Statistical Postprocessing of High-Resolution Regional Climate Model Output

PABLO A. MENDOZA
Department of Civil, Environmental, and Architectural Engineering and Cooperative Institute for Research in Environmental Sciences, University of Colorado Boulder, and Research Applications Laboratory, National Center for Atmospheric Research, Boulder, Colorado

BALAJI RAJAGOPALAN
Department of Civil, Environmental, and Architectural Engineering and Cooperative Institute for Research in Environmental Sciences, University of Colorado Boulder, Boulder, Colorado

MARTYN P. CLARK, KYOKO IKEDA, AND ROY M. RASMUSSEN
Research Applications Laboratory, National Center for Atmospheric Research,* Boulder, Colorado

(Manuscript received 20 May 2014, in final form 31 October 2014)

ABSTRACT

Statistical postprocessing techniques have become essential tools for downscaling large-scale information to the point scale, and also for providing a better probabilistic characterization of hydrometeorological variables in simulation and forecasting applications at both short and long time scales. In this paper, the authors assess the utility of statistical postprocessing methods for generating probabilistic estimates of daily precipitation totals, using deterministic high-resolution outputs obtained with the Weather Research and Forecasting (WRF) Model. After a preliminary assessment of WRF simulations over a historical period, the performance of three post-processing techniques is compared: multinomial logistic regression (MnLR), quantile regression (QR), and Bayesian model averaging (BMA)—all of which use WRF outputs as potential predictors. Results demonstrate that the WRF Model has skill in reproducing observed precipitation events, especially during fall/winter. Furthermore, it is shown that the spatial distribution of skill obtained from statistical postprocessing is closely linked with the quality of WRF precipitation outputs. A detailed comparison of statistical precipitation postprocessing approaches reveals that, although the poorest performance was obtained using MnLR, there is not an overall best technique. While QR should be preferred if skill (i.e., small probability forecast errors) and reliability (i.e., match between forecast probabilities and observed frequencies) are target properties, BMA is recommended in cases when discrimination (i.e., prediction of occurrence versus nonoccurrence) and statistical consistency (i.e., equiprobability of the observations within their ensemble distributions) are desired. Based on the results obtained here, the authors believe that future research should explore frameworks reconciling hierarchical Bayesian models with the use of the extreme value theory for high precipitation events.

1. Introduction

Over the last few decades, there have been many advances in numerical weather prediction (NWP) models, resulting in substantial improvements in local weather and climate simulations (Friederichs and Hense 2007). However, it is well known that NWP models are far from being perfect, having common problems associated with differences between the real world and the interpretation provided in dynamical models, incomplete representations of physical processes of the atmosphere, and systematic biases and small spread in the case of ensemble systems (Hamill and Whitaker 2006; Grimit and Mass 2007). Therefore, statistical postprocessing techniques have emerged as an attractive alternative to provide a quantitative reinterpretation of raw NWP outputs,
Based on meteorological observations. These methods can be particularly useful when NWP models are deterministic, because they allow the generation of probabilistic information that can be valuable for decision making (e.g., Delle Monache et al. 2013).

Among these approaches, model output statistics (MOS) methods (Glahn and Lowry 1972; Carter et al. 1989) have been typically preferred. The basis of MOS techniques is the assumption of a relatively stationary climate, such that long historical time series of NWP simulations and observations can be used to develop forecast equations using dynamical model outputs as predictors. Some advantages of MOS approaches are (Wilks 2011) (i) unobserved quantities can be used as predictors (e.g., dynamical model-calculated wind vertical velocity, total column precipitable water), (ii) systematic errors exhibited in NWP models are accounted in the development of MOS equations, (iii) MOS products can be quite extensive (several sites, variables, forecast lead times, etc.), and (iv) equations can be seasonally and/or geographically stratified. The literature is quite rich in examples that make use of MOS techniques for postprocessing NWP outputs, including applications with multiple linear regression (e.g., Clark and Hay 2004), logistic regression (e.g., Hamill et al. 2004; Wilks 2009), and quantile regression (e.g., Bremnes 2004; Friederichs and Hense 2007), among other methods.

Although many authors have reported promising results for precipitation downscaling and postprocessing using MOS approaches, these techniques are unable to provide a full predictive probability density function (PDF), unless equations for a large number of thresholds/quantiles are adjusted. It is important to note that having a full PDF is particularly convenient, because several probabilistic quantities (e.g., prediction intervals, probabilities of exceeding a threshold) can be obtained in a consistent way (Scheuerer 2014). To address this issue, Raftery et al. (2005) introduced Bayesian model averaging (BMA) as a statistical postprocessing method that uses ensembles as input to generate calibrated predictive PDFs. Accordingly, the predictive PDF of the variable of interest is estimated as a weighted average of PDFs centered on the individual-bias corrected forecasts, where the weights are a measure of the predictive skill of the member forecasts over a training period (Sloughter et al. 2007; Fraley et al. 2010). Sloughter et al. (2007) extended this method to precipitation by modeling the predictive distribution for each ensemble member as a mixture of point mass at zero and a gamma distribution, and Fraley et al. (2010) extended the BMA approach to cases when one or more of the member forecasts are exchangeable and when there are missing ensemble members.

Several past studies have aimed to provide a comparison between different probabilistic postprocessing methods for meteorological variables. For instance, Wilks (2006) performed a comparison between a large number of statistical techniques, including direct model output, ensemble dressing, logistic regression, nonhomogeneous Gaussian regression, and BMA, among others. Wilks and Hamill (2007) compared the accuracy and skill of logistic regression with nonhomogeneous Gaussian regression and ensemble Gaussian dressing using a multidecadal dataset of temperature and precipitation, concluding that nonhomogeneous Gaussian regression generally performed better for medium-range temperature forecasts, and that logistic regression was better for daily temperature forecasts and for medium-range precipitation forecasts. Schmeits and Kok (2010) compared the performance of extended logistic regression (Wilks 2009), the Sloughter et al. (2007) version of BMA, and a modified BMA based on an alternative formula for computing the probability of precipitation (PoP) and additive bias correction. More recently, Bentzien and Friederichs (2012) analyzed the applicability of logistic regression, quantile regression, and parametric mixture models for generating probabilistic quantitative precipitation forecasts.

Nevertheless, none of the studies mentioned above has compared the performance of logistic regression-based techniques and quantile regression against BMA. Moreover, analyses of spatial biases or a robust comparison in terms of probabilistic properties such as skill (what is the magnitude of probability forecast errors?), reliability (how well do predicted probabilities correspond to observed frequencies?), discrimination (what is the ability of a forecasting system to distinguish the occurrence and nonoccurrence of an event?), and statistical consistency (are observations equally likely to be any ensemble member?) have been typically disregarded. Therefore, the main goal of this study is to apply and compare the utility of three statistical postprocessing methods (multinomial logistic regression, quantile regression, and Bayesian model averaging) for extracting probabilistic information from deterministic precipitation outputs (aggregated to daily amounts) obtained through high-resolution simulations with the Weather Research and Forecasting (WRF) Model. After a preliminary assessment of the WRF simulation outputs over the historical period October 2000–September 2008 (Rasmussen et al. 2014), we compare the postprocessing techniques in terms of several probabilistic properties. This paper is organized as follows: a description of the three statistical postprocessing methods compared in this study is presented in section 2, section 3 includes a full description of the experimental
design, results are provided in section 4, and the main conclusions are summarized in section 5.

