found in a given bin is equal for all bins (Hersbach 2000). The $\mathrm{CRPS}_{\mathrm{pot}}$ measures the CRPS that one would obtain for a perfect reliable system. It is sensitive to the average spread of the ensemble and outliers. For instance, the narrower the spread of the ensemble is, the smaller the $\mathrm{CRPS}_{\mathrm{pot}}$ becomes. As indicated by Hersbach (2000), provided a certain degree of unpredictability, a balance between the ensemble spread and the statistics of outliers will result in the optimal value of the $\mathrm{CRPS}_{\mathrm{pot}}$.

## REFERENCES

Bremnes, J. B., 2004: Probabilistic forecasts of precipitation in terms of quantiles using NWP model output. *Mon. Wea. Rev.*, **132**, 338–347, doi:10.1175/1520-0493(2004)132<0338:PFOPIT>2.0.CO;2.

Bröcker, J., and L. A. Smith, 2008: From ensemble forecasts to predictive distribution functions. *Tellus*, **60A**, 663–678, doi:10.1111/j.1600-0870.2008.00333.x.

Brown, J. D., and D.-J. Seo, 2010: A nonparametric postprocessor for bias correction of hydrometeorological and hydrologic ensemble forecasts. *J. Hydrometeor.*, **11**, 642–665, doi:10.1175/2009JHM1188.1.

——, J. Demargne, D.-J. Seo, and Y. Liu, 2010: The Ensemble Verification System (EVS): A software tool for verifying ensemble forecasts of hydrometeorological and hydrologic variables at discrete locations. *Environ. Model. Software*, **25**, 854–872, doi:10.1016/j.envsoft.2010.01.009.

Buizza, R., P. L. Houtekamer, G. Pellerin, Z. Toth, Y. Zhu, and M. Wei, 2005: A Comparison of the ECMWF, MSC, and NCEP Global Ensemble Prediction Systems. *Mon. Wea. Rev.*, **133**, 1076–1097, doi:10.1175/MWR2905.1.

Clark, M. P., and L. E. Hay, 2004: Use of medium-range numerical weather prediction model output to produce forecasts of streamflow. *J. Hydrometeor.*, **5**, 15–32, doi:10.1175/1525-7541(2004)005<0015:UOMNWP>2.0.CO;2.

Dempster, A. P., N. M. Laird, and D. B. Rubin, 1977: Maximum likelihood from incomplete data via the EM algorithm. *J. Roy. Stat. Soc.*, **39B**, 1–38.

Erickson, M. J., B. A. Colle, and J. J. Charney, 2012: Impact of bias-correction type and conditional training on Bayesian model averaging over the northeast United States. *Wea. Forecasting*, **27**, 1449–1469, doi:10.1175/WAF-D-11-00149.1.

Fraley, C., A. E. Raftery, and T. Gneiting, 2010: Calibrating multimodel forecast ensembles with exchangeable and missing members using Bayesian model averaging. *Mon. Wea. Rev.*, **138**, 190–202, doi:10.1175/2009MWR3046.1.

Friederichs, P., and A. Hense, 2007: Statistical downscaling of extreme precipitation events using censored quantile regression. *Mon. Wea. Rev.*, **135**, 2365–2378, doi:10.1175/MWR3403.1.

Glahn, H. R., and D. A. Lowry, 1972: The use of model output statistics (MOS) in objective weather forecasting. *J. Appl. Meteor.*, **11**, 1203–1211, doi:10.1175/1520-0450(1972)011<1203:TUOMOS>2.0.CO;2.

Greene, E. A., A. E. LaMotte, and K.-A. Cullinan, 2005: Groundwater vulnerability to nitrate contamination at multiple thresholds in the mid-Atlantic region using spatial probability models. Scientific Investigations Rep. 2004-5118, U.S. Department of the Interior, USGS, 24 pp.

Greybush, S. J., S. E. Haupt, and G. S. Young, 2008: The regime dependence of optimally weighted ensemble model consensus forecasts of surface temperature. *Wea. Forecasting*, **23**, 1146–1161, doi:10.1175/2008WAF2007078.1.

