How Suitable is Quantile Mapping For Postprocessing GCM Precipitation Forecasts?

TONGTIEGANG ZHAO, JAMES C. BENNETT, AND Q. J. WANG

CSIRO Land and Water, Clayton, Victoria, Australia

ANDREW SCHEPEN

CSIRO Land and Water, Dutton Park, Queensland, Australia

ANDREW W. WOOD

National Center for Atmospheric Research, Boulder, Colorado

DAVID E. ROBERTSON

CSIRO Land and Water, Clayton, Victoria, Australia

MARIA-HELENA RAMOS

Irstea, Hydrosystems and Bioprocesses Research Unit, Antony, France

(Manuscript received 31 August 2016, in final form 11 January 2017)

ABSTRACT

GCMs are used by many national weather services to produce seasonal outlooks of atmospheric and oceanic conditions and fluxes. Postprocessing is often a necessary step before GCM forecasts can be applied in practice. Quantile mapping (QM) is rapidly becoming the method of choice by operational agencies to postprocess raw GCM outputs. The authors investigate whether QM is appropriate for this task. Ensemble forecast postprocessing methods should aim to 1) correct bias, 2) ensure forecasts are reliable in ensemble spread, and 3) guarantee forecasts are at least as skillful as climatology, a property called ''coherence.'' This study evaluates the effectiveness of QM in achieving these aims by applying it to precipitation forecasts from the POAMA model. It is shown that while QM is highly effective in correcting bias, it cannot ensure reliability in forecast ensemble spread or guarantee coherence. This is because QM ignores the correlation between raw ensemble forecasts and observations. When raw forecasts are not significantly positively correlated with observations, QM tends to produce negatively skillful forecasts. Even when there is significant positive correlation, QM cannot ensure reliability and coherence for postprocessed forecasts. Therefore, QM is not a fully satisfactory method for postprocessing forecasts where the issues of bias, reliability, and coherence pre-exist. Alternative postprocessing methods based on ensemble model output statistics (EMOS) are available that achieve not only unbiased but also reliable and coherent forecasts. This is shown with one such alternative, the Bayesian joint probability modeling approach.

1. Introduction

Ensemble forecasts of seasonal precipitation from coupled ocean–atmosphere general circulation models (GCMs) have mostly replaced traditional statistical forecasts as the basis of operational outlooks issued by many national weather services. For example, the National Centers for Environmental Prediction (NCEP) in the United States has operated its Climate Forecast System (CFS) since 2004 (Saha et al. 2014), the European Centre for Medium-Range Weather Forecasts (ECMWF) has operated its Seasonal Forecast System since 1997 (Molteni et al. 2011), and the Bureau of Meteorology (BOM) in Australia has operated its Predictive Ocean and Atmosphere Model for Australia

Corresponding author e-mail: Tongtiegang Zhao, tony.zhao@csiro.au

DOI: 10.1175/JCLI-D-16-0652.1

(POAMA) since 2002 (Marshall et al. 2014). GCMs are now widely considered the state-of-the-art ''dynamical'' method for predicting seasonal atmospheric and oceanic conditions and fluxes, in that they are beginning to offer forecast skill that is similar to, or better than, long-standing empirical, analog, or statistical methods for climate prediction (Barnston et al. 2012; DelSole et al. 2014; Loikith and Broccoli 2015).

While raw ensemble GCM forecasts are informative, their usefulness is hampered by three well-known deficiencies: they are usually biased, making them unsuitable for use in decision-support tools; the spread of ensembles can be either too narrow or too wide, leading to operational risks or overly conservative decisions; and the forecasts may not always be ''skillful,'' meaning that GCM forecasts can be less accurate than the naïve climatology forecasts traditionally used by decision-makers (Shukla and Lettenmaier 2013; Schepen and Wang 2014). Postprocessing is thus a necessary step before GCM forecasts can be practically applied (Wood et al. 2002, 2005; Gneiting et al. 2005; Wilks and Hamill 2007; Lerch and Thorarinsdottir 2013; Yuan et al. 2015; Baran and Lerch 2015).

Quantile mapping (QM), also called quantile–quantile transformation or distribution mapping, is a popular method for postprocessing ensemble GCM forecasts (e.g., Wood et al. 2002, 2005; Wood and Lettenmaier 2006; Hopson and Webster 2010; Shukla and Lettenmaier 2013; Yuan 2016). One of its first applications for GCM outputs was to correct seasonal climate forecasts for streamflow forecasting (Wood et al. 2002). Its popularity in seasonal forecasting has since grown, in part due to its extensive use to correct climatological bias in studies projecting future climate change (e.g., Wood et al. 2004; Piani et al. 2010; Bürger et al. 2013; Gudmundsson et al. 2012; Lafon et al. 2013; Bennett et al. 2014; Li et al. 2014; Mehrotra and Sharma 2016; Rajczak et al. 2016).

Seasonal climate forecasts differ from long-range climate change projections in a crucial way: seasonal forecasts can be paired with observations. This allows the estimation of forecast skill and supports a wide range of strategies for postprocessing that include not only bias correction but also statistical calibration (Gneiting et al. 2007). In contrast, long-range GCM projections are not synchronous with observations (Hawkins and Sutton 2011), and thus the skill of future climate projections cannot be ''verified'' (Maraun 2016). Verification is a key concept for forecasts, including the consideration of reliability, which is the ability of forecast probabilities (or quantiles) to match their observed frequencies over time (Hagedorn et al. 2005; Doblas-Reyes et al. 2005; Gneiting et al. 2007). In forecasting, the term ''skill'' is usually used to mean the degree to which a forecast outperforms a benchmark or reference forecast, which for climate variables is typically defined as the climatology of observations (Murphy 1993; Hersbach 2000; Wilks and Hamill 2007).

Here, we present a case study to illustrate the strengths and weaknesses of QM in the context of seasonal GCM forecasts and facilitate discussion of the extent to which quantile-mapped GCM ensemble forecasts may or may not be ''better'' than climatology forecasts. Specifically, we examine the ability of QM to 1) remove bias from raw GCM forecasts, 2) make the forecast ensemble spread reliable, and 3) yield forecasts that are equivalent to climatology where there is no evident skill in raw GCM forecasts, a property termed ''coherence'' (Krzysztofowicz 1999). To this end, we use QM to postprocess raw ensemble precipitation forecasts from one GCM for the Australian continent. We identify where QM works well and where it does not, and describe which factors influence the performance of QM. We also compare the performance of QM to a full statistical forecast calibration using the Bayesian joint probability (BJP) modeling approach (Wang and Robertson 2011; Hawthorne et al. 2013; Schepen and Wang 2014).

2. Methods

a. *Quantile mapping*

QM matches the cumulative distribution function (CDF) of raw forecasts to the CDF of observations. For ensemble forecasts, the matching takes place at the level of individual ensemble members. Let x and x' denote raw and postprocessed forecasts, respectively, and $F_X(\,)$ and $F_O(\,)$ denote the CDFs of raw forecasts and observations. QM is formulated as

$$x' = F_O^{-1}[F_X(x)]. \tag{1}$$

Applying $F_O(\,)$ to the left- and right-hand sides of Eq. (1) yields

$$F_O(x') = F_X(x). \tag{2}$$

According to Eqs. (1) and (2), a new raw forecast value is postprocessed in two steps. First, a quantile fraction (or cumulative probability) is determined for the raw forecast by its position in the CDF of (preceding) forecasts. Second, a new postprocessed value of the forecast ensemble member is generated by ''looking up'' that quantile in the CDF of the observations.