2. Review of statistical postprocessing methods

MOS methods range from simple bias corrections to complex parametric and nonparametric statistical methods (Gangopadhyay et al. 2004). These techniques were developed at a time when NWP precipitation predictions were unreliable, and it was necessary to focus attention on atmospheric variables that were reliably simulated by the NWP model (e.g., zonal wind, total column precipitable water, convective available potential energy). However, modern NWP models, especially those run at regional scales and fine spatial resolution, have reliable precipitation simulations (e.g., Rasmussen et al. 2011, 2014), allowing the direct use of NWP precipitation output in postprocessing frameworks.

In this study, we consider three statistical methods to extract probabilistic information from deterministic NWP outputs, and characterize uncertainty in daily NWP precipitation: multinomial logistic regression (MnLR), quantile regression (QR), and Bayesian model averaging (BMA). The first two methods build on the development of regression equations: while MnLR is aimed to predict the probability of a variable to fall within different categories, QR directly provides conditional quantile estimates. On the other hand, BMA provides the full predictive PDF by weighting an ensemble of PDFs (which are centered in each ensemble member) based on their relative performance during a training period. We provide a detailed description of these postprocessing methods in sections 2a–c.

a. Multinomial logistic regression

A popular MOS method is logistic regression (Wilks 2011), which allows for the estimation of the probability of exceeding thresholds (e.g., Sokol 2003; Clark et al. 2004; Gangopadhyay et al. 2004; Hamill et al. 2004; Hamill and Whitaker 2006; Wilks 2009). By selecting different thresholds, the PDF of a specific variable can be constructed at each time step of the analysis period. Wilks (2009) noted that although good results can be obtained from logistic regression, there are some inconsistencies when fitting several equations for different thresholds (e.g., higher exceedance probabilities for larger values of the variable, in comparison with those obtained for lower values; differing estimates of uncertainties for different thresholds at a specific time step, etc.). Therefore, he introduced the extended logistic regression, which considers the threshold of interest as a new predictor (Wilks 2009; Roulin and Vannitsem 2012).

An alternative way to deal with the above problems is to use multinomial logistic regression, which predicts the probability of category membership of a dependent variable at each time step. Multinomial logistic regression may be considered a generalization of logistic regression analysis to nominal variables with more than two categories (Rodríguez 2007). Let Y be a random variable that may take one of several discrete values indexed as 1, 2, . . . , J. Defining the probability that the response falls in the jth category as

\[
\pi_j = P(Y = j)
\]

and assuming that the response categories are mutually exclusive and exhaustive, it holds that \(\sum_{j=1}^{J} \pi_j = 1\). It is important to note that there are only \(J - 1\) parameters in the model.

The multinomial logistic model can then be formulated for each time step as

\[
\pi_j = P(Y = j) = \frac{\exp(\eta_j)}{\sum_{m=1}^{J} \exp(\eta_m)}
\]

for \(j = 1, \ldots, J\), where

\[
\eta_j = \log\left(\frac{\pi_j}{\pi_j}\right) = \alpha_j + \mathbf{x}_j^T \mathbf{\beta}_j,
\]

where \(\alpha_j\) is a constant, \(\mathbf{\beta}_j\) is a vector of regression coefficients, with \(j = 1, 2, \ldots, J\), and \(\mathbf{x}\) is the vector of covariates (or predictors) for the time step of interest. The \(J - 1\) multinomial logistic regression equations compare each of categories 1, 2, . . . , \(J - 1\) with category J, whereas the logistic regression model is a contrast between successes and failures.

A key advantage of MnLR is that, as all categories are mutually exclusive by construction, possible numerical inconsistencies associated with the application of logistic regression for several thresholds (Wilks 2009) are alleviated. Hence, it is possible to get a consistent estimation of a cumulative distribution function (CDF) when several categories are defined, by converting the probabilities obtained for the different categories into the probability of exceeding a threshold. On the other hand, having a large number of categories might be a problem for obtaining reliable estimates of model coefficients if only short training datasets are available. For more details on MnLR, including model selection and fitting, we refer the readers to Hastie et al. (2009) and McCullagh and Nelder (1989).

b. Quantile regression

Another well-known MOS method is quantile regression (Koenker and Bassett 1978; Koenker and D’Orey 1987; Koenker and Hallock 2001), which has been used in
many past geophysical applications, including precipitation downscaling (e.g., Bremnes 2004; Tareghian and Rasmussen 2013), wind forecasting (e.g., Nielsen et al. 2006; Juban et al. 2007), and predictive hydrologic uncertainty (e.g., Weerts et al. 2011). Whereas standard regression methods attempt to predict the conditional mean of a response variable \( y \), QR seeks individual models for user-selected conditional quantiles. In other words, given a cumulative probability \( \theta \) and a vector of covariates \( x \), a linear function can be adjusted for \( q_\theta(y|x) \), which denotes the \( \theta \)th conditional quantile [i.e., the value satisfying the condition \( P(y \leq q_\theta(x) = \theta) \)]. Additional advantages of QR over traditional linear regression are as follows: no distribution assumptions, more robustness when handling extreme value points and outliers, and invariance of QR to monotonic transformations [i.e., if \( h \) is a monotonic transformation, \( q_\theta(h(y)|x) = h(q_\theta(y)|x)) \).

