On the Joint Calibration of Multivariate Seasonal Climate Forecasts from GCMs

ANDREW SCHEPEN

CSIRO Land and Water, Brisbane, and James Cook University, Townsville, Australia

YVETTE EVERINGHAM

James Cook University, Townsville, Australia

QUAN J. WANG

University of Melbourne, Melbourne, Australia

(Manuscript received 21 February 2019, in final form 18 September 2019)

ABSTRACT

Multivariate seasonal climate forecasts are increasingly required for quantitative modeling in support of natural resources management and agriculture. GCM forecasts typically require postprocessing to reduce biases and improve reliability; however, current seasonal postprocessing methods often ignore multivariate dependence. In low-dimensional settings, fully parametric methods may sufficiently model intervariable covariance. On the other hand, empirical ensemble reordering techniques can inject desired multivariate dependence in ensembles from template data after univariate postprocessing. To investigate the best approach for seasonal forecasting, this study develops and tests several strategies for calibrating seasonal GCM forecasts of rainfall, minimum temperature, and maximum temperature with intervariable dependence: 1) simultaneous calibration of multiple climate variables using the Bayesian joint probability modeling approach; 2) univariate BJP calibration coupled with an ensemble reordering method (the Schaake shuffle); and 3) transformation-based quantile mapping, which borrows intervariable dependence from the raw forecasts. Applied to Australian seasonal forecasts from the ECMWF System4 model, univariate calibration paired with empirical ensemble reordering performs best in terms of univariate and multivariate forecast verification metrics, including the energy and variogram scores. However, the performance of empirical ensemble reordering using the Schaake shuffle is influenced by the selection of historical data in constructing a dependence template. Direct multivariate calibration is the second-best method, with its far superior performance in in-sample testing vanishing in cross validation, likely because of insufficient data relative to the number of parameters. The continued development of multivariate forecast calibration methods will support the uptake of seasonal climate forecasts in complex application domains such as agriculture and hydrology.

1. Introduction

Seasonal forecasts of climate variables are in high demand around the globe for informing decision-making in climate-sensitive industries and for water resources management. These days, global climate model forecasting systems (GCMs) are widely used for seasonal forecasting, in part, because they generate a detailed global view of the climate state and, in part, because they output a broad spectrum of climate variables of importance to sectors including water management, agriculture, and public health. Many different GCMs have been developed internationally, with differences in component models (i.e., ocean, atmosphere, land surface, and sea ice), data assimilation strategies, ensemble generation schemes, scales, dynamics, and physics; leading to systems with vastly different biases and forecasting skill (e.g., Kim et al. 2012; Pegion et al. 2019). Even at the global scale, GCMs differ to some degree in their characterization of dominant climate

Supplemental information related to this paper is available at the Journals Online website: https://doi.org/10.1175/MWR-D-19-0046.s1.

Corresponding author: Andrew Schepen, andrew.schepen@csiro.au

DOI: 10.1175/MWR-D-19-0046.1

patterns such as ENSO (Barnston and Tippett 2013; Shi et al. 2012). Moreover, at the local scale, GCMs vary in their representations of key climate variables (e.g., rainfall and temperature) and associations with seasonal climate drivers (Kim et al. 2012; Lim et al. 2009; White et al. 2014; Zhao and Hendon 2009). Consequently, individual GCMs present nuanced outlooks around broader climate patterns.

For local decision-making and risk-taking on the basis of GCM forecasts, raw GCM forecasts require statistical postprocessing to rectify model biases, reduce skill deficits and to improve overall reliability (e.g., Feddersen et al. 1999; Gneiting et al. 2005; Weisheimer and Palmer 2014; Zhao et al. 2017). GCM forecast ensemble spread typically is too narrow relative to the true forecast uncertainty and doesn't vary appropriately from one forecast to the next (Barnston et al. 2015; Weisheimer and Palmer 2014). Moreover, where quantitative modeling is to be undertaken using GCM outputs, it is vital that ensemble members have a physically coherent structure across the relevant variables and, depending on the application, in space and time as well. Scheuerer and Hamill (2015) give the perfunctory example of snowmelt in spring being dependent on both rainfall and temperature, suggesting the joint distribution of rainfall and temperature is, therefore, an important consideration. Regression-based calibration and other forms of statistical postprocessing are often only practical to apply to individual locations, time periods and variables (e.g., Doblas-Reyes et al. 2005). More problematically, GCM-modeled relationships between these dimensions are easily lost in postprocessing where random sampling from statistical distributions occurs, requiring reestablishment of covariance structures through nonparametric ensemble reordering techniques such as ensemble copula coupling (Schefzik et al. 2013) or the Schaake shuffle (Clark et al. 2004). For example, Luo and Wood (2008) and Yuan and Wood (2012) injected the spatiotemporal covariance from observations into rainfall and temperature forecasts generated by a Bayesian linear-regression technique to obtain forecasts suitable for use in hydrological applications.