QM is conceptually simple and can be implemented relatively easily: $F_X(\,)$ and $F_O(\,)$ can be derived

parametrically by fitting a distribution to the data (Piani et al. 2010; Gudmundsson et al. 2012) or nonparametrically through an empirical distribution function, "lookup table," or kernel density estimation (Wood et al. 2002, 2004; Bennett et al. 2014). Parametric distributions have the advantage of being less influenced by sampling errors, particularly with small samples, and generally produce mapping functions that are more stable (Lafon et al. 2013). They also provide a ready means for extrapolation when new forecast values are beyond the limits of the sample data used to form the CDFs. In this study, we use the setup of QM as described by Piani et al. (2010). The CDFs take the form of a mixed Bernoulli–gamma distribution: the Bernoulli distribution handles the probability of precipitation, while the gamma distribution characterizes precipitation amounts greater than zero. In cases where there exist a substantial number of outliers and the mixed distribution cannot be fitted, a nonparametric empirical cumulative distribution function is derived from the data (Gudmundsson et al. 2012).

b. *Full statistical calibration using BJP*

Numerous methods are available to statistically calibrate GCM climate forecasts; that is, to correct bias and unreliable ensemble spread, and to ensure coherence. Calibration methods typically use techniques that are more mathematically complex than QM to consider more explicitly the relationship between raw forecasts and observations (Gneiting et al. 2005; Wilks and Hamill 2007; Voisin et al. 2010). These include methods akin to model output statistics (MOS; Glahn and Lowry 1972) and more recently the extension of MOS to ensemble predictions (EMOS; Gneiting et al. 2005). Here the terms MOS and EMOS refer to methods that recognize that the correlation between ensemble forecasts and observations is imperfect, and thus that postprocessing requires a random component. We note that the term MOS has also been used in the climate projection literature to refer to any statistical postprocessing methods that relate model output to observations (Maraun et al. 2010). This more general use of MOS encompasses QM. However, the meaning of MOS is much narrower in the forecasting literature, where MOS refers to regression methods that account for the correlation between forecasts and observations (Glahn and Lowry 1972; Gneiting et al. 2005; Wilks and Hamill 2007). In this study, we use this narrower definition.

We apply BJP, an EMOS-type method. BJP formulates a joint probability distribution to characterize the relationship between the raw GCM ensemble mean and observations (Wang and Robertson 2011). The log-sinh transformation is employed to normalize the precipitation variables:

$$\begin{cases} \hat{x} = \dfrac{1}{\beta_x}\ln(\sinh[\alpha_x + \beta_x x]) \\ \hat{y} = \dfrac{1}{\beta_y}\ln(\sinh[\alpha_y + \beta_y y]) \end{cases}, \tag{3}$$

where $\hat{x}$ and $\hat{y}$ respectively represent transformed forecasts and observations; α_x, β_x, α_y, and β_y are transformation parameters. The transformation also ensures that the marginal distributions are homoscedastic so that the joint distribution can be used (Wang et al. 2012). The joint probability distribution is assumed to take the form of a bivariate normal distribution:

$$p(\hat{x}, \hat{y}) \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma}), \tag{4}$$

where $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ are respectively the mean vector and the covariance matrix; that is, $\boldsymbol{\mu} = \begin{bmatrix} \mu_{\hat{x}} \\ \mu_{\hat{y}} \end{bmatrix}$ ($\mu_{\hat{x}}$ and $\mu_{\hat{y}}$ are respectively the mean of $\hat{x}$ and $\hat{y}$) and $\boldsymbol{\Sigma} = \begin{bmatrix} \sigma_{\hat{x}}^2 & \rho_{\hat{x}\hat{y}}\sigma_{\hat{x}}\sigma_{\hat{y}} \\ \rho_{\hat{x}\hat{y}}\sigma_{\hat{x}}\sigma_{\hat{y}} & \sigma_{\hat{y}}^2 \end{bmatrix}$ ($\sigma_{\hat{x}}$ and $\sigma_{\hat{y}}$ are respectively the standard deviation of $\hat{x}$ and $\hat{y}$, and $\rho_{\hat{x}\hat{y}}$ is the correlation between $\hat{x}$ and $\hat{y}$).

In BJP, we obtain the transformation parameters using the maximum a posteriori probability (MAP) estimation and sample the parameters of the joint distribution with a Bayesian Markov chain Monte Carlo (MCMC) algorithm. A crucial feature of BJP is the covariance matrix, which explicitly describes how well transformed raw GCM forecasts are correlated with transformed observations. The presence of zero values, which is typical of precipitation in dry regions, is handled using data censoring. Zero values are treated as unknown values less than or equal to zero, which facilitates the use of the continuous bivariate normal distribution. The parameters of the BJP are estimated using past forecasts and observations as training samples. When given a new raw forecast, a univariate normal conditional distribution can be derived from the joint bivariate normal distribution. To generate a postprocessed forecast, random samples are drawn from the univariate conditional distribution and then back-transformed.

The mathematical techniques of data transformation, joint probability, and data censoring efficiently deal with heteroscedastic data, and with zero and missing values. They make BJP flexible for climatic and hydrologic modeling. BJP was originally developed for seasonal streamflow forecasting for the Australian Bureau of Meteorology (http://www.bom.gov.au/water/ssf). In recent years, this method has also been applied to

postprocess seasonal climate forecasts (Hawthorne et al. 2013; Peng et al. 2014; Schepen and Wang 2014) and subdaily numerical weather predictions (Robertson et al. 2013; Shrestha et al. 2015). Inferring BJP parameters is reasonably fast (less than 1 minute per grid cell), and forecast sampling takes only a few seconds per grid cell (Bennett et al. 2016). Therefore, compared to QM, BJP does not incur a prominent computational cost.

3. Data and experiment

a. GCM forecasts and observations

GCM precipitation forecasts are obtained from POAMA, the operational seasonal forecasting GCM run by the Bureau of Meteorology. POAMA couples the Australian Community Ocean Model version 2 (ACOM2) and the Bureau of Meteorology's atmospheric model version 3 (BAM3) (Marshall et al. 2014). ACOM2 is based on the Geophysical Fluid Dynamics Laboratory modular ocean model version 2 and has 25 vertical levels. BAM3 is a spectral transform model with triangular truncation 47 and 17 vertical sigma levels. POAMA uses perturbed initial conditions and differing model physics to produce a 33-member ensemble. The forecasts are available at a daily time step, but we choose to use monthly forecasts here. This has the advantage of removing much of the noise in the data, which can confound predictions of monthly or seasonal variables. The horizontal resolution of forecasts is 2.5° × 2.5°, approximately 250 km × 250 km. Monthly POAMA reforecasts are available for 1981–2011, which provides a means of training and assessing the postprocessing methods. POAMA forecasts are initialized on the first day of each calendar month and run for 9 months into the future.

Precipitation observations are obtained from the Australian Water Availability Project (AWAP; http://www.bom.gov.au/jsp/awap/). AWAP integrates in situ precipitation observations from up to 7000 stations across Australia and provides high-quality monthly spatial precipitation datasets from 1900 until now. The dataset is at a spatial resolution of 0.05° × 0.05°, approximately 5 km × 5 km. For this study, we regrid AWAP precipitation to match the POAMA horizontal grid resolution.

b. Experimental design

QM and BJP are applied to postprocess POAMA monthly precipitation reforecasts for 12 target months from 1981 to 2011. The QM and BJP methods are developed for each target month and lead time. For each lead time, there are 1464 cases (122 grid cells × 12 target months). Here we only present detailed results for 0-month lead time forecasts—that is, forecasts for the month that immediately follows the issue time. Limited commentary is provided on results for lead times of 1–8 months, because the performance of postprocessing methods at the 0-month lead time is consistent with all other lead times. All 33 ensemble members are used to fit the CDFs for QM, while BJP is trained using the ensemble mean. POAMA forecasts are postprocessed through leave-one-year-out cross-validation. Specifically, when postprocessing a forecast in one year, QM and BJP are trained with forecasts and observations from the other years. Then, the cross-validated forecasts are pooled in verification. In this way, we do not artificially inflate the performances of QM and BJP.

c. Forecast verification

Forecasts are verified in terms of bias, reliability, and skill. To illustrate bias, forecast and observed climatological means are plotted together in a scatterplot. Unbiased forecasts will display a 1:1 relationship.