In this paper, the quantile regression models have the following form:

\[ q_\theta(y|x) = x^T\beta_\theta, \quad (4) \]

where \( 0 < \theta < 1 \), and \( \beta_\theta \) is a vector of parameters for the \( \theta \)th quantile regression equation. These parameters may also include an intercept, in which case the first element of \( x \) is a 1. Given a sample of \( N \) observations for the variable \( y \) (in this paper, daily precipitation total) and covariates \( x \), a set of parameters \( \beta_\theta \) for the \( \theta \)th quantile regression model can be estimated by solving the following minimization problem:

\[ \hat{\beta}_\theta = \arg\min_{\beta} \sum_{i=1}^{N} \rho_\theta(y_i - x_i^T\beta), \quad (5) \]

where \((x_i, y_i)\) are observations for covariates and the response variable at time steps \( t = 1, \ldots, N \), \( \hat{\beta}_\theta \) is the vector of parameters \( \beta \) that minimizes the cost function in Eq. (5), and \( \rho_\theta \)—also known as the "check" function (Nielsen et al. 2006)—is given as

\[ \rho_\theta(u) = \begin{cases} u\theta & u \geq 0 \\ u(\theta - 1) & u < 0 \end{cases}. \quad (6) \]

In Eq. (5), the absolute value of the difference between the observation \( y_i \) and the estimated \( \theta \)th conditional quantile is weighted by \( \theta \) if the observation is above the quantile plane (i.e., if \( y_i > x_i^T\beta \)), and weighted by \((1 - \theta)\), otherwise. If \( \theta = 0.5 \) (i.e., the median regression model), the vector of parameters \( \beta_{0.5} \) is obtained by minimizing the sum of absolute errors (Tareghian and Rasmussen 2013).

c. Bayesian model averaging

The principle of BMA (Raftery et al. 2005) is that given an ensemble forecast with \( M \) members, each ensemble member \( f_i \) \((i = 1, 2, \ldots, M)\) is associated with a conditional PDF \( h_i(y|f_i) \), which can be interpreted as the PDF of the meteorological variable \( y \) given \( f_i \). Thus, the BMA predictive model is

\[ p(y|f_1, \ldots, f_M) = \sum_{i=1}^{M} w_i h_i(y|f_i), \quad (7) \]

where the BMA weight \( w_i \) is the posterior probability of forecast \( i \) and is obtained based on its relative performance during the training period. Therefore, the weights \( w_i \)'s are nonnegative and add up to 1 [i.e., \( \sum_{i=1}^{M} w_i = 1 \); Raftery et al. (2005)].

In this study, we use the same model \( h_i(y|f_i) \) proposed by Sloughter et al. (2007), which has two components: one that specifies PoP as a function of the forecast \( f_i \), and another that specifies the PDF of precipitation amount given that it is not zero. The first component (PoP) is estimated using logistic regression as a function of the cube root of precipitation forecasts \( f_i \), adding a predictor variable \( \delta \), that is equal to 1 if \( f_i = 0 \) and equal to 0, otherwise. Hence, the logistic regression model is given by

\[ \logit P(y = 0|f_i) = \log \frac{P(y = 0|f_i)}{P(y > 0|f_i)} = a_{0i} + a_{1i}f_i^{1/3} + a_{2i}\delta_i. \quad (8) \]

Although Sloughter et al. (2007) started exploring gamma models to raw precipitation values, they found that employing the cube root of the precipitation amount yields a more appropriate model fit. Hence, the conditional PDF \( g_i(y|f_i) \) of the cube root precipitation amount \( y \) given that it is positive is a gamma distribution with PDF:

\[ g_i(y|f_i) = \frac{1}{\beta_i^\alpha_i \Gamma(\alpha_i)} y^{\alpha_i - 1} \exp(-y/\beta_i). \quad (9) \]

The shape parameter \( \alpha_i = \mu_i^2/\sigma_i^2 \) and the scale parameter \( \beta_i = \sigma_i/\mu_i \) of the gamma distribution depend on \( f_i \) following the relationships:

\[ \mu_i = b_{0i} + b_{1i}f_i^{1/3}, \quad (10) \]
\[ \alpha_i^2 = c_0 + c_1f_i, \quad (11) \]

which specify the mean and the variance of the distribution, respectively. The parameters \( b_{0i} \) and \( b_{1i} \) are determined separately for each ensemble member by applying linear regression between the cube root of the observed amount of precipitation (dependent variable) and the cube root of the forecasted precipitation amount (predictor
variable), considering only nonzero precipitation observations. On the other hand, the parameters $w_i (i = 1, 2, \ldots, M)$, $c_0$ and $c_1$ are estimated by maximum likelihood, following Sloughter et al. (2007).

Finally, the conditional PDF $h_i(y | f_i)$ combining the two components is

$$h_i(y | f_i) = P(y = 0 | f_i)I[y = 0] + P(y > 0 | f_i)g_i(y | f_i)I[y > 0],$$  

(12)

where $y$ is the cube root of the precipitation, $P(y = 0 | f_i)$ is specified by Eq. (8), and the function $I[s]$ is given by

$$I[s] = \begin{cases} 1 & s = \text{true} \\ 0 & s = \text{false} \end{cases}.$$  

(13)

3. Experimental design

a. Study region and data

The area of interest is the Colorado Headwaters region (Fig. 1), which is a major water resource for the southwestern United States. In this region, approximately 85% of the streamflow comes from snowmelt, and there is a high runoff sensitivity to changes in precipitation and temperature (Christensen and Lettenmaier 2007). This vulnerability, together with the importance of the Colorado River basin for water management, has motivated several climate change studies in this area, driven with different methodological choices and, therefore, resulting in a diverse set of conclusions (Milly et al. 2005; Christensen and Lettenmaier 2007; Hoerling and Eischeid 2007; Ray et al. 2008; Rasmussen et al. 2011; Vano et al. 2012). Although there are other major sources of uncertainty in climate change studies (e.g., choice of climate models, initial conditions, greenhouse emission scenarios), statistical techniques for precipitation downscaling and/or postprocessing have also been found to be important contributors to overall uncertainty (Wilby and Harris 2006; Chen et al. 2011; Gutmann et al. 2014; Vano et al. 2014).

The datasets used here come from the WRF historical simulation with a horizontal grid spacing of 4 km described in Rasmussen et al. (2014). The initial and 3-hourly lateral boundary conditions for the model run were taken from the North American Regional Reanalysis (NARR; Mesinger et al. 2006), whose spatial resolution is 0.3° (~32 km). The model physics options used in this simulation included the Noah land surface model.
(Noah-LSM) version 3.2 with upgraded snow physics (Chen and Dudhia 2001; Barlage et al. 2010), the Thompson mixed-phase cloud microphysics scheme (Thompson et al. 2008), the Yonsei University planetary boundary layer (Hong et al. 2006), and the Community Atmosphere Model’s (CAM) longwave and shortwave radiation schemes (Collins et al. 2006). These WRF simulations have been previously validated against SNOTEL sites, and precipitation spatial variability, timing, and intensities are well represented by the model (Ikeda et al. 2010; Prein et al. 2013).