Elsewhere, the Bayesian joint probability modeling approach (BJP; Wang and Robertson 2011; Wang et al. 2009) has been applied to calibrate seasonal GCM forecasts in Australia (Hawthorne et al. 2013; Schepen and Wang 2013), China (Peng et al. 2014) and the United States (Strazzo et al. 2019). Rather than being a typical regression, BJP is designed to model the full joint distribution of any number of predictor and predictand climate variables after allowing for the independent transformation of the marginal distributions (hereafter, marginals). Postprocessed ensemble members are obtained through a sequence of conditional sampling of the posterior distribution, which includes parameter uncertainty, and back-transformation. Various studies have found that BJP produces reliable probabilistic forecasts that capture inherent GCM skill; however, these studies have been limited to a univariate configuration (in the sense of dealing with a single variable). For example, BJP-calibrated seasonal forecasts of rainfall have been subjected to the Schaake shuffle and used to generate reliable long-range ensemble streamflow forecasts. Very little attention appears to have been given to the multivariate calibration of seasonal climate forecasts, which is essential for more complex applications such as agricultural crop-modeling, which requires coherent forecasts of rainfall, temperature and solar radiation.

In contrast to seasonal forecasting, the joint postprocessing of weather variables in short-term (NWP) forecasting has become a topic of increasing interest in recent years. Several studies have investigated the bivariate calibration of the u and v components of wind vectors (McLean Sloughter et al. 2013; Pinson 2012; Schuhen et al. 2012) and the joint calibration of temperature and wind speed forecasts (Baran and Möller 2015, 2017; Schefzik 2016). In particular, Baran and Möller (2015) introduced a Bayesian model averaging methodology and, later (Baran and Möller 2017), an ensemble model output statistics (EMOS) methodology for temperature/wind speed calibration, both relying on a truncated bivariate normal construction. Earlier, Möller et al. (2013) presented a more general methodology that first calibrates the marginals independently, thereafter constructing the intervariable dependence structure using Gaussian copulas. Baran and Möller (2017) concluded that all three aforementioned methods (EMOS, BMA, and copula-reconstruction) yielded similar reliability and accuracy improvements over raw temperature/wind speed forecasts, and, therefore, they advocated for the bivariate EMOS approach for efficiency reasons.

Schefzik (2016) surmised that there are two broad approaches to multivariate postprocessing of weather forecasts. The first is univariate postprocessing followed by nonparametric ensemble reordering methods to establish spatial, temporal and intervariable correlation structures. The second is fully parametric postprocessing, which is usually tailored for low-dimensional settings. Consequently, Schefzik (2016) proposed a hybrid postprocessing approach that jointly postprocesses related variables in low-dimensional settings and thereafter applies an ensemble reordering method with a multivariate ranking to obtain final aggregated, postprocessed forecasts for higher-dimensional spaces (e.g., across different locations or lead times). Similarly to earlier studies,

the focus was on the truncated-bivariate-normal model for temperature and wind speed.

In this study, we investigate the merits of postprocessing multivariate seasonal climate forecasts using several parametric and nonparametric methods. We propose a comparison of 1) directly postprocessing multiple climate variables simultaneously using one BJP model; 2) postprocessing each variable with a univariate BJP model and subsequently restoring the intervariable correlations via the Schaake shuffle; and 3) a quantile-mapping approach as another comparison. It is anticipated that testing these three different strategies will expose the numerous trade-offs that exist between the efficiency and dimensionality of parametric approaches, and the amenity of historical data to fit the parametric model and/or provide realistic covariance structures. While it has been suggested that parametric approaches are quite suitable for low-dimensional forecast calibration problems (Schefzik 2016; Vannitsem et al. 2018), a priori, we do not suspect which approach will perform better for seasonal forecast calibration. Direct multivariate calibration may be challenged by the number of parameters relative to a small number of data points available (typically 20–40 for seasonal postprocessing). Indeed, Doblas-Reyes et al. (2005) found difficulties establishing robust regression coefficients when using multiple regression for combining multiple seasonal forecasts. That said, studies using BJP for hydrology have successfully exploited its ability to model multiple predictands for forecasting streamflow at multiple sites (Wang and Robertson 2011; Wang et al. 2009) and for multiple months ahead (Zhao et al. 2016), situations where the covariances are likely to be well structured.

In this study, we target one-month-lead-time forecasts of seasonal (3-month average) rainfall, minimum temperature, and maximum temperature for Australia. These variables are core products in seasonal forecast services globally. Our remit is restricted to modeling of intervariable correlations—models are developed for each month and grid point individually. Forecast skill and reliability are assessed using ECMWF System4 hindcasts from 1981 to 2016, establishing separate models for each start month from January to December, and with a forecast lead time of 1 month. Forecast skill is quantified as the improvement over a seasonally dependent climatology reference formed from observations. As another comparison for the performance of BJP calibration, we develop a novel version of quantile mapping that is consistent with BJP in terms of modeling the marginals. Quantile mapping adjusts the location and ensemble spread of the GCM forecasts but simply transfers information about intervariable relationships from the raw model output into the observation space; thus, it does not involve a correction based on the correlation between forecasts and observations, but it has the benefit of fewer parameters. Hereafter we present the modeling and verification methods, followed by a continental-scale study, results, discussion, and conclusions.