Forecast reliability describes the ability of the ensemble spread to accurately represent the predictive uncertainty (Murphy 1993; Gneiting et al. 2007). Forecast ensembles that are too narrow (overconfident) can lead to operational risk, while ensembles that are too wide (underconfident) can result in decisions that are overly cautious (Arnal et al. 2016). A reliable forecast ensemble is neither too narrow nor too wide. To illustrate reliability, we pool the pairwise forecasts and observations for all locations and all months to construct two types of plots. The reliability of full ensemble forecasts is assessed by the histogram of the probability integral transform (PIT) values under each pair of forecast and observation:

$$\mathrm{PIT} = \begin{cases} F(y_{\mathrm{obs}}) & (y_{\mathrm{obs}} > 0) \\ u \times F(0) & (y_{\mathrm{obs}} = 0) \end{cases}, \tag{5}$$

where $F(\,)$ is the CDF obtained from the ensemble forecast, y_{obs} is the corresponding observation, and $u \sim U(0, 1)$ is a random number for censored data. When the ensemble forecast reliably captures the distribution of observation, the observation y_{obs} can statistically be regarded as random samples drawn from $F(\,)$. That is, the individual PIT values for different observations should collectively follow a uniform distribution. Therefore, the reliability of ensemble spread is shown by the uniformity of the PIT histogram. In addition, the reliability of forecast probability is examined by treating forecasts as binary, in this case whether precipitation exceeds the median, and is assessed by plotting an attributes diagram (Hsu and Murphy 1986).

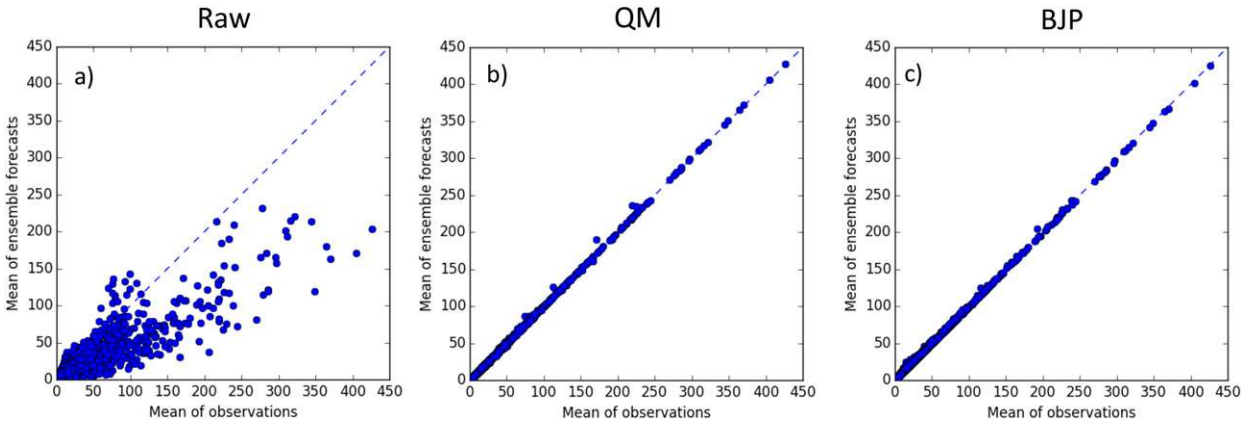

FIG. 1. Diagnostic plots of bias of raw and postprocessed forecasts at 0-month lead time, generated by pooling all POAMA grid cells and all months.

The forecast probabilities are reliable when they are consistent with the relative frequency of observations, manifested by points that follow the diagonal 1:1 line.

Forecast skill is quantified using a skill score based on the continuous ranked probability score (CRPS; Hersbach 2000; Gneiting et al. 2007):

$$\mathrm{CRPS} = \int_{-\infty}^{+\infty} [F(y) - H(y - y_{\mathrm{obs}})]^2 \, dy, \qquad (6)$$

where $H(\ )$ is a Heaviside step function that indicates the cumulative distribution of observation y_{obs}. That is, the value of $H(y - y_{\mathrm{obs}})$ is 0 when $y - y_{\mathrm{obs}} < 0$ and is 1 when $y - y_{\mathrm{obs}} \geq 0$. Thus, CRPS measures the difference between the observed distribution $H(\ )$ and the forecast distribution $F(\ )$. It is essentially a probability-weighted average of the error of each ensemble member. We use the CRPS of a naïve climatology forecast as the reference and define the skill score as the ratio (%) of the reduction in CRPS for QM- and BJP-postprocessed forecasts. The climatology forecasts are generated by fitting a log-sinh transformed normal distribution to observations. The climatology distributions are fitted under the same leave-one-year-out cross-validation scheme used to train the QM- and BJP-postprocessed forecasts.

4. Results

a. Forecast bias and reliability

Raw POAMA forecasts are generally biased and have a tendency to underestimate observed precipitation (Fig. 1a). As a result, reliability is poor. This is strongly evident in the PIT histogram describing raw forecasts for all grid cells (Fig. 2a), which is heavily left-skewed. The effect of bias on reliability is also strongly evident in the corresponding attributes diagram (Fig. 2b), with obvious mismatches between forecast probabilities and observed relative frequencies.

QM is extremely effective at correcting bias in the raw forecasts. The climatological mean of the QM-postprocessed forecasts for each grid cell is virtually indistinguishable from the corresponding climatological mean of observations (Fig. 1b). Removing bias can improve forecast reliability, reducing the skewness of the PIT histogram (Fig. 2c). However, because QM does not explicitly address spread errors in the raw ensemble, QM-postprocessed forecasts are still not reliable: the PIT histogram is U-shaped. This indicates that the ensemble spread tends to be too narrow. In other words, the forecasts are overconfident. The attributes diagram for QM-postprocessed forecasts provides corroborating evidence that QM relieves forecast bias but does not lead to fully reliable probabilistic forecasts (Fig. 2d).

By contrast, BJP generates unbiased ensemble forecasts that are also reliable. As with QM-postprocessed forecasts, the climatological mean of BJP-postprocessed forecasts is virtually indistinguishable from the climatological mean of observations (Fig. 1c). The PIT histogram is practically uniform (Fig. 2e). The BJP-postprocessed forecast ensemble spread is slightly too wide, but nonetheless close to being reliable in representing the distribution of observations. The reliability of BJP-postprocessed forecasts of the probability for exceeding the climatological median is clearly shown in the attributes diagram, with points very close to the 1:1 line (Fig. 2f).