Daily precipitation records at 93 SNOTEL sites located in the Colorado Headwaters region are used as verification data during eight water years (period 1 October 2000–30 September 2008). The resolution of the precipitation gauges is 2.54 mm (0.1 in.), but measurement error can be 10%–15% in snow due to undercatch related to windy conditions (Serreze et al. 1999; Rasmussen et al. 2012).

b. Evaluation of deterministic NWP outputs

We first evaluate the ability of the WRF Model to correctly simulate the occurrence of daily precipitation events of several magnitudes [i.e., the relative accuracy with respect to some set of standard control or reference climatology: Wilks (2011)]. For this evaluation, WRF precipitation data for each station were obtained by interpolating the WRF outputs from the four nearest grid points using the inverse distance weighting (IDW) interpolation method, following recent WRF assessments by Prein et al. (2013) and Rasmussen et al. (2014). To evaluate the ability of WRF for simulating precipitation events of different magnitudes, we use skill scores derived from the 2 × 2 contingency table, which provides information regarding the match between the model outputs and observations of dichotomous variables. Daily precipitation amounts defining different events (i.e., thresholds), are obtained as multiples of 2.54 mm (0.1 in.), which is the resolution of precipitation measurements. Therefore, contingency tables are constructed for each precipitation threshold of interest, and these results are used to compute the bias ratio and three skill scores: the Heidke skill score (HSS), the Peirce skill score (PSS), and the Clayton skill score (CSS). These three metrics were selected because they describe different aspects of model performance, by comparing a specific accuracy measure with that obtained by a set of reference model estimates. Specifically, HSS compares the proportion of correct model estimates (accuracy) with the correct proportion that would be achieved from random predictions that are statistically independent from the observations (reference). PSS also evaluates the proportion correct of model estimates, but with respect to that obtained by random predictions that are constrained to be unbiased [i.e., when proportions of wet (dry) model estimates and wet (dry) observations are constrained to be the same]. Finally, CSS compares the probability that the event occurred when it was simulated (accuracy) with the probability that the event occurred and it was not simulated (reference). The formulation of the three skill scores used in this paper for evaluating WRF precipitation outputs is detailed in appendix A.

c. Implementation of statistical postprocessing techniques

For each SNOTEL station, we selected the four nearest points from the 4-km grid spacing WRF domain, and then obtained potential predictors or ensemble forecasts depending on the postprocessing methodology to be used. For MnLR and QR, we obtained a set of predictors for each station by applying the IDW method to the four nearest grid points (see Table 1 for list of predictors from WRF). The BMA method as developed combines PDFs of multiple forecasts of the meteorological variable of interest, in this case, daily precipitation amounts. For this purpose, daily precipitation outputs from the WRF model at the four nearest grid points of a station are considered as four forecasts of the station precipitation, which can be combined through BMA. This technique has the clear advantage that observed extremes can be reproduced, in opposition to other simplistic approaches (e.g., arithmetic mean of the four nearest neighbors, IDW). The three methods are applied at each station separately in order to obtain the best postprocessing models.

Following the idea of Hamill et al. (2004), we performed several experiments with MnLR and QR at individual SNOTEL sites to get a suitable power transformation \( x' = x^\xi \) for WRF precipitation inputs. We found that \( \xi = 0.3 \) for MnLR, and \( \xi = 0.9 \) for QR were the best choices for model fitting. In the case of BMA, we followed Sloughter et al. (2007) with \( \xi = 1/3 \). We divided our methodology into three main steps: (i) estimation of the daily CDF of precipitation at each station, (ii) generation of daily precipitation ensembles at each location, and (iii) probabilistic verification using confidence limits.

<table>
<thead>
<tr>
<th>Variable</th>
<th>Description</th>
<th>Units</th>
</tr>
</thead>
<tbody>
<tr>
<td>PPT</td>
<td>Precipitation</td>
<td>mm</td>
</tr>
<tr>
<td>TEMP</td>
<td>2-m air temperature</td>
<td>K</td>
</tr>
<tr>
<td>Q2</td>
<td>Specific humidity</td>
<td>kg kg(^{-1})</td>
</tr>
<tr>
<td>PRESS</td>
<td>Air pressure</td>
<td>hPa</td>
</tr>
<tr>
<td>UWND</td>
<td>Zonal wind speed ( U ) at 10 m</td>
<td>m s(^{-1})</td>
</tr>
<tr>
<td>VWND</td>
<td>Meridional wind speed ( V ) at 10 m</td>
<td>m s(^{-1})</td>
</tr>
</tbody>
</table>
1) **ESTIMATION OF THE CDF**

As each postprocessing method has different characteristics, estimation of the daily CDF at each station [step (i) above] varies for each technique:

1) **MnLR**: At each station, we classify daily precipitation observations in $J = 41$ categories (defined from $J - 1 = 40$ precipitation thresholds), from 1 for null values to $J$ if the maximum threshold is exceeded. Precipitation thresholds are obtained as multiples of 2.54 mm (0.1 in.), which is the minimum resolution of precipitation observations. MnLR models are adjusted for several combinations of predictors, and the “best” model is selected by using Akaike’s Information Criterion (AIC; Akaike 1974). Finally, the probability of exceeding each threshold $P_{\text{thresh}}$ (i.e., $P(P \geq P_{\text{thresh}})$) is estimated by adding the probabilities of all the categories associated with precipitation amounts larger than the value of interest.

2) **QR**: We select $109 \theta$ values ranging from 0 to 1 (100 of them ranging from 0 to 0.99 at intervals of 0.01, and a finer grid with size 0.001 for the interval 0.991–0.999 for more precise estimates of extremes), and then we fit quantile regression equations for daily precipitation totals at each station, for each $\theta$ value and each possible combination of predictors (i.e., having defined a probability of nonexceeding, we identify the associated precipitation threshold). Hence, the final quantile regression model (and final set of predictors) for each $\theta$ value will be the one that minimizes the cost function in Eq. (5). Finally, the CDF curves for the thresholds of interest (i.e., those defined for MnLR) are obtained by linear interpolation from the CDF constructed through quantile regression.

3) **BMA**: For each station location, the ensemble of daily precipitation forecasts is given by the four nearest WRF points. Prior information (i.e., weights) for the ensemble members is provided based on the inverse of the distance between each WRF grid point and the station of interest. Instead of using a sliding window as a training period (e.g., Sloughter et al. 2007; Schmeits and Kok 2010), we use the same calibration/verification periods as MnLR and QR (see details below). The final CDF at each time step is constructed by computing the probability of exceedance for the same precipitation thresholds defined for the other methods.

With the aim to fairly compare the different post-processing approaches via cross-validation, we use the same training/verification periods for this step. Hence, each water year in the time series is held out of the sample used to fit the models, and the equations are used to predict the data for the withheld year. We repeat this process for all water years in the record, following the procedure in Clark et al. (2004). Additionally, model equations/weights are adjusted separately for fall/winter (October–March) and spring/summer (April–September) seasons, mainly because winter precipitation is dominated by synoptically forced stratiform-type storms, while summer precipitation is convective and occurs at a smaller scale over the Colorado Headwaters. All models are fitted using the packages VGAM version 0.8-4.1 for MnLR (Yee 2010), quantreg version 4.76 for QR (Koenker 2012), and ensemble BMA version 5.0.3 for BMA (Fraley et al. 2013), which are implemented in the statistical software R (R Development Core Team 2011).