2. Methods

a. Multivariate calibration strategies

Before getting into the detailed methods, we introduce the three general approaches that are developed and tested in this study for multivariate calibration of Tmin, Tmax, and rainfall:

1) Simultaneous calibration of all climate variables in one BJP model; termed multivariate BJP (MBJP).
2) Independent BJP calibration for each variable followed by restoration of intervariable correlations via the Schaake shuffle ensemble reordering method; termed univariate BJP plus Schaake shuffle (UBJP + SS).
3) Quantile mapping of transformed variables (TQM).

The workflow for each of these three approaches is shown in Fig. 1.

b. Marginal transformation

The three postprocessing methods are constructed with the working assumption that the marginal distributions are able to be modeled as normal distributions after being subjected to variance-stabilizing transformations. The assumption is patently reasonable for variables like temperature, except that the normal distribution has infinite support and, therefore, the tails may not represent extremes precisely. For rainfall, which ostensibly has a mixed discrete-continuous distribution, the way forward is not immediately obvious. Nevertheless, the ability to model its distribution using a transformed-normal is highly desirable because it allows postprocessing of rainfall in the same framework as temperature. The solution adopted here is to treat rainfall data as being left-censored. That is, rainfall data with a value of 0, or some other minimum measurable amount, are assumed to have a true value of less than or equal to that amount, with the precise value unknown. Standard statistical methods are available for the normal distribution and censored data and, therefore, it is possible to use variance-stabilizing transformations for all variables in BJP.

The degree, or the ''strength,'' of the transformation required to achieve normality, depends on several factors including the range, scale, and skewness of the data. We employ two flexible variance-stabilizing transformations

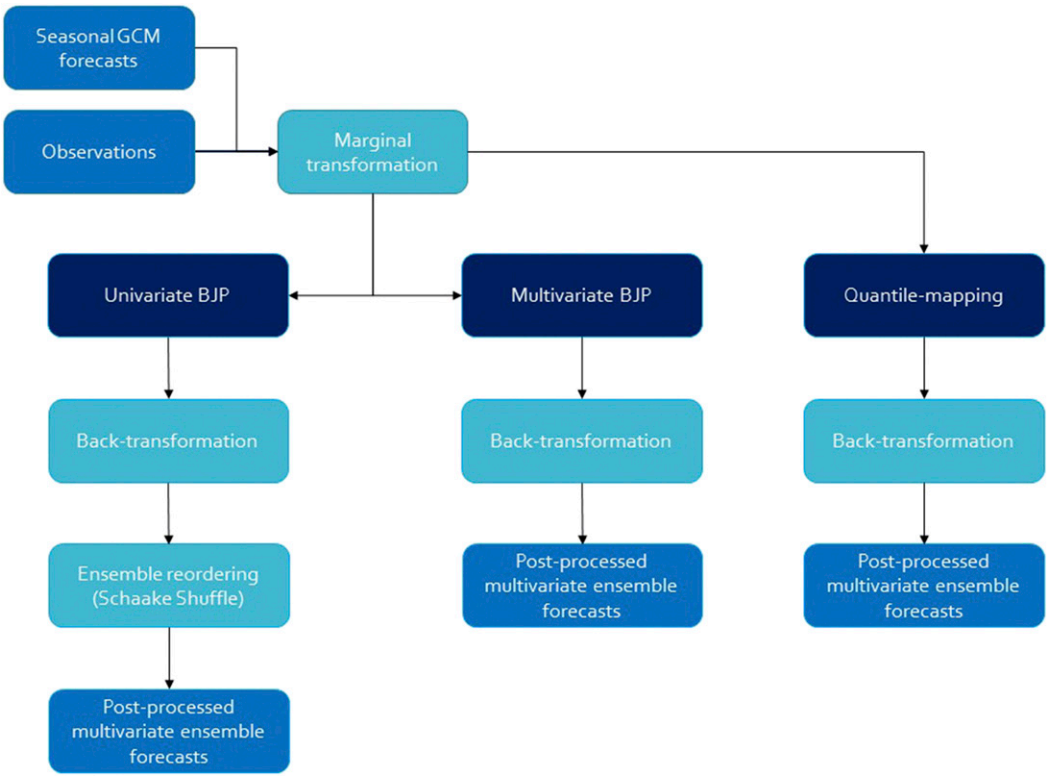

FIG. 1. Schematic of the three different modeling approaches tested for producing calibrated multivariate forecasts of Tmin, Tmax, and rainfall.

in this work. The reason for using two different transformations is because we use the log–sinh transformation (Wang et al. 2012b) for rainfall, which was developed specifically for hydrological variables. Temperature variables use the Yeo–Johnson transformation (Yeo and Johnson 2000). While temperature is often modeled using a normal distribution, which suggests no transformation is required, preliminary investigations revealed statistically significant skewness in temperature distributions in some regions and seasons in Australia (not shown) and, therefore, we allow for transformation if needed. The flexibility of the variance-stabilizing transformations effectively allows for little or no transformation if need be.