To summarize, both QM and BJP are highly effective at removing climatological biases in raw POAMA forecasts. However, they differ in their abilities to correct forecast spread and provide reliable forecasts. QM-postprocessed forecasts exhibit overly narrow ensemble spreads, and are therefore not reliable. In comparison,

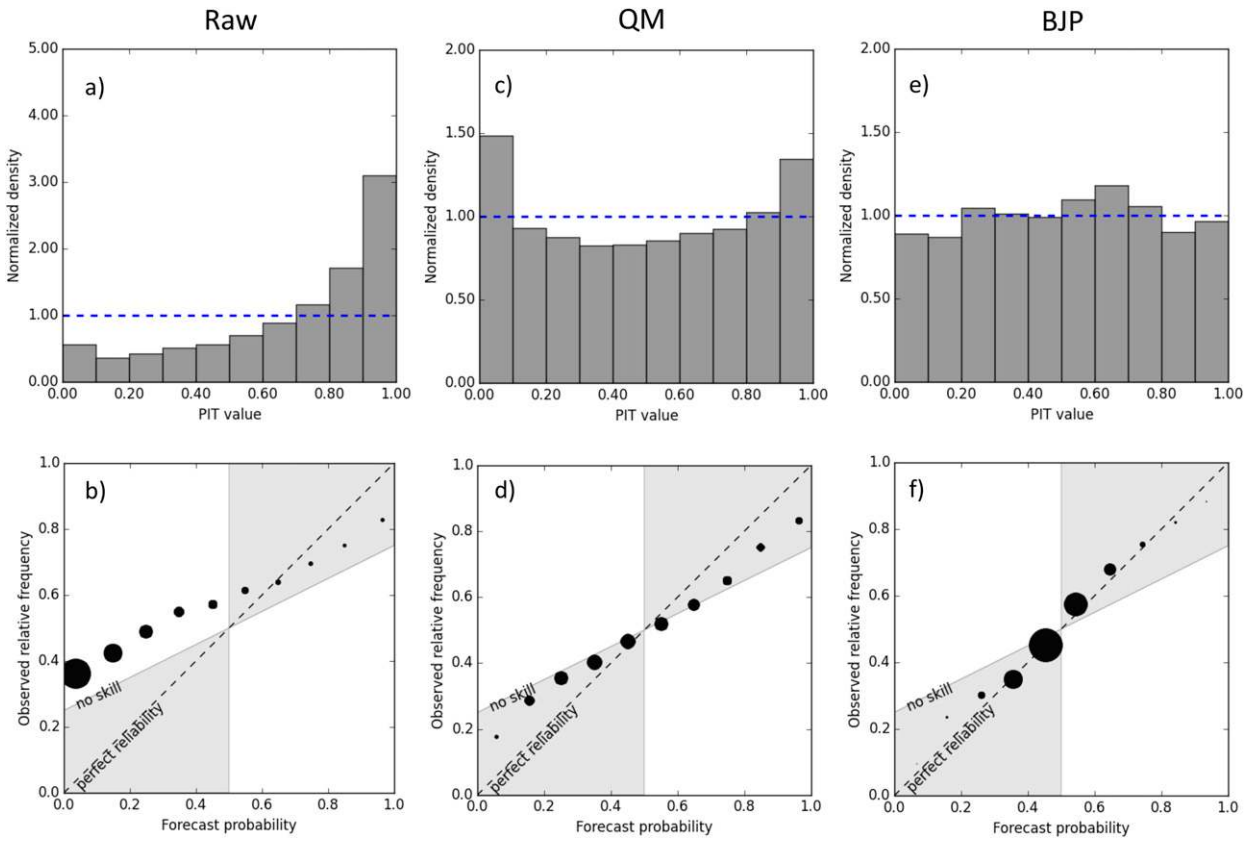

FIG. 2. Diagnostic plots of reliability of raw and postprocessed forecasts at 0-month lead time, generated by pooling all POAMA grid cells and all months, showing (top) PIT histograms and (bottom) attributes diagrams for forecasts of probabilities of exceeding the median observation. Forecasts in attributes diagrams are divided into ten forecast probability bins; point sizes indicate relative number of forecasts in each bin.

BJP-postprocessed forecasts are generally reliable. Note that although we have concentrated our analysis on forecasts at 0-month lead time, all other lead times (1–8 months) give similar results.

b. *Forecast skill*

The CRPS skill scores of postprocessed forecasts for each target month are mapped in Fig. 3. QM yields many instances of positive skill (blue pixels in Fig. 3a). However, QM-postprocessed forecasts can also exhibit strongly negative skill (red pixels in Fig. 3a). In other words, QM-postprocessed forecasts can, in some cases, have larger errors than naïve climatology forecasts: they do not exhibit ''coherence.'' In contrast, BJP-postprocessed forecasts are overall at least as skillful as climatology forecasts and therefore coherent (Fig. 3b). There are only a few cases where slightly negative forecast skills are observed (red pixels in Fig. 3b), which is attributable to cross-validation and sampling uncertainty.

The differences in forecast skill arise primarily from how the relationship between raw GCM forecasts and observations is formulated in postprocessing. QM does not consider the strength of the association between raw GCM forecasts and observations and consequently generates both positively and negatively skillful forecasts. In contrast, BJP ensures that CRPS skill scores, measured relative to climatology, will be nearly zero; that is, if raw forecasts are uncorrelated with observations, a forecast approximating climatology will be returned (Wang and Robertson 2011; Hawthorne et al. 2013; Schepen and Wang 2014).

c. *Influence of raw forecast performance*

To show the influence of raw forecast performance on postprocessed forecasts, Spearman's rank correlation is calculated for the ensemble mean of raw GCM forecasts and the corresponding observations. The rank correlation quantifies the strength of the forecast–observation relationship without assuming that it is linear. The CRPS skill scores of QM- and BJP-postprocessed forecasts are plotted against rank correlations in Fig. 4.

Both QM and BJP tend to produce forecasts that have positive CRPS skill scores if the raw forecasts are