2) **GENERATING PRECIPITATION ENSEMBLES**

In hydrologic modeling applications, uncertainties in meteorological forcings, especially precipitation, are a major source of hydrologic model errors (e.g., Carpenter and Georgakakos 2004; Nasonova et al. 2011), translating into uncertainties in modeled states (e.g., soil moisture, snow water equivalent) and fluxes (e.g., streamflow, evapotranspiration). A typical approach for explicitly considering forcing uncertainties and quantifying hydrologic model error is to run hydrologic model simulations with an ensemble of inputs (e.g., Clark et al. 2006; De Lannoy et al. 2006). In this context, probabilistic precipitation estimates in the form of ensembles are particularly useful. Therefore, we use the results from the previous step to generate ensembles of daily precipitation totals at each precipitation station. For each time step, $M = 100$ random numbers are generated from a uniform distribution $U[0, 1]$, and the corresponding precipitation amounts are estimated from each of the three CDFs obtained in step (i).

3) **PROBABILISTIC VERIFICATION WITH CONFIDENCE INTERVALS**

Probabilistic forecasts of hydrometeorological variables are commonly used to quantify uncertainty and to supplement the information provided by point-value predictions (e.g., Hamill 1997; Clark et al. 2004; Clark and Slater 2006; Laio and Tamea 2007; Stensrud and Yussouf 2007; Pappenberger et al. 2009; Renner et al. 2009; Mendoza et al. 2012, 2014). In many cases, forecast verification is limited to a simple quantitative comparison between the observed and the ensemble time series plots, using a suite of scalar measures for the evaluation of model performance in reproducing some events. However, these comparisons do not provide a quantitative assessment of skill and uncertainty of internal calculations (Mascaro et al. 2008).

In this study, statistical postprocessing methods are compared using four different verification criteria: the Brier (skill) score (Brier 1950), reliability diagrams,
discrimination diagrams (Wilks 2011), and rank histograms (Hamil 2001). To create confidence intervals for the verification statistics, we use a Monte Carlo resampling approach based on bootstrapping with replacement (Wilks 2011). That is, \( N \) pairs of estimated probabilities and observed occurrences were resampled from the original joint distribution (\( N \) is the total number of events for which probabilistic estimates are available). This process is repeated 1000 times, and all statistics are therefore computed for each realization and ranked in order to obtain confidence limits. Probabilistic verification is performed separately over the two periods defined above (fall/winter and spring/summer).

4. Results

a. Assessment of historical WRF outputs

Figure 2 presents three skill scores of raw WRF precipitation outputs for fall/winter and spring/summer along with confidence intervals, for various daily precipitation thresholds. The bias ratio is also included in Fig. 2 to examine the (mis)match between “wet” model events and “wet” observations. We disregard the use of a zero precipitation threshold in this study since it has been observed that WRF outputs may contain very small daily precipitation amounts when no precipitation has been observed or precipitation was too light to be measured by the SNOTEL precipitation gauges. Moreover, there is evidence that the inclusion of 0 mm (e.g., Schmeits and Kok 2010) or very light (e.g., Ruiz et al. 2009; Pappenberger et al. 2009) precipitation thresholds many times lead to poor skill in postprocessing results, which likely has more to do with instrument limitations rather than capabilities of different postprocessing methods. For these reasons, we decide to define precipitation occurrence at the gauge resolution following Colle et al. (1999, 2000, 2003) and Jones et al. (2007). The accuracy measures displayed in Fig. 2 are the Heidke skill score (HSS), the Peirce skill score (PSS), and the Clayton skill score (CSS), which assign a maximum value of 1 for perfect simulations, positive values if forecast accuracy is better than that provided by a set of reference predictions (which is different for each score, as pointed in section 3b and appendix A), and negative values otherwise. Results show higher skill in fall/winter compared to spring/summer, especially for simulating heavier precipitation events. Moreover, these results are consistent with those shown in Ikeda et al. (2010), Rasmussen et al. (2011, 2014), and Prein et al. (2013), demonstrating the good skill of high-resolution dynamical downscaling for simulating events of several magnitudes in the Colorado Headwaters region, especially during the fall/winter period. As expected, skill values tend to decrease for higher precipitation thresholds, particularly during spring/summer. One can also note considerable differences between skill scores among precipitation events during fall/winter, which is due to differences in the reference accuracy used to formulate the skill measures included in this analysis.

Based on the bias ratio (BR) results displayed in Fig. 2, it was obtained that, over fall/winter, WRF tends to underforecast (BR < 1) light precipitation events \( [P_{\text{thresh}} = 2.54 \text{ mm (0.1 in.)}] \), and tends to overforecast (BR > 1) all events with \( P_{\text{thresh}} \geq 7.62 \text{ mm (0.3 in.)} \).
During spring/summer, WRF is prone to overforecast all precipitation events with $P_{\text{thresh}} \geq 5.08 \text{ mm (0.2 in.)}$. To examine the spatial distribution of skill associated with dynamically downscaled precipitation in the area of interest, we generated spatial maps with HSS, PSS, and CSS values obtained at all stations for different precipitation thresholds. As a similar spatial pattern was obtained with all thresholds, we only report skill maps for $P_{\text{thresh}} = 7.62 \text{ mm (0.3 in.); Fig. 3}$. Scores up to 0.8 at some locations are evident during fall/winter, indicating remarkably skillful WRF simulations over this season. Consistent with Fig. 2, skill is lower during spring/summer. Additionally, similar spatial patterns are obtained for all scores. During fall/winter, higher skill is obtained at sites in the southwestern area of the Colorado Headwaters region, while during spring/summer slightly better skill is obtained at high-altitude sites in the northeastern region. Analogous skill spatial patterns for winter and summer seasons were reported by Prein et al. (2013), who assessed the ability of WRF for simulating heavy precipitation events.

b. Comparison of statistical postprocessing techniques

1) PROBABILISTIC SKILL

It should be noted that the skill scores presented in section 4a are formulated for nonprobabilistic variables (recall that WRF outputs are deterministic) describing dichotomous events. Therefore, a different

![Image of spatial distribution of skill associated with raw WRF precipitation outputs over the Colorado Headwaters region for a daily precipitation threshold of 7.62 mm (0.3 in.). The verification metrics displayed (from left to right) are Heidke skill score (HSS), Peirce skill score (PSS), and Clayton skill score (CSS). All scores were calculated at each SNOTEL site (top) for fall/winter days (October–March) over the period October 2000–September 2008 and (bottom) for spring/summer days (April–September) during the same period.](image-url)
accuracy measure should be selected if the variable of interest is probabilistic (Jolliffe and Stephenson 2003; Cloke and Pappenberger 2008; Pappenberger et al. 2011). In this paper, we choose the Brier skill score (Brier 1950) to assess the relative skill of the probabilistic precipitation estimates over that of climatology, in terms of predicting whether or not a precipitation event with a particular magnitude occurred. BSS values range from 1 (perfect skill) to negative infinity. Positive values indicate that the model is more skillful than the mean observed climatology, negative values reflect the opposite, and a value of zero indicates no skill relative to reference climatology (see appendix B for details).