Temperature variables are transformed by the single parameter Yeo–Johnson transformation (Yeo and Johnson 2000):

$$\psi_\lambda(y) = \begin{cases} [(y+1)^\lambda - 1]/\lambda & \lambda \neq 0, y \geq 0 \\ \log(y+1) & \lambda = 0, y \geq 0 \\ -[(-y+1)^{2-\lambda} - 1]/(2-\lambda) & \lambda \neq 2, y < 0 \\ -\log(-y+1) & \lambda = 2, y < 0 \end{cases}. \quad (1)$$

The Yeo–Johnson transformation is highly flexible and can be used to transform both positively and negatively skewed data. It incorporates a range of useful transformations, including the log, square root and inverse transformations and embeds the historically popular Box–Cox transformation (Box and Cox 1964). In this study, transformations are established by using Bayesian maximum a posteriori (MAP) estimation of λ for the posterior probability of (λ, μ, σ) where μ and σ are the normal distribution mean and standard deviation parameters. The full details of the Bayesian estimation procedure, including specification of the prior distributions, is given by Schepen et al. (2016).

As mentioned, rainfall is transformed by a two-parameter log–sinh transform (Wang et al. 2012b):

$$\psi_{\varepsilon,\lambda}(y) = \frac{1}{\lambda}\log[\sinh(\varepsilon + \lambda y)], \quad (2)$$

where ε and λ are transformation parameters. The log–sinh transformation was developed to handle the pattern of errors in hydrological predictions. The log–sinh transformation has been widely applied to transform rainfall and streamflow data in statistical modeling of hydrological data (e.g., Bennett et al. 2016; Del Giudice et al. 2013; Robertson et al. 2013). MAP estimation of ε and λ is carried out for the posterior probability of

$(\varepsilon, \lambda, \mu, \sigma^2)$ using the same type of procedure as for the Yeo–Johnson transformation.

c. Multivariate BJP calibration (MBJP)

Multivariate BJP calibration is when several different climate variables are calibrated jointly in the one model, with covariance explicitly modeled. The BJP modeling approach uses a multivariate normal distribution to model the relationship between the transformed predictor and predictand variables (hereafter referred to as predictors and predictands). We note that the predictors and predictands are transformed separately. In this study, BJP predictors are ensemble-mean GCM forecasts and predictands are observations. The collection of d transformed predictors and predictands form the vector $\mathbf{z}^{\mathrm{T}} = [z_1 \quad z_2 \quad \cdots \quad z_d]$. Once the marginals have been transformed using a variance-stabilizing transformation, it is assumed that the joint distribution is multivariate normal:

$$\mathbf{z} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma}), \tag{3}$$

where $\boldsymbol{\mu}$ is the mean vector:

$$\boldsymbol{\mu}^{\mathrm{T}} = [\mu_1 \quad \mu_2 \quad \cdots \quad \mu_d], \tag{4}$$

$\boldsymbol{\Sigma}$ is the covariance matrix:

$$\boldsymbol{\Sigma} = \mathbf{D}(\boldsymbol{\sigma}) \times \mathbf{P} \times \mathbf{D}(\boldsymbol{\sigma}), \tag{5}$$

$\mathbf{D}(\boldsymbol{\sigma})$ is a diagonal matrix from the standard deviation vector:

$$\boldsymbol{\sigma}^{\mathrm{T}} = [\sigma_1 \quad \sigma_2 \quad \cdots \quad \sigma_d], \tag{6}$$

and $\mathbf{P}$ is the symmetric correlation matrix:

$$\mathbf{P} = \begin{bmatrix} 1 & \rho_{1,2} & \cdots & \rho_{1,d} \\ \rho_{2,1} & 1 & \cdots & \rho_{2,d} \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{d,1} & \rho_{d,2} & \cdots & 1 \end{bmatrix}, \tag{7}$$

giving a total of $2d + d(d-1)/2$ parameters in addition to the transformation parameters. Previous descriptions of BJP in the literature detail an inference method based on a Metropolis sampler (Wang and Robertson 2011; Wang et al. 2009). Here, we use a more efficient Gibbs sampler to infer $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ (Wang et al. 2019). The following uninformative prior is specified to complete the Bayesian formulation:

$$p(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \propto |\boldsymbol{\Sigma}|^{-(d+1)/2}. \tag{8}$$

Beyond the description included here, BJP includes treatments to allow inference in the presence of missing values and censored data. These treatments are described by Wang and Robertson (2011) and Wang et al. (2019).