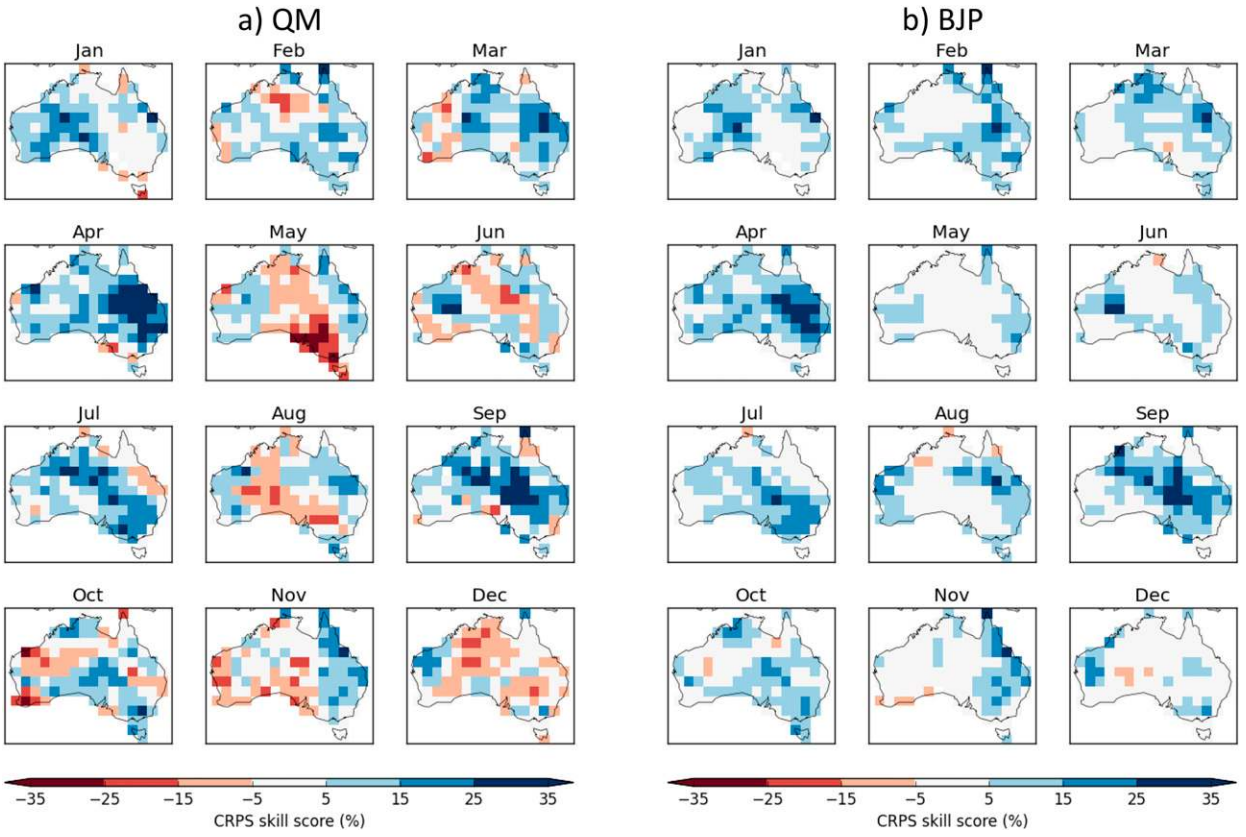

FIG. 3. CRPS skill score of (a) QM- and (b) BJP-postprocessed forecasts at 0-month lead time (climatology forecasts are used as the reference). Blue pixels (>5%) show positively skillful forecasts; white pixels (>−5%, <5%) show neutrally skillful forecasts; red pixels (<−5%) show negatively skillful forecasts.

significantly positively correlated with observations (p value < 0.10). However, QM-postprocessed forecasts can have negative skill scores when the correlation between raw forecasts and observations is low and/or nonsignificant. By contrast, BJP-postprocessed forecasts have similar accuracy to climatology forecasts when there is little or even negative correlation between raw forecasts and observations. We note that the skill of raw GCM forecasts typically diminishes as lead time increases, leading to more instances where raw forecasts are not strongly correlated with observations (Schepen and Wang 2014; Peng et al. 2014; Kumar et al. 2015). This means that at longer lead times there are more instances when QM-postprocessed forecasts are not coherent than shown for the 0-month lead time in Fig. 4.

We also consider the reliability of postprocessed forecasts in light of whether raw POAMA forecasts are significantly positively correlated with observations or not (Fig. 5). Forecasts are separated into 1) cases where positive correlations between the raw forecast ensemble mean and observations are statistically significant (p value < 0.10) and 2) all other cases. The PIT histograms for QM-postprocessed forecasts are U-shaped, indicating overconfident forecasts, regardless of whether the raw forecasts are significantly positively correlated with observations (Fig. 5a) or not (Fig. 5c). Attributes diagrams also show that the QM-postprocessed forecasts are overconfident (Figs. 5b, d), although to different degrees: when correlations are not significantly positive (Fig. 5d), the attributes diagram deviates more markedly from the 1:1 line.

By contrast, BJP-postprocessed forecasts tend to be reliable, irrespective of whether raw POAMA forecasts are significantly positively correlated with observations or not. The PIT histograms in Figs. 5e and 5g show that the BJP-postprocessed forecasts have similarly reliable ensemble spread in all cases. The attributes diagram for the cases when raw POAMA forecasts are significantly positively correlated with observations (Fig. 5f) demonstrates that the forecast probabilities of exceeding the climatological median are consistent with observed relative frequencies (i.e., the forecasts are reliable). When raw forecasts are not significantly positively correlated with observations, the attributes diagram shows that

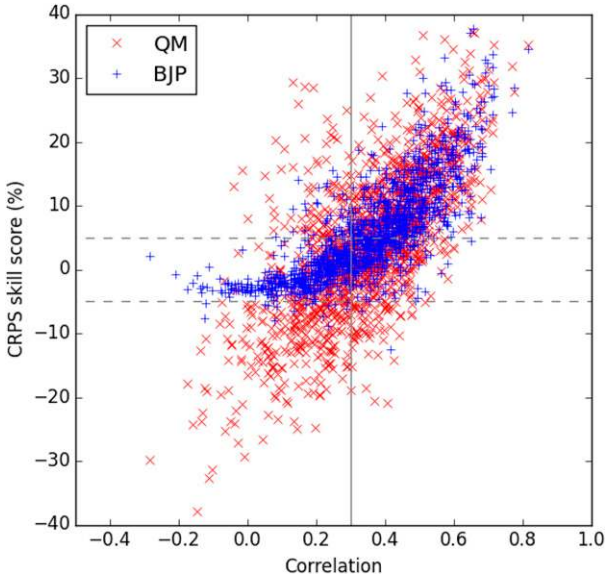

FIG. 4. Relationship between CRPS skill score of postprocessed forecasts and the correlation between raw ensemble mean and observation at 0-month lead time. The vertical solid line divides correlation into two categories: grid cells with significantly positive correlation (p value < 0.10) (right of line) and all other grid cells (left of line). The two horizontal dashed lines divide CRPS skill score into three groups: positively skillful, neutrally skillful, and negatively skillful.

BJP forecast probabilities concentrate around the climatological probability of 0.5 (Fig. 5h). This shows that the BJP is performing as expected: when raw forecasts are essentially uninformative, the BJP returns a forecast akin to climatology.

5. Discussion

QM can be useful in instances when bias is the only, or main, deficiency of raw forecasts, as it is highly effective at removing bias (Crochemore et al. 2016). By removing bias, QM also improves the reliability of GCM forecasts to some extent. However, as we have shown, QM cannot correct overconfidence in the raw ensemble spread, and thus cannot guarantee reliable forecasts. It is usual for GCMs to produce ensemble forecasts that are not only biased but also have ensemble spread that is overconfident (Gneiting et al. 2005; Wilks and Hamill 2007; Lerch and Thorarinsdottir 2013; Baran and Lerch 2015; Yuan 2016). Forecasts that are not reliable, even if they are unbiased, may be of little use for decisions where users need to weigh the risk of action against the uncertainty inherent in a forecast (e.g., Arnal et al. 2016; Raftery 2016).

The property of coherence is a prerequisite for forecasts to be valuable to decision makers. When introducing the term "coherence," Krzysztofowicz (1999) argued that forecasts should be at least as informative as a Bayesian prior—that is, a naïve reference forecast—to allow them to have formal economic value to decision makers. Climatology has long played the role of the naïve reference forecast in theoretical discussions of forecast verification and economic value (Murphy 1993; Hersbach 2000; Wilks and Hamill 2007). Climatology is also the de facto benchmark in most practical applications of seasonal precipitation forecasts. For example, resampled historical rainfall, a form of climatological precipitation forecast, has been used for decades in seasonal ensemble streamflow prediction (ESP) systems (e.g., Day 1985). The fact that QM does not ensure coherent forecasts, as we have shown, is thus a serious limitation to its value as a postprocessor of seasonal GCM precipitation forecasts.

The inability of QM to produce fully reliable and coherent forecasts has been pointed out before in contexts other than seasonal precipitation forecasting. Wood and Schaake (2008) compared QM and regression-based postprocessing for ensemble streamflow forecasts, showing that QM did not produce reliable and coherent streamflow forecasts while regression provided calibration benefits similar to those afforded here by BJP. In addition, the ability of a variety of EMOS-type methods to produce reliable and coherent ensemble forecasts is well established. These include the logistic regression model, which postprocesses raw forecasts and generates probability forecasts for selected quantiles (Wilks and Hamill 2007); the nonhomogeneous Gaussian regression (NGR) model, which employs multivariate linear regression to account for the relationship between raw ensemble forecasts and observations (Gneiting et al. 2005); and variants of the NGR model that use other distributions, such as generalized extreme value and lognormal distributions (Lerch and Thorarinsdottir 2013; Baran and Lerch 2015). Similarly to BJP, these EMOS methods explicitly consider the relationship between raw forecasts and observations in postprocessing, and any of these methods is likely to be preferable to QM for postprocessing seasonal climate forecasts.