Hamill, T. M., J. S. Whitaker, and X. Wei, 2004: Ensemble reforecasting: Improving medium-range forecast skill using retrospective forecasts. *Mon. Wea. Rev.*, **132**, 1434–1447, doi:10.1175/1520-0493(2004)132<1434:ERIMFS>2.0.CO;2.

——, G. T. Bates, J. S. Whitaker, D. R. Murray, M. Fiorino, T. J. Galarneau, Y. Zhu, and W. Lapenta, 2013: NOAA's second-generation global medium-range ensemble reforecast dataset. *Bull. Amer. Meteor. Soc.*, **94**, 1553–1565, doi:10.1175/BAMS-D-12-00014.1.

——, ——, ——, ——, ——, and ——, 2016: A description of the 2nd-generation NOAA global ensemble reforecast data set. NOAA/Earth System Research Laboratory, 10 pp. [Available online at https://www.esrl.noaa.gov/psd/forecasts/reforecast2/README.GEFS_Reforecast2.pdf.]

Hersbach, H., 2000: Decomposition of the continuous ranked probability score for ensemble prediction systems. *Wea. Forecasting*, **15**, 559–570, doi:10.1175/1520-0434(2000)015<0559:DOTCRP>2.0.CO;2.

Jolliffe, I. T., and D. B. Stephenson, Eds., 2012: *Forecast Verification: A Practitioner's Guide in Atmospheric Science*. 2nd ed. John Wiley & Sons, 292 pp., doi:10.1002/9781119960003.

Kleiber, W., A. E. Raftery, and T. Gneiting, 2011: Geostatistical model averaging for locally calibrated probabilistic quantitative precipitation forecasting. *J. Amer. Stat. Assoc.*, **106**, 1291–1303, doi:10.1198/jasa.2011.ap10433.

Lawrence, B. A., M. I. Shebsovich, M. J. Glaudemans, and P. S. Tilles, 2003: Enhancing precipitation estimation capabilities at National Weather Service field offices using multi-sensor precipitation data mosaics. Preprints, *19th Conf. on Interactive Information Processing Systems for Meteorology, Oceanography, and Hydrology,* Long Beach, CA, Amer. Meteor. Soc., 15.1. [Available online at https://ams.confex.com/ams/annual2003/techprogram/paper_54867.htm.]

McLachlan, G. J., and T. Krishnan, 1997: *The EM Algorithm and Extensions*. 2nd ed. Wiley Series in Probability and Statistics, Vol. 389, Wiley, 400 pp.

Mendoza, P. A., B. Rajagopalan, M. P. Clark, K. Ikeda, and R. M. Rasmussen, 2015: Statistical postprocessing of high-resolution regional climate model output. *Mon. Wea. Rev.*, **143**, 1533–1553, doi:10.1175/MWR-D-14-00159.1.

Messner, J. W., G. J. Mayr, A. Zeileis, and D. S. Wilks, 2014a: Heteroscedastic extended logistic regression for postprocessing of ensemble guidance. *Mon. Wea. Rev.*, **142**, 448–456, doi:10.1175/MWR-D-13-00271.1.

——, ——, D. S. Wilks, and A. Zeileis, 2014b: Extending extended logistic regression: Extended versus separate versus ordered versus censored. *Mon. Wea. Rev.*, **142**, 3003–3014, doi:10.1175/MWR-D-13-00355.1.

Murphy, A. H., 1973: A new vector partition of the probability score. *J. Appl. Meteor.*, **12**, 595–600, doi:10.1175/1520-0450(1973)012<0595:ANVPOT>2.0.CO;2.

Politis, D. N., and J. P. Romano, 1994: The stationary bootstrap. *J. Amer. Stat. Assoc.*, **89**, 1303–1313, doi:10.1080/01621459.1994.10476870.

Polsky, C., J. Allard, N. Currit, R. Crane, and B. Yarnal, 2000: The Mid-Atlantic Region and its climate: Past, present, and future. *Climate Res.*, **14**, 161–173, doi:10.3354/cr014161.

Prat, O., and B. Nelson, 2015: Evaluation of precipitation estimates over CONUS derived from satellite, radar, and rain gauge data sets at daily to annual scales (2002–2012). *Hydrol. Earth Syst. Sci.*, **19**, 2037–2056, doi:10.5194/hess-19-2037-2015.