To use BJP as a forecasting tool, the multivariate normal distribution is conditioned on the predictors. For a single set of parameters $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$, consider the transformed predictors $\mathbf{z}_1$ and predictands $\mathbf{z}_2$ organized as

$$\mathbf{z} = \begin{bmatrix} \mathbf{z}_1 \\ \mathbf{z}_2 \end{bmatrix} \tag{9}$$

and the mean vector and covariance matrix correspondingly partitioned as follows:

$$\boldsymbol{\mu} = \begin{bmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{bmatrix}, \tag{10}$$

$$\boldsymbol{\Sigma} = \begin{bmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{bmatrix}. \tag{11}$$

The conditional distribution of the predictands given the predictors is also a multivariate normal distribution:

$$\mathbf{z}_2 | \mathbf{z}_1 \sim N(\boldsymbol{\mu}', \boldsymbol{\Sigma}'), \tag{12}$$

where

$$\boldsymbol{\mu}' = \boldsymbol{\mu}_2 + \boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}[\mathbf{z}_1 - \boldsymbol{\mu}_1], \tag{13}$$

$$\boldsymbol{\Sigma}' = \boldsymbol{\Sigma}_{22} - \boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12}. \tag{14}$$

Forecast values are sampled from the distribution given by Eq. (14) and back transformed to the original space. Gibbs sampling is used to obtain one sample from $\mathbf{z}_2|\mathbf{z}_1$ for M different sets of parameters, thus generating an ensemble of size M that incorporates parameter uncertainty. In this study, $M = 200$.

d. Univariate BJP calibration plus Schaake shuffle (UBJP+SS)

Univariate BJP calibration is when there is only one climate variable under consideration (although there are technically two variables in the model: the BJP predictor and the BJP predictand). To establish coherent multivariate forecasts after applying univariate BJP to each variable, we apply the Schaake shuffle ensemble reordering method (Clark et al. 2004). The Schaake shuffle imposes the rank correlation structure of randomly selected historical observations into forecasts. We describe the essential steps of the procedure here. For a given forecast time period (e.g., month), consider an ensemble forecast of size M denoted by

$$\mathbf{X} = (x_1, x_2, \ldots, x_M), \tag{15}$$

which can be sorted to obtain

$$\boldsymbol{\chi} = [x_{(1)}, x_{(2)}, \ldots, x_{(M)}] \quad x_{(1)} \leq x_{(2)} \cdots \leq x_{(M)}. \tag{16}$$

Consider also a vector of observations from the historical record for the same time period (e.g., the same season in other years), also of size M:

$$\mathbf{Y} = (y_1, y_2, \ldots, y_M), \tag{17}$$

which can be sorted to obtain

$$\boldsymbol{\gamma} = [y_{(1)}, y_{(2)}, \ldots, y_{(M)}] \quad y_{(1)} \leq y_{(2)} \cdots \leq y_{(M)}. \tag{18}$$

Furthermore, let rank be a function that determines the position of a value from $\boldsymbol{\gamma}$ in the original unsorted vector $\mathbf{Y}$. The shuffled forecast ensemble is constructed as

$$\mathbf{X}_{\mathrm{SS}} = (x_{\mathrm{ss},1}, \ldots, x_{\mathrm{ss},M}), \tag{19}$$

where $x_{\mathrm{ss},q} = x_{(n)}$ and $q = \mathrm{rank}[\mathbf{Y}, y_{(n)}]$ $n = 1, \ldots, M$. When $\mathbf{Y}$ is constructed consistently using the same dates for all variables, the Schaake shuffle reconstructs the intervariable correlations.

In this study, because BJP forecasts have 200 ensemble members, two different strategies are applied to acquire $\mathbf{Y}$ of sufficient size. The first strategy is to expand the selection of dates by allowing offsets of −30, −15, 15, and 30 days from the start of the seasonal forecast in addition to dates aligning with the beginning of the forecast. A random sample of 200 dates is taken. Aggregates of daily observations matching the length of the seasonal forecasts are derived accordingly for use in the Schaake shuffle. This strategy is termed the window Schaake shuffle (WSS). The second strategy is to use only dates aligning with the forecast start date. The ensemble is then shuffled in blocks. For example, if there are 40 years of historical data, 200 members are shuffled in 5 blocks, assuming the forecast ensemble members are initially in a random order. This strategy is termed the block Schaake shuffle (BSS).

e. Transformed quantile mapping (TQM)

Quantile mapping is a popular method for bias-correcting climate model outputs in impacts studies. It has no model of covariance. Instead, it relies on the intervariable correlations in the GCM being approximately correct, and, therefore, it isn't a full calibration method (Maraun 2013; Zhao et al. 2017). However, it is a method currently supported by the Australian Bureau of Meteorology and being investigated in agricultural applications of seasonal forecasts (e.g., Brown et al. 2018; Western et al. 2018) and, therefore, it is a useful method for comparison purposes.

Quantile mapping comes in many forms, which boil down to two main types: empirical quantile mapping and parametric quantile mapping. In this study, we develop a new, parametric quantile-mapping methodology using the fitted log–sinh or Yeo–Johnson transformed normal distributions from section 2b to represent the marginal distributions. Hence, we call it transformed quantile mapping (TQM). Accordingly, the TQM and BJP methodologies model the marginals of each variable in an entirely consistent way, meaning that the results of BJP and QM postprocessing are more comparable than if we used another QM implementation. The TQM steps are described in the appendix.