We note that EMOS-type methods are not without potential limitations. The statistical assumptions underlying any EMOS method underpin its effectiveness. For example, the BJP model assumes a joint bivariate normal distribution for the transformed forecasts and observations. Similarly, the NGR model assumes that the observation follows a normal distribution conditional on the ensemble mean. If the forecasts and observations do not satisfy the underlying assumptions, the performance of EMOS methods will be impacted. Further, EMOS methods all depend, to some degree, on "doubting" the forecasts (by applying a regression of some form) and adding uncertainty. If the underlying

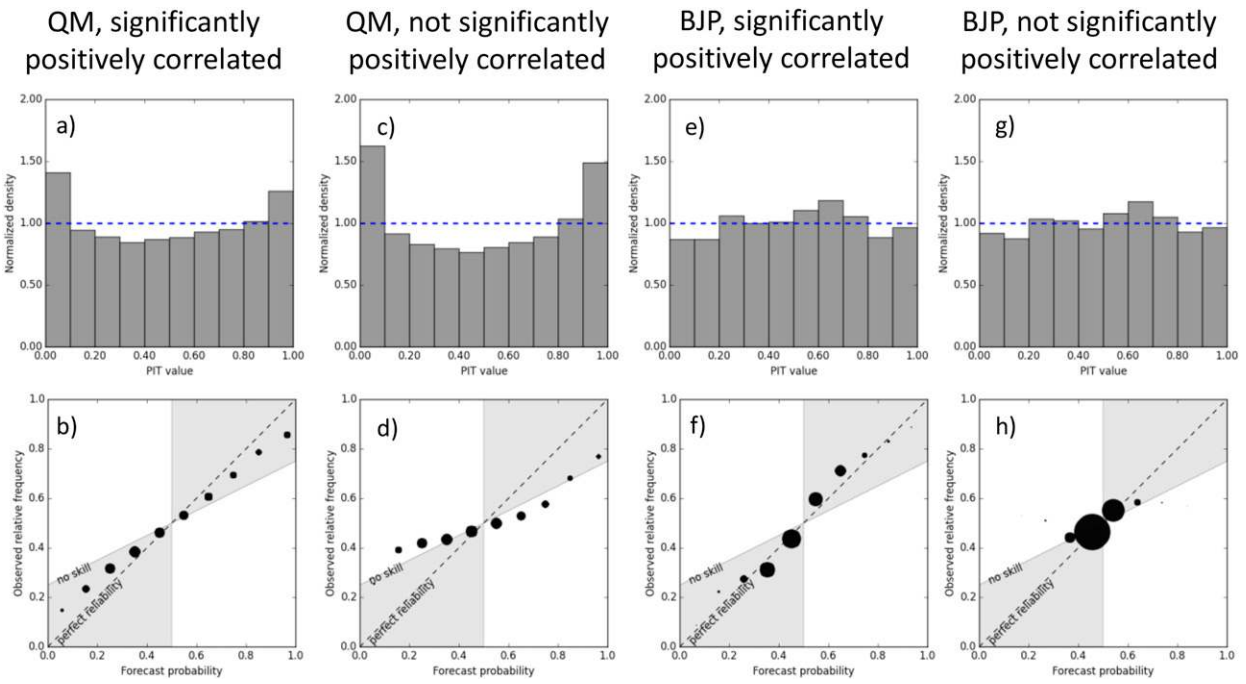

FIG. 5. Reliability conditioned on raw forecast performance. (left) QM-corrected forecasts, showing (a),(b) cases in which raw POAMA forecasts are significantly positively correlated with observations (p value < 0.10) and (c),(d) all other cases. (right) BJP-postprocessed forecasts, showing (e),(f) cases where raw forecasts are significantly positively correlated with observations and (g),(h) all other cases.

forecasts are extremely good, it may be better to simply ''trust'' the forecast in the way that QM does—for example, in a few cases where forecasts have high skill, QM-postprocessed POAMA forecasts have (slightly) lower errors than BJP-postprocessed forecasts (Figs. 3 and 4). On the whole, however, EMOS-type methods are conceptually preferable to QM for seasonal forecast postprocessing.

For general applications, QM has a few other limitations not discussed so far. Maraun (2013) pointed out that in using QM as a tool for downscaling, it can produce outputs that have larger variance than observations when the outputs are rescaled to the original domain (''variance inflation''). This problem arises from the same root cause as the reliability and coherence issues described above—that is, because QM does not consider the correlation between forecasts and observations. Another way of putting this is that QM assumes that raw forecasts will be perfectly correlated with observations, and thus it is justifiable to apply a deterministic correction. EMOS methods recognize that the correlation between forecasts and observations is imperfect, and thus that postprocessing requires a random component [as recommended by Maraun (2013) and Wong et al. (2014)]. This not only ensures that forecasts are coherent and reliable, it also means EMOS methods can be adapted to downscaling applications without causing variance inflation (Robertson et al. 2013; Schepen and Wang 2014; Shrestha et al. 2015).

Our findings on the usefulness of QM for postprocessing seasonal forecasts are not readily transferable to the cases where QM is applied to correct biases in climate projections. A number of studies have shown that QM is highly effective for correcting the bias of climate projections (e.g., Wood et al. 2004; Piani et al. 2010; Mehrotra and Sharma 2016). When QM is applied to climate projections from an ensemble of models, it can also adjust the ensemble spread of the projections (e.g., Maraun et al. 2010; Wong et al. 2014; IPCC 2015). When a larger ensemble spread is desired, Kim et al. (2016) suggests a procedure that applies the QM in a more selective manner, to only some of the ensemble members. However, it is not possible to objectively evaluate how reliable the ensemble spread is, because climate projections are not synchronous with observations (Maraun 2016). Accordingly, it is not possible to calibrate climate projections using the BJP method as we have presented it in this paper. Note that use of QM for climate projections also has other limitations as reported, for example, in Bürger et al. (2013), Maraun (2013), and IPCC (2015), which we do not discuss here in the interests of brevity.

Our findings may be relevant to specific aspects of climate projections. For example, climate projections

aggregated over time may be able to be treated as synchronous with observations, for example the mean climate over longer periods (e.g., 20-yr mean temperature), or trends in climate variables taken over a sufficiently long period (e.g., the rate of change in temperature per decade). If these variables can be considered synchronous with observations, it may be possible to apply EMOS-type methods to long-range climate projections. Better representations of the spread of climate projections may be useful, for example, in the attribution of extreme events to human influence (Bellprat and Doblas-Reyas 2016).

We have illustrated that QM does not ensure forecast reliability and coherence, while EMOS-type methods, exemplified by the BJP, do. This may already be appreciated by experienced forecast practitioners, but perhaps because QM's limitations vis-à-vis EMOS-type methods have not been demonstrated for GCM precipitation forecasts, QM continues to be applied (e.g., Hopson and Webster 2010; Mo et al. 2012; Shukla and Lettenmaier 2013; Yuan 2016). Given the serious limitations of QM and the ready availability of more comprehensive postprocessing methods, we caution against the use of QM in forecasting applications.

6. Summary and conclusions

In this study, we investigate the performance of QM for postprocessing seasonal GCM precipitation forecasts. QM is highly effective at correcting bias in the raw GCM forecasts. However, QM ignores the correlation between raw ensemble forecasts and observations. When raw forecasts are not significantly positively correlated with observations, QM tends to produce negatively skillful forecasts. Even when there is significantly positive correlation, QM cannot ensure reliability and guarantee that postprocessed forecasts are ''coherent''—that is, at least as skillful as climatology forecasts. The flaws of QM in postprocessing GCM forecasts cast doubt on the usefulness of this popular method. Based on these findings, we contend that QM is not a wholly satisfactory method for postprocessing GCM precipitation forecasts. These findings can be generalized to postprocessing other types of forecasts. Alternative EMOS-type postprocessing methods are available for producing forecasts that are not only unbiased but also reliable in ensemble spread and at least as skillful as climatology.