Raftery, A. E., T. Gneiting, F. Balabdaoui, and M. Polakowski, 2005: Using Bayesian model averaging to calibrate forecast ensembles. *Mon. Wea. Rev.*, **133**, 1155–1174, doi:10.1175/MWR2906.1.

Roulin, E., and S. Vannitsem, 2012: Postprocessing of ensemble precipitation predictions with extended logistic regression based on hindcasts. *Mon. Wea. Rev.*, **140**, 874–888, doi:10.1175/MWR-D-11-00062.1.

Roulston, M., and L. A. Smith, 2003: Combining dynamical and statistical ensembles. *Tellus*, **55A**, 16–30, doi:10.1034/j.1600-0870.2003.201378.x.

Schefzik, R., T. L. Thorarinsdottir, and T. Gneiting, 2013: Uncertainty quantification in complex simulation models using ensemble copula coupling. *Stat. Sci.*, **28**, 616–640, doi:10.1214/13-STS443.

Scheuerer, M., 2014: Probabilistic quantitative precipitation forecasting using ensemble model output statistics. *Quart. J. Roy. Meteor. Soc.*, **140**, 1086–1096, doi:10.1002/qj.2183.

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, doi:10.1175/2010MWR3285.1.

Sharma, S., and Coauthors, 2017: Eastern U.S. verification of ensemble precipitation forecasts. *Wea. Forecasting*, **32**, 117–139, doi:10.1175/WAF-D-16-0094.1.

Siddique, R., A. Mejia, J. Brown, S. Reed, and P. Ahnert, 2015: Verification of precipitation forecasts from two numerical weather prediction models in the Middle Atlantic Region of the USA: A precursory analysis to hydrologic forecasting. *J. Hydrol.*, **529**, 1390–1406, doi:10.1016/j.jhydrol.2015.08.042.

Sloughter, J. M. L., A. E. Raftery, T. Gneiting, and C. Fraley, 2007: Probabilistic quantitative precipitation forecasting using Bayesian model averaging. *Mon. Wea. Rev.*, **135**, 3209–3220, doi:10.1175/MWR3441.1.

Toth, Z., O. Talagrand, G. Candille, and Y. Zhu, 2003: Probability and ensemble forecasts. *Forecast Verification: A Practioner's Guide in Atmospheric Science*, I. T. Jolliffe and D. B. Stephenson, Eds., Wiley, 137–163.

Tracton, M. S., and E. Kalnay, 1993: Operational ensemble prediction at the National Meteorological Center: Practical aspects. *Wea. Forecasting*, **8**, 379–398, doi:10.1175/1520-0434(1993)008<0379:OEPATN>2.0.CO;2.

Wang, X., and C. H. Bishop, 2005: Improvement of ensemble reliability with a new dressing kernel. *Quart. J. Roy. Meteor. Soc.*, **131**, 965–986, doi:10.1256/qj.04.120.

Wei, M., Z. Toth, R. Wobus, and Y. Zhu, 2008: Initial perturbations based on the ensemble transform (ET) technique in the NCEP global operational forecast system. *Tellus*, **60A**, 62–79, doi:10.1111/j.1600-0870.2007.00273.x.

Wilks, D. S., 2006a: Comparison of ensemble-MOS methods in the Lorenz '96 setting. *Meteor. Appl.*, **13**, 243–256, doi:10.1017/S1350482706002192.

——, 2006b: *Statistical Methods in the Atmospheric Sciences.* Elsevier Academic Press, 627 pp.

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

——, 2010: Use of stochastic weather generators for precipitation downscaling. *Wiley Interdiscip. Rev.: Climate Change*, **1**, 898–907, doi:10.1002/wcc.85.

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

Wu, L., D.-J. Seo, J. Demargne, J. D. Brown, S. Cong, and J. Schaake, 2011: Generation of ensemble precipitation forecast from single-valued quantitative precipitation forecast for hydrologic ensemble prediction. *J. Hydrol.*, **399**, 281–298, doi:10.1016/j.jhydrol.2011.01.013.

Zhu, J., F. Kong, L. Ran, and H. Lei, 2015: Bayesian model averaging with stratified sampling for probabilistic quantitative precipitation forecasting in northern China during summer 2010. *Mon. Wea. Rev.*, **143**, 3628–3641, doi:10.1175/MWR-D-14-00301.1.