3. Application and verification

a. Study data

We now evaluate the multivariate postprocessing of GCM seasonal forecasts of rainfall, minimum temperature maximum temperature for Australia. These three variables form the basis for seasonal outlooks in Australia and routinely have their predictability assessed (e.g., Hudson et al. 2011; Marshall et al. 2014a; Marshall et al. 2014b). Australia is currently switching to a new GCM and doesn't yet have long hindcasts available for verification and calibration studies. In this study, GCM forecasts are obtained from the ECMWF System4 (Sys4) seasonal forecast system, which has been widely evaluated globally.

Sys4 is a coupled system of ocean, atmosphere and land surface models with sea ice concentration conditionally resampled from climatology. It implements the NEMO (Nucleus for European Modeling of the Ocean) v3.0 ocean model at a 1° resolution in the extratropics. It implements the IFS (Integrated Forecasting System) cycle 36r4 atmospheric model with an approximate horizontal resolution of 80 km. The Hydrology Tiled ECMWF Scheme of Surface Exchanges over Land (H-TESSEL) land surface model is integrated into IFS.

Hindcasts are available from 1981 to 2010 with each model run initialized on the 1st of each month and enduring for 7 months. The hindcast dataset is augmented by an archive of real-time forecasts from 2011 to 2016. In hindcast mode, the ensemble generation scheme outputs 15 ensemble members. In forecast mode, the ensemble size increases to 51. Throughout this study we make use of the first 15 ensemble members for all years. Hindcasts and archived real-time forecasts are treated as equivalent. All members are treated as statistically exchangeable.

Gridded observed data come from the Silo database (Jeffrey et al. 2001). Silo is constructed from Bureau of

Meteorology observational records and has been infilled to create a temporally complete record for all locations. We use the Silo data as the reference observations, noting that the data quality is dependent on the degree of quality control in Silo processing, the amount of processing, and the density and quality of the original observations. Silo data are available on a 0.05° grid. We regrid the Silo observations to match the Sys4 data at 0.75° resolution.

In this study, we choose to focus on three-month-average forecasts, with a lead time of 1 month. These types of forecasts represent a true seasonal outlook beyond the current information available about the weather. BJP models are established separately for 12 overlapping seasons from January–February–March (JFM) to December–January–February (DJF). With this configuration, there are 35 data points available to fit each calibration model at each grid cell.

As a preview to the intervariable relationships in seasonal observations, we calculate the absolute Kendall correlation for all grid cells and months. Between Tmin and Tmax, the median Kendall correlation is 0.34 and the 90th percentile is 0.58. Between Tmax and rainfall (which tend to be negatively correlated), these values are 0.35 and 0.55. For Tmin and rainfall, the result is 0.18 and 0.4. These preliminary results suggest it is prudent to handle intervariable dependencies in seasonal forecast postprocessing of rainfall and temperature.

b. *Univariate and multivariate probabilistic forecast verification*

We first apply univariate bias and reliability scores to check the consistency of forecasts and observations for the individual variables. We then apply two multivariate probabilistic scores to assess the overall skill and performance for all variables. In general, quality seasonal forecasts will have little or no bias, be reliable in terms of ensemble spread and supply skill in excess of a climatological reference forecast. All of these aspects of forecast quality are verified here using a leave-one-year-out cross-validation approach for all postprocessing steps.

Forecast bias is recognized as the long-term mean error between forecasts means and observations. For a single variable, we calculate the percentage bias:

$$\mathrm{PBIAS} = \frac{\sum_{t=1}^{T}(\bar{x}_t - y_t)}{\sum_{t=1}^{T}(y_t)} \times 100(\%), \tag{20}$$

where $\bar{x}_t$ is the forecast ensemble mean for event t, and y_t is the corresponding observation. Positive PBIAS indicates systematic overforecasting whereas negative PBIAS indicates systematic underforecasting.

Reliability is the property of statistical consistency between probabilistic forecasts and observations. A reliable forecasting system will accurately estimate the likelihood of an event. Reliability is checked by analyzing the distribution of probability integral transformations or PIT values (Gneiting et al. 2007). The PIT for a forecast CDF (F_t) and paired observation (y_t) is defined by

$$\pi_t = F_t(y_t). \tag{21}$$

In the case that $y_t = 0$, a pseudoPIT value is sampled from a uniform distribution with a range $[0, \pi_t]$ (Wang and Robertson 2011) and this value then supplants the original π_t. If a forecasting system is reliable and the forecasts are continuous, then the PIT values for a set of forecasts follow a standard uniform distribution. Hence, we quantitate reliability using a score that measures the deviation of the PIT values from the theoretical standard uniform values (Renard et al. 2010):

$$\mathrm{REL_{PIT}} = 1.0 - \frac{2}{T}\sum_{i=1}^{T}\left|\pi_{(i)} - \frac{i}{T+1}\right|, \tag{22}$$

where $\pi_{(i)}$ is the ith ranked PIT value. $\mathrm{REL_{PIT}}$ ranges from 0 (worst reliability) to 1 (perfect reliability). Visualization of $\mathrm{REL_{PIT}}$ and its interpretation in the context of PIT uniform probability plots are given by Renard et al. (2010).