Acknowledgments. This research has been supported by the Water Information Research and Development Alliance (WIRADA) between the Australian Bureau of Meteorology and CSIRO Land and Water. The precipitation data are available via the AWAP data portal http://www.bom.gov.au/jsp/awap/. The POAMA forecast archive is available at http://poama.bom.gov.au/. We thank three anonymous reviewers for constructive and thorough reviews.

REFERENCES

Arnal, L., M.-H. Ramos, E. Coughlan de Perez, H. L. Cloke, E. Stephens, F. Wetterhall, S. J. van Andel, and F. Pappenberger, 2016: Willingness-to-pay for a probabilistic flood forecast: A risk-based decision-making game. *Hydrol. Earth Syst. Sci.*, **20**, 3109–3128, doi:10.5194/hess-20-3109-2016.

Baran, S., and S. Lerch, 2015: Log-normal distribution based ensemble model output statistics models for probabilistic wind-speed forecasting. *Quart. J. Roy. Meteor. Soc.*, **141**, 2289–2299, doi:10.1002/qj.2521.

Barnston, A. G., M. K. Tippett, M. L. L'Heureux, S. H. Li, and D. G. DeWitt, 2012: Skill of real-time seasonal ENSO model predictions during 2002–11: Is our capability increasing? *Bull. Amer. Meteor. Soc.*, **93**, 631–651, doi:10.1175/BAMS-D-11-00111.1.

Bellprat, O., and F. Doblas-Reyes, 2016: Attribution of extreme weather and climate events overestimated by unreliable climate simulations. *Geophys. Res. Lett.*, **43**, 2158–2164, doi:10.1002/2015GL067189.

Bennett, J. C., M. R. Grose, S. P. Corney, C. J. White, G. K. Holz, J. J. Katzfey, D. A. Post, and N. L. Bindoff, 2014: Performance of an empirical bias-correction of a high-resolution climate dataset. *Int. J. Climatol.*, **34**, 2189–2204, doi:10.1002/joc.3830.

——, Q. J. Wang, M. Li, D. E. Robertson, and A. Schepen, 2016: Reliable long-range ensemble streamflow forecasts: Combining calibrated climate forecasts with a conceptual runoff model and a staged error model. *Water Resour. Res.*, **52**, 8238–8259, doi:10.1002/2016WR019193.

Bürger, G., S. R. Sobie, A. J. Cannon, A. T. Werner, and T. Q. Murdock, 2013: Downscaling extremes: An intercomparison of multiple methods for future climate. *J. Climate*, **26**, 3429–3449, doi:10.1175/JCLI-D-12-00249.1.

Crochemore, L., M.-H. Ramos, and F. Pappenberger, 2016: Bias correcting precipitation forecasts to improve the skill of seasonal streamflow forecasts. *Hydrol. Earth Syst. Sci.*, **20**, 3601–3618, doi:10.5194/hess-20-3601-2016.

Day, G. N., 1985: Extended streamflow forecasting using NWSRFS. *J. Water Resour. Plann. Manage.*, **111**, 157–170, doi:10.1061/(ASCE)0733-9496(1985)111:2(157).

DelSole, T., X. Yan, P. A. Dirmeyer, M. Fennessy, and E. Altshuler, 2014: Changes in seasonal predictability due to global warming. *J. Climate*, **27**, 300–311, doi:10.1175/JCLI-D-13-00026.1.

Doblas-Reyes, F. J., R. Hagedorn, and T. N. Palmer, 2005: The rationale behind the success of multi-model ensembles in seasonal forecasting—II. Calibration and combination. *Tellus*, **57A**, 234–252, doi:10.1111/j.1600-0870.2005.00104.x.

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.

Gneiting, T., A. E. Raftery, A. H. Westveld, and T. Goldman, 2005: Calibrated probabilistic forecasting using ensemble model output statistics and minimum CRPS estimation. *Mon. Wea. Rev.*, **133**, 1098–1118, doi:10.1175/MWR2904.1.

——, F. Balabdaoui, and A. E. Raftery, 2007: Probabilistic forecasts, calibration and sharpness. *J. Roy. Stat. Soc.*, **69B**, 243–268, doi:10.1111/j.1467-9868.2007.00587.x.

Gudmundsson, L., J. B. Bremnes, J. E. Haugen, and T. Engen-Skaugen, 2012: Technical note: Downscaling RCM precipitation to the station scale using statistical transformations—A comparison of methods. *Hydrol. Earth Syst. Sci.*, **16**, 3383–3390, doi:10.5194/hess-16-3383-2012.

Hagedorn, R., F. J. Doblas-Reyes, and T. N. Palmer, 2005: The rationale behind the success of multi-model ensembles in seasonal forecasting—I. Basic concept. *Tellus*, **57A**, 219–233, doi:10.1111/j.1600-0870.2005.00103.x.

Hawkins, E., and R. Sutton, 2011: The potential to narrow uncertainty in projections of regional precipitation change. *Climate Dyn.*, **37**, 407–418, doi:10.1007/s00382-010-0810-6.

Hawthorne, S., Q. J. Wang, A. Schepen, and D. Robertson, 2013: Effective use of general circulation model outputs for forecasting monthly rainfalls to long lead times. *Water Resour. Res.*, **49**, 5427–5436, doi:10.1002/wrcr.20453.

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.

Hopson, T. M., and P. J. Webster, 2010: A 1–10-day ensemble forecasting scheme for the major river basins of Bangladesh: Forecasting severe floods of 2003–07. *J. Hydrometeor.*, **11**, 618–641, doi:10.1175/2009JHM1006.1.

Hsu, W.-R., and A. H. Murphy, 1986: The attributes diagram—A geometrical framework for assessing the quality of probability forecasts. *Int. J. Forecast.*, **2**, 285–293, doi:10.1016/0169-2070(86)90048-8.

IPCC, 2015: IPCC workshop on regional climate projections and their use in impacts and risk analysis studies. T. F. Stocker et al., Eds., IPCC Workshop Rep., 171 pp. [Available online at http://www.ipcc.ch/pdf/supporting-material/RPW_WorkshopReport.pdf.]

Kim, K. B., H.-H. Kwon, and D. W. Han, 2016: Precipitation ensembles conforming to natural variations derived from a regional climate model using a new bias correction scheme. *Hydrol. Earth Syst. Sci.*, **20**, 2019–2034, doi:10.5194/hess-20-2019-2016.

Krzysztofowicz, R., 1999: Bayesian theory of probabilistic forecasting via deterministic hydrologic model. *Water Resour. Res.*, **35**, 2739–2750, doi:10.1029/1999WR900099.

Kumar, A., M. Y. Chen, Y. Xue, and D. Behringer, 2015: An analysis of the temporal evolution of ENSO prediction skill in the context of the equatorial Pacific Ocean observing system. *Mon. Wea. Rev.*, **143**, 3204–3213, doi:10.1175/MWR-D-15-0035.1.

Lafon, T., S. Dadson, G. Buys, and C. Prudhomme, 2013: Bias correction of daily precipitation simulated by a regional climate model: A comparison of methods. *Int. J. Climatol.*, **33**, 1367–1381, doi:10.1002/joc.3518.

Lerch, S., and T. L. Thorarinsdottir, 2013: Comparison of non-homogeneous regression models for probabilistic wind speed forecasting. *Tellus*, **65A**, 21206, http://dx.doi.org/10.3402/tellusa.v65i0.21206.