Brier skill scores with 95% confidence intervals are presented in Fig. 4 for the three postprocessing methods and various daily precipitation thresholds. Overall, all approaches are better than taking the mean observed climatology as the estimated probability (i.e., BSS > 0). Skill scores are higher during fall/winter with all methods when compared to spring/summer, and also tend to decrease from a threshold \( P_{\text{thresh}} = 7.62 \) mm (0.3 in.), which is consistent with the assessment of raw WRF outputs previously presented (Fig. 2). During fall/winter, higher skill scores are obtained with QR, followed by MnLR and BMA. However, during spring/summer the most skillful method seems to be threshold dependent: while MnLR is in general the most skillful for low precipitation events \( [P_{\text{thresh}} \leq 10.16 \) mm (0.4 in.)], QR and BMA become better as precipitation threshold increases. It is worth noting that the uncertainty introduced by sampling variability (i.e., size of confidence intervals) in MnLR and QR is much larger than that in BMA, especially for very high precipitation events. This suggests that the impact of sampling uncertainty on skill results depends on the type of postprocessing technique. In this application, skill results obtained with threshold-based (e.g., MnLR) or quantile-based (e.g., QR) methods are more sensitive to sampling variability than a PDF-centered approach. Schmeits and Kok (2010) also found that bootstrap confidence intervals for the BMA formulation by Sloughter et al. (2007) were smaller than other postprocessing techniques.

BSS maps obtained with all postprocessing methods for a threshold \( P_{\text{thresh}} = 7.62 \) mm (0.3 in.) were produced in order to look for possible regionalized skill patterns in the area of interest (Fig. 5). Similar to skill results from raw WRF outputs (Fig. 3), higher BSS values were obtained at sites located in the southwestern area during the fall/winter period and slightly lower BSS were observed in the northeastern area. However, during spring/summer skill is more spatially uniform across the study region. Since the spatial distributions of skill scores are very similar with all postprocessing methods regardless of the season, we can infer that they are directly correlated with the quality of WRF predictors.

The overall better skill obtained by postprocessing methods over fall/winter is better than that for spring/summer (Figs. 4 and 5) for the combination of two reasons. First, better skill should be expected during fall/winter because, in this region, the number of precipitation events observed is larger over this period, and therefore statistical (data driven) postprocessing techniques will be “better trained” for simulating them in comparison with spring/summer
On the other hand, the higher skill of raw WRF precipitation over fall/winter (Figs. 2 and 3) should also benefit postprocessing skill results over this season, since all these techniques make use of them as inputs.

2) RELIABILITY

The reliability diagram is a graphic device that describes how well the predicted probabilities of a binary predictand correspond to their observed frequencies. Therefore, the reliability diagram displays the conditional probability that an event occurred (also known as “observed relative frequency”) as a function of discrete values representative of forecast probability categories (e.g., 0%–10%, 10%–20%, 20%–30%, etc.). A characteristic signature of perfectly reliable forecasts is given when observed relative frequencies equal forecast probabilities, resulting in 1:1 diagonal line, except for deviations related to sample variability. Typically, the frequency of forecast probabilities in each category is displayed in a histogram (“sharpness diagram”) together with the reliability diagram. A “sharp forecast” will be represented by high frequencies for the forecast probabilities 0 and 1, and low frequencies in between. Recall that a sharper forecast can distinguish better between an event and a nonevent (Renner et al. 2009).
Figure 6 presents reliability diagrams for three precipitation thresholds (rows) and the three post-processing methods (columns) during the fall/winter period. When predicting the occurrence of precipitation [i.e., larger than $P_{\text{thresh}} = 2.54 \text{ mm (0.1 in.)}$], QR is more reliable, followed by MnLR and BMA, which have the tendency to underpredict (i.e., relative observed frequencies are larger than the associated forecast probabilities, or “dry bias”) and overpredict (i.e., relative observed frequencies smaller than the associated forecast probabilities, or “wet bias”) precipitation events, respectively. On the other hand, MnLR has a better sharpness (higher frequency of forecasts probabilities close to zero) when compared to QR and BMA. For a threshold $P_{\text{thresh}} = 7.62 \text{ mm (0.3 in.)}$, results reflect a good sharpness in all methods, a dry bias with MnLR, good reliability with QR, and a slight underconfidence of BMA (i.e., forecasts are less extreme than observed frequencies, given by an “S” shape). When heavier precipitation events [$P_{\text{thresh}} = 15.24 \text{ mm (0.6 in.)}$] are analyzed, MnLR and BMA are characterized by wet and dry bias, respectively, while QR is the most reliable postprocessing method. It is interesting to note that the poor reliability of BMA for this precipitation threshold is consistent with the results reported by Schmeits and Kok (2010).

The reliability diagrams for spring/summer (Fig. 7) illustrate that QR is more reliable compared to the other two methods for the three thresholds analyzed. It is also interesting to note the mismatch between estimated probabilities and observed frequencies for the
highest forecast probabilities, especially for QR, which can be attributed to insufficient sampling size. When predicting precipitation occurrence \([P_{\text{thresh}} = 2.54 \text{ mm (0.1 in.)}]\), MnLR and BMA present dry and wet biases, respectively. However, an opposite behavior is obtained for larger precipitation thresholds: MnLR and BMA start showing wet and dry biases, respectively, for the highest forecast probabilities.

Missing points associated with categories for high estimated probabilities in Figs. 6 and 7 are related with the frequency at which some precipitation events occur in each season, and also to the reliability of each postprocessing method for specific thresholds. For instance, the top-left panel in Fig. 7 \([P_{\text{thresh}} = 2.54 \text{ mm (0.1 in.)}]\), MnLR] has points for all categories (i.e., 10 points) because MnLR generates probabilities for all magnitudes during spring/summer, whereas QR and BMA only produce estimated probabilities lower than 0.9 (i.e., only 9 points in reliability diagrams). Note that, for the same threshold during fall/winter (top panels in Fig. 6), all the methods can predict high probabilities of precipitation occurrence (i.e., 10 points in all reliability diagrams) because such events are more frequent during that season, and therefore all techniques are better calibrated for reproducing them. On the other hand, lack of points for high probability categories is expected to be obtained if the precipitation event analyzed is less frequent (e.g., bottom panels in Fig. 7 for \([P_{\text{thresh}} = 15.24 \text{ mm (0.6 in.)}])\), which is consistent with the results obtained by Clark and Slater (2006). Note that MnLR is an exception because it is clearly unreliable as the threshold increases, tending to overestimate the occurrence of higher precipitation events (i.e., when a high probability of an event is forecasted, the actual occurrence of that event is less common).