The overall skill and performance evaluation of the multivariate forecasts is done using multivariate scores, namely the energy score (ES; Gneiting and Raftery 2007) and the variogram score (VS; Scheuerer and Hamill 2015). For an M ensemble member forecast for N variables and multivariate observations y:

$$\mathrm{ES} = \frac{1}{M}\sum_{k=1}^{M}\|x_k - y\| - \frac{1}{2M^2}\sum_{k=1}^{M}\sum_{l=k}^{M}\|x_k - x_l\|, \tag{23}$$

where x_k is the forecast for ensemble member k and $||\cdot||$ denotes a Euclidean norm. In a single dimension, the energy score reduces to the widely used continuous ranked probability score (CRPS) for single-variable verification.

The ES is an effective measure for determining the aggregate skill of many individual components; however, it is rather insensitive to the miscalibration of dependencies between components (Scheuerer and Hamill 2015). The VS can be much more sensitive to such miscalibration. Using the same notations as for the ES, the VS based on variograms of order p can be estimated for an ensemble forecast by

$$\mathrm{VS} = \sum_{i=1}^{N}\sum_{j=1}^{N} w_{ij}\left(|y_i - y_j|^p - \frac{1}{M}\sum_{k=1}^{M}|x_{k,i} - x_{k,j}|^p\right)^2, \quad (24)$$

where w_{ij} are weights to promote/demote certain pairs in the calculation of the VS. For example, in the spatial case, it can be used to up-weight proximate pairs and down-weight distant pairs. Here, we set $w_{ij} = 1$ to consider all pairings of variables equally; and $p = 0.5$ as commonly used.

The calculate ES and VS will be calculated for variables with different units, which makes the results more challenging to interpret than, for example, applications to one variable across space and/or time. To make the comparison more meaningful, the variables are made dimensionless before calculating the scores. Rainfall is standardized by dividing by the mean of observations. Temperature variables are standardized by a z-score transform.

For ES and VS we calculate a skill score where S is the average score of the postprocessed forecasts over a set of events and S_{ref} is the average score over the same events for a climatological reference set of forecasts:

$$\text{Skill Score} = \frac{\overline{S_{\mathrm{ref}}} - S}{\overline{S_{\mathrm{ref}}}} \times 100(\%). \quad (25)$$

Reference forecasts are leave-one-year-out observation data for the same period as the forecasts.

4. Results and discussion

a. Bias, reliability, and skill of individual variables

The percentage bias (PBIAS), reliability score ($\mathrm{REL_{PIT}}$), and CRPS skill score metrics are summarized for each variable (Tmin, Tmax, and rainfall), for raw forecasts (RSYS4), and for three sets of postprocessed forecasts (UBJP, MBJP, and TQM) (Fig. 2). Univariate verification results are invariant to ensemble member order; hence, we do not refer to the Schaake shuffle in this section. The summaries plot the proportion of cases where a score value is exceeded and are constructed after pooling the scores for all grid cells and seasons.

Regarding bias (Fig. 2, left column), RSYS4 forecasts are (as expected) biased for all three climate variables: Tmin, Tmax, and rainfall. RSYS4 Tmax forecasts have a propensity to be negatively biased, although the bias magnitude is normally less than 10%. RSYS4 Tmin forecasts can be either positively or negatively biased with magnitudes greater than 10% in approximately 30% of cases. RSYS4 rainfall forecasts are biased positively and negatively in approximately equal measure with magnitudes exceeding 25% not uncommon.

Postprocessing substantially reduces PBIAS for all three climate variables. For Tmin and Tmax, bias is reduced to near zero regardless of the postprocessing method. For rainfall, some biases remain after postprocessing with UBJP and MBJP, which is mainly a problem in very dry grid cells where small absolute biases manifest as a large percentage bias; further discussion is given in section 4c. For UBJP and MBJP, the median bias for rainfall is around 2%–3%, although it can exceed 10%; MBJP performing slightly worse for bias correcting rainfall than UBJP. TQM effectively reduces the bias to near zero in nearly all rainfall cases.

Regarding reliability (Fig. 2, middle column), a gray, dashed, vertical line is plotted at $\mathrm{REL_{PIT}} = 0.9$ as a guiding threshold for highly reliable forecasts. Although the choice is arbitrary, it means that on a PIT uniform probability plot (e.g., Renard et al. 2010; Wang et al. 2009) the points would line up closely along the 1:1 line. RSYS4 forecasts of all three climate variables are frequently unreliable, which is in accordance with the observed biases.

Postprocessing substantially improves the reliability of the forecasts by reducing bias and improving ensemble spread. The UBJP and MBJP forecasts are almost always highly reliable. TQM forecasts are also frequently highly reliable, although they are overall less reliable than the BJP forecasts.