Li, C., E. Sinha, D. E. Horton, N. S. Diffenbaugh, and A. M. Michalak, 2014: Joint bias correction of temperature and precipitation in climate model simulations. *J. Geophys. Res. Atmos.*, **119**, 13 153–13 162, doi:10.1002/2014JD022514.

Loikith, P. C., and A. J. Broccoli, 2015: Comparison between observed and model-simulated atmospheric circulation patterns associated with extreme temperature days over North America using CMIP5 historical simulations. *J. Climate*, **28**, 2063–2079, doi:10.1175/JCLI-D-13-00544.1.

Maraun, D., 2013: Bias correction, quantile mapping, and downscaling: Revisiting the inflation issue. *J. Climate*, **26**, 2137–2143, doi:10.1175/JCLI-D-12-00821.1.

——, 2016: Bias correcting climate change simulations—A critical review. *Curr. Climate Change Rep.*, **2**, 211–220, doi:10.1007/s40641-016-0050-x.

——, and Coauthors, 2010: Precipitation downscaling under climate change: Recent developments to bridge the gap between dynamical models and the end user. *Rev. Geophys.*, **48**, RG3003, doi:10.1029/2009RG000314.

Marshall, A. G., D. Hudson, M. C. Wheeler, O. Alves, H. H. Hendon, M. J. Pook, and J. S. Risbey, 2014: Intra-seasonal drivers of extreme heat over Australia in observations and POAMA-2. *Climate Dyn.*, **43**, 1915–1937, doi:10.1007/s00382-013-2016-1.

Mehrotra, R., and A. Sharma, 2016: A multivariate quantile-matching bias correction approach with auto- and cross-dependence across multiple time scales: Implications for downscaling. *J. Climate*, **29**, 3519–3539, doi:10.1175/JCLI-D-15-0356.1.

Mo, K. C., S. Shukla, D. P. Lettenmaier, and L.-C. Chen, 2012: Do climate forecast system (CFSv2) forecasts improve seasonal soil moisture prediction? *Geophys. Res. Lett.*, **39**, L23703, doi:10.1029/2012GL053598.

Molteni, F., and Coauthors, 2011: The new ECMWF Seasonal Forecast System (System 4). ECMWF Tech. Memo. 656, 49 pp.

Murphy, A. H., 1993: What is a good forecast? An essay on the nature of goodness in weather forecasting. *Wea. Forecasting*, **8**, 281–293, doi:10.1175/1520-0434(1993)008<0281:WIAGFA>2.0.CO;2.

Peng, Z. L., Q. J. Wang, J. C. Bennett, A. Schepen, F. Pappenberger, P. Pokhrel, and Z. R. Wang, 2014: Statistical calibration and bridging of ECMWF System4 outputs for forecasting seasonal precipitation over China. *J. Geophys. Res. Atmos.*, **119**, 7116–7135, doi:10.1002/2013JD021162.

Piani, C., J. O. Haerter, and E. Coppola, 2010: Statistical bias correction for daily precipitation in regional climate models over Europe. *Theor. Appl. Climatol.*, **99**, 187–192, doi:10.1007/s00704-009-0134-9.

Raftery, A. E., 2016: Use and communication of probabilistic forecasts, statistical analysis and data mining. *Stat. Anal. Data Min.: ASA Data Sci. J.*, **9**, 397–410, doi:10.1002/sam.11302.

Rajczak, J., S. Kotlarski, and C. Schär, 2016: Does quantile mapping of simulated precipitation correct for biases in transition probabilities and spell lengths? *J. Climate*, **29**, 1605–1615, doi:10.1175/JCLI-D-15-0162.1.

Robertson, D. E., D. L. Shrestha, and Q. J. Wang, 2013: Post-processing rainfall forecasts from numerical weather prediction models for short-term streamflow forecasting. *Hydrol. Earth Syst. Sci.*, **17**, 3587–3603, doi:10.5194/hess-17-3587-2013.

Saha, S., and Coauthors, 2014: The NCEP Climate Forecast System version 2. *J. Climate*, **27**, 2185–2208, doi:10.1175/JCLI-D-12-00823.1.

Schepen, A., and Q. J. Wang, 2014: Ensemble forecasts of monthly catchment rainfall out to long lead times by post-processing coupled general circulation model output. *J. Hydrol.*, **519**, 2920–2931, doi:10.1016/j.jhydrol.2014.03.017.

Shrestha, D. L., D. E. Robertson, J. C. Bennett, and Q. J. Wang, 2015: Improving precipitation forecasts by generating ensembles through postprocessing. *Mon. Wea. Rev.*, **143**, 3642–3663, doi:10.1175/MWR-D-14-00329.1.

Shukla, S., and D. P. Lettenmaier, 2013: Multi-RCM ensemble downscaling of NCEP CFS winter season forecasts: Implications

for seasonal hydrologic forecast skill. *J. Geophys. Res. Atmos.*, **118**, 10 770–10 790, doi:10.1002/jgrd.50628.

Voisin, N., J. C. Schaake, and D. P. Lettenmaier, 2010: Calibration and downscaling methods for quantitative ensemble precipitation forecasts. *Wea. Forecasting*, **25**, 1603–1627, doi:10.1175/2010WAF2222367.1.

Wang, Q. J., and D. E. Robertson, 2011: Multisite probabilistic forecasting of seasonal flows for streams with zero value occurrences. *Water Resour. Res.*, **47**, W02546, doi:10.1029/2010WR009333.

——, D. L. Shrestha, D. E. Robertson, and P. Pokhrel, 2012: A log-sinh transformation for data normalization and variance stabilization. *Water Resour. Res.*, **48**, W05514, doi:10.1029/2011WR010973.

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

Wong, G., D. Maraun, M. Vrac, M. Widmann, J. M. Eden, and T. Kent, 2014: Stochastic model output statistics for bias correcting and downscaling precipitation including extremes. *J. Climate*, **27**, 6940–6959, doi:10.1175/JCLI-D-13-00604.1.

Wood, A. W., and D. P. Lettenmaier, 2006: A test bed for new seasonal hydrologic forecasting approaches in the western United States. *Bull. Amer. Meteor. Soc.*, **87**, 1699–1712, doi:10.1175/BAMS-87-12-1699.

——, and J. C. Schaake, 2008: Correcting errors in streamflow forecast ensemble mean and spread. *J. Hydrometeor.*, **9**, 132–148, doi:10.1175/2007JHM862.1.

——, E. P. Maurer, A. Kumar, and D. P. Lettenmaier, 2002: Long-range experimental hydrologic forecasting for the eastern United States. *J. Geophys. Res.*, **107**, 4429, doi:10.1029/2001JD000659.

——, L. R. Leung, V. Sridhar, and D. P. Lettenmaier, 2004: Hydrologic implications of dynamical and statistical approaches to downscaling climate model outputs. *Climatic Change*, **62**, 189–216, doi:10.1023/B:CLIM.0000013685.99609.9e.

——, A. Kumar, and D. P. Lettenmaier, 2005: A retrospective assessment of National Centers for Environmental Prediction climate model–based ensemble hydrologic forecasting in the western United States. *J. Geophys. Res.*, **110**, D04105, doi:10.1029/2004JD004508.

Yuan, X., 2016: An experimental seasonal hydrological forecasting system over the Yellow River basin—Part 2: The added value from climate forecast models. *Hydrol. Earth Syst. Sci.*, **20**, 2453–2466, doi:10.5194/hess-20-2453-2016.

——, E. F. Wood, and Z. G. Ma, 2015: A review on climate-model-based seasonal hydrologic forecasting: Physical understanding and system development. *Wiley Interdiscip. Rev.: Water*, **2**, 523–536, doi:10.1002/wat2.1088.