**Fig. 7.** As in Fig. 6, but for spring/summer days (April–September) during the period October 2000–September 2008.
3) DISCRIMINATION

The discrimination diagram (Wilks 2011) shows the ability of a forecast system to clearly distinguish situations leading to the occurrence of an event of interest from those leading to the nonoccurrence of the event. Thus, given a dichotomous predictand (i.e., \( J = 2 \)), the discrimination diagram consists on the superimposed plots of the two conditional distributions \( p(y_i | o_j), j = 0, 1 \) as functions of forecast probabilities \( y_i \). A perfect discrimination between the two events will require no overlap between their likelihoods (PDFs). Additionally, a good discrimination will depend on the separation of means of conditional distributions, and on the variance within conditional distributions. In this paper, we use the discrimination distance (Wilks 2011) to quantify the separation of the two likelihood distributions, which is given by the absolute difference between their means:

\[
    d = |\mu_{y_i|o_0} - \mu_{y_i|o_1}|,
\]

where \( \mu_{y_i|o_0} \) (\( \mu_{y_i|o_0} \)) is the mean of forecast probabilities for those cases where the event was (not) observed.

Figure 8 displays discrimination diagrams for three precipitation thresholds (rows) and the three post-processing approaches (columns) during fall/winter.
The metrics displayed in each panel are discrimination distance ($d$), overlapping area ($A$), and standard deviations ($\sigma_0$ and $\sigma_1$) from the two conditional likelihoods. Recall that a good discrimination will be given by sharp (i.e., small $\sigma_0$ and $\sigma_1$), well separated (large $d$), and no overlapping (small $A$) likelihoods. Confidence limits are omitted in all discrimination diagrams since there were only minor differences in the resampled PDFs. Results show that none of the three methods is the best when looking at all performance measures simultaneously. For example, for $P_{\text{thresh}} = 2.54 \text{ mm (0.1 in.)}$, BMA has the largest discrimination distance, QR shows the sharpest PDF for event occurrence (i.e., $o = 1$), and MnLR shows the sharpest PDF for nonoccurrence and the smallest overlapping area. However, one can see that, unlike QR and BMA, the shapes of the two likelihoods obtained from MnLR are quite similar, and the PDF for cases when $o = 1$ is always skewed to the left, reflecting the failure to assign high probabilities to precipitation events of different magnitudes. Although MnLR provides the smallest values for $A$, there are not substantial differences in this quantity among methods, and BMA shows in general better defined (i.e., sharp) PDFs for event occurrence ($o = 1$). As precipitation threshold increases, the likelihoods for nonoccurrence ($o = 0$) from all methods tend to be sharper and also—with the exception of BMA—mimic the shape of the PDF for occurrence ($o = 1$).

Discrimination diagrams for the spring/summer period are presented in Fig. 9. Again, the performance of MnLR is poor when contrasted to QR and BMA (i.e., although values of $d$ and $A$ are better, the conditional likelihoods are hardly distinguishable). BMA has a larger discrimination distance $d$ when compared to QR for nonzero precipitation events. Nevertheless, QR performs better than BMA in terms of $d$ for larger

---

**Fig. 9.** As in Fig. 8, but for spring/summer days (April–September) during the period October 2000–September 2008.
thresholds. Also, BMA has better defined (i.e., sharper) PDFs for event occurrence ($\sigma = 1$) when compared to the rest of the methods. A final comparison of the results in Figs. 8 and 9 reveals that, although sharper likelihoods (i.e., smaller $\sigma_0$ and $\sigma_1$) and slightly smaller overlapping areas are obtained over spring/summer, the discrimination distance is larger for all methods/thresholds over fall/winter, and a clearer separation of PDFs can be observed during this period.

4) STATISTICAL CONSISTENCY

Rank histograms are used to diagnose whether the observations are equally distributed throughout the ensemble. A rank histogram can be generated by collecting the rank (i.e., “position”) of the observation with respect to the ensemble members over a number of cases. If the ensemble is statistically consistent, the “truth” (i.e., observation) should fall into each bin with equal probability (flat rank histogram). Positive or negative biases in the ensemble forecast will produce overpopulation of the lowest or highest ranks (Hamill 2001; Mascaro et al. 2008); an excess of dispersion (overdispersion) implies overpopulation of the middle ranks, and a lack of variability in the ensemble will show up as a U-shaped, or concave, rank population. It should be noted that a flat rank histogram is a necessary but not sufficient condition for statistical consistency (Hamill 2001).

In addition to the rank histogram, we also include a discrepancy index (DI; Delle Monache et al. 2006) to quantify the departure of the histogram from uniformity:

$$DI = \frac{100}{M+1} \sum_{i=1}^{M+1} \left| \frac{\text{count}_i}{N} - \frac{1}{M+1} \right|$$  \hspace{1cm} (15)

In the above equation, $M$ is the number of ensemble members (so $M + 1$ is the number of bins in the rank histogram), count$_i$ is the number of times that the observation falls into the $i$th bin, and $N = \sum_{i=1}^{M+1} \text{count}_i$ is the sample size. Lower DI means that the condition of statistical consistency is better achieved by the ensemble.

Figure 10 displays the rank histograms for fall/winter and spring/summer daily precipitation over the period October 2000–September 2008. The shape of fall/winter rank histograms (top row) and DI values demonstrate that a better spread is generated using precipitation CDFs computed with BMA. During the

![Fig. 10. Rank histograms for (top) fall/winter and (bottom) spring/summer during the period October 2000–September 2008. Columns are associated with different statistical postprocessing methods. DI denotes the discrepancy index.](image)
same period, MnLR presents a good spread except for the 15% upper bins (i.e., negative bias), while the U-shaped rank histogram associated with QR reflects lack of variability in the ensemble, with DI = 51.32%. Similar rank histograms are obtained during spring/summer days (Fig. 10, bottom row), although DI improves for MnLR (from 17.53% to 11.33%) and BMA (from 6.46% to 4.54%) compared to fall/winter, while it degrades for QR (from 51.32% to 63.97%) over that period.

5. Summary and conclusions

In this paper, we applied statistical precipitation postprocessing methods for generating probabilistic information from high-resolution deterministic outputs obtained with the WRF Model. First, an evaluation of historical WRF runs (period October 2000–September 2008) was performed in terms of their ability to reproduce events of several magnitudes. Second, we applied and compared the performance of three statistical postprocessing techniques over the same period: multinomial logistic regression (MnLR), quantile regression (QR), and Bayesian model averaging (BMA). The application of all postprocessing methods involved (i) estimation of the cumulative distribution function (CDF) of precipitation at each station and each day for the period October 2000–September 2008, (ii) generation of daily precipitation ensembles at each location, and (iii) probabilistic verification using confidence intervals through bootstrapping with replacement.

The evaluation of raw WRF precipitation outputs via three skill scores demonstrated the accuracy of high-resolution dynamically downscaled precipitation for several thresholds, with a better performance over fall/winter. Moreover, it was found that WRF tends to produce more precipitation events of most magnitudes during fall/winter and spring/summer compared with the observations. The spatial distribution of skill measures over the Colorado Headwaters region was also analyzed, finding that during fall/winter a better skill is obtained at sites located in the southwestern area, while the skillful area covers the northeast over spring/summer. The posterior assessment of statistical postprocessing methods showed a similar spatial distribution of skill, proving the dependence of postprocessing results on the quality of predictors.

The comparison of statistical postprocessing techniques indicates that, although the poorest performance was obtained using MnLR, there is not an overall best approach. In other words, the “best” method will depend on the target probabilistic property. For the particular case of postprocessing deterministic daily precipitation, if the goal is to get small errors in probability forecasts (skill), or a good match between forecast probabilities and relative observed frequencies (reliability), QR should be preferred. On the other hand, if the primary concern is to correctly discriminate the occurrence and nonoccurrence of precipitation events (discrimination), or to obtain ensemble precipitation estimates whose spread is comparable with that given by observations (statistical consistency)—and therefore precipitation ensembles more appropriate for hydrologic uncertainty quantification—the technique to use should be BMA. Note that these results also illustrate that there is not necessarily a correspondence of probabilistic verification results (i.e., Brier skill scores, reliability diagrams, etc.) across postprocessing techniques, since the verification methods included in this paper evaluate different aspects of probabilistic information content in the postprocessed WRF outputs.

A general degradation of probabilistic properties with increasing precipitation thresholds was observed with all methods, which is consistent with results reported in other studies (e.g., Clark and Slater 2006; Stensrud and Yussouf 2007; Sloughter et al. 2007; Schmeits and Kok 2010). In particular, poor reliability was obtained using BMA for high thresholds, which is in agreement with the later study. Based on these results, improvements in extreme event estimates may be possible using a hybrid approach that makes use of a different probability distribution for precipitation amounts above a specific threshold (e.g., Furrer and Katz 2008; Bentzien and Friederichs 2012). Even more, we hypothesize that, given the potential of the Bayesian approach explored here, a framework connecting (i) a formal Bayesian hierarchical model including time–space-varying parameters and additional circulation variables (e.g., Bannerjee et al. 2003), and (ii) elements from the extreme value theory for high precipitation events, could be extremely powerful for more consistent probabilistic precipitation estimates.

The techniques evaluated in this paper have potential for both weather forecasting and climate downscaling applications. In real-time weather forecasting applications, the methods tested here may be incorporated by using sliding windows, whose data (WRF forecasts and observations) can be used to estimate parameters (MnLR and QR) or weights (BMA) for subsequent forecasting periods (e.g., Raftery et al. 2005; Sloughter et al. 2007; Fraley et al. 2010; Schmeits and Kok 2010; Sweeney et al. 2013). Applications involving climate change scenarios are certainly more challenging, since the assumptions that the statistical relationships between predictors and predictand remain constant (i.e.,
stationarity), or that predictors carry the climate change “signal” and natural climatic trends may not be valid (Benestad 2010; Vannitsem 2011), though careful predictor selection, together with an evaluation of coefficients/weights over different hydroclimatic conditions, may help to build robust postprocessing frameworks under nonstationary climate conditions (Schmith 2008; Gutiérrez et al. 2013).

Acknowledgments. The authors thank the constructive comments from Naoki Mizukami and three anonymous reviewers, and James McCreight for his help with the implementation of Bayesian model averaging. The first author received support from the CIRES Graduate Fellowship Award, a research grant from the U.S. Bureau of Reclamation (USBR), and a contract from the U.S. Army Corps of Engineers (USACE).

APPENDIX A

Skill Scores from Contingency Tables

Given a precipitation threshold of interest, contingency tables (Table A1) can be constructed in order to assess the correspondence between nonprobabilistic forecast values and the discrete observable predictand values to which they pertain (Wilks 2011). Based on the information provided by 2 × 2 contingency tables, the bias ratio and some relative accuracy measures (skill scores) can be formulated. The bias ratio (BR) is simply the ratio of the number of wet forecasts to the number of wet observations:

\[ BR = \frac{a + b}{a + c}. \]  

(A1)

Unbiased forecasts exhibit BR = 1. If BR > 1, the event was forecasted more often than observed (“overforecasting”), while BR < 1 indicates the opposite (“underforecasting”).

The Heidke skill score (HSS) is based on the correct proportion referenced with the correct fraction that would be achieved by random forecasts that are statistically independent of the observations. The HSS is given as

\[ HSS = \frac{2(ad - bc)}{(a + c)(c + d) + (a + b)(b + d)}. \]  

(A2)

where values of a, b, c, and d come from the contingency table. For this metric, a perfect forecast system receives skill values equal to 1, forecasts equivalent to the reference receive 0, and negative scores are associated with forecasts worse than the reference climatology.

The Peirce skill score (PSS) is similar to the HSS, except that the reference hit rate in the denominator is for random and unbiased forecasts:

\[ PSS = \frac{ad - bc}{(a + c)(b + d)}. \]  

(A3)

For this metric, perfect forecasts receive a score of 1, random and constant forecasts receive a score of 0, and forecasts inferior to the random forecast receive negative scores.

The Clayton skill score (CSS) indicates positive skill to the extent that the event is observed more frequently when forecasted than when not forecasted. Perfect forecast yield CSS = 1, and random forecasts yield CSS = 0. The CSS is

\[ CSS = \frac{ad - bc}{(a + b)(c + d)}. \]  

(A4)

APPENDIX B

Brier Score and Brier Skill Score

The Brier score (Brier 1950) represents the magnitude of the probability forecast errors for dichotomous predictands. Let \( p_t \) be the estimated probability that a particular threshold is exceeded at time step \( t \), where \( t = 1, \ldots, N \). Observations can be treated in a similar way, by letting \( o_t = 1 \) if the event occurred (i.e., the observed value is larger than the threshold), and \( o_t = 0 \), otherwise. Then, the Brier score can be obtained with the following formula:

\[ BS_f = \frac{1}{N} \sum_{t=1}^{N} (p_t - o_t)^2. \]  

(B1)

Brier score values range from 0 to 1, with a perfect score of \( BS_f = 0 \). If one wants to compare with other stations or datasets, it is recommended to use the Brier skill score (BSS):

\[ BSS = 1 - \frac{BS_f}{BS_c}, \]  

(B2)

where \( BS_f \) denotes the Brier score of the forecast and \( BS_c \) is the Brier score obtained from climatological
estimates. BSS values equal to 1 indicate perfect estimates, while negative scores indicate that the estimate is worse than what could be obtained from climatology.

REFERENCES


Ray, A., J. Barsugli, K. Averty, K. Wolter, M. Hoerling, N. Doesken, B. Udall, and R. Webb, 2008: Climate change in Colorado: A synthesis to support water resources management and adaptation. Intermountain West Climate Summary,