Regarding skill (Fig. 2, right column), a gray, dashed line is plotted at a CRPS skill score value of 0.0 to indicate the skill of the climatology reference forecasts. Skill is positive for the postprocessed forecasts in the majority of cases; however, Tmin and Tmax forecasts are overall more skillful than rainfall forecasts. Out of the different postprocessing models, UBJP produces the most skillful forecasts with the median CRPS skill score being higher than every other model for every climate variable, even if only by a small margin. UBJP skill scores are rarely negative and when they are, they are not worse than about −5% to −10%, which can be attributable to cross-validation effects. The MBJP model produces forecasts that are overall less skillful than UBJP and occasionally negative to about −20%, suggesting overfitting may occur; further investigation is given in section 4c. TQM skill is marginally better than MBJP overall but worse than UBJP; TQM is sometimes seen to produce skill scores that are considerably negative, particularly for Tmin; however, unlike with MBJP, overfitting is unlikely to be the problem. More likely, it is the inability of TQM to return negatively skillful forecasts to climatology.

b. Overall performance of multivariate forecasts

Geographical maps of the energy score (ES) skill scores for the multivariate (Tmin, Tmax, rainfall) forecasts are shown for each season and for the

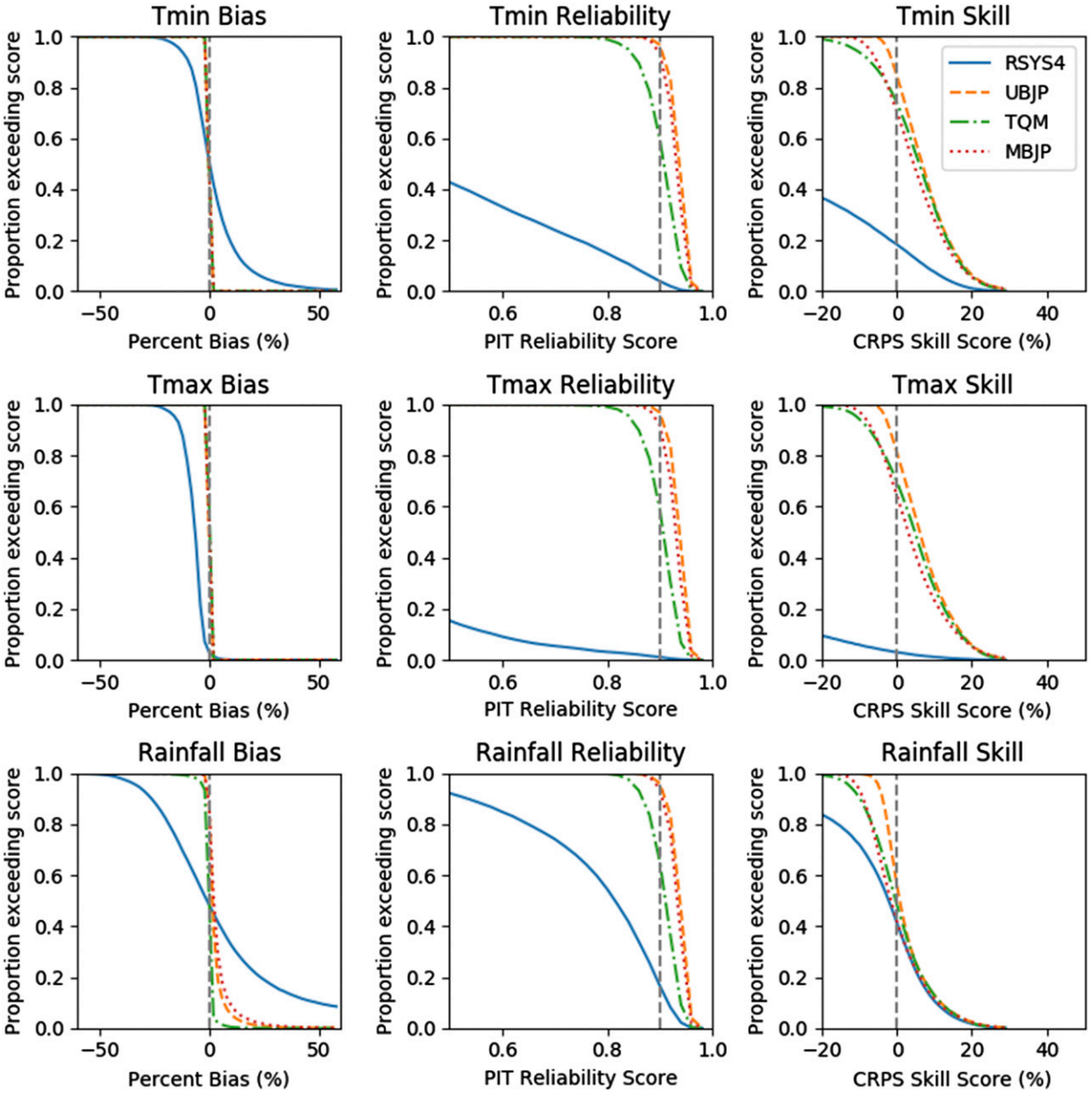

FIG. 2. Plots comparing the overall performance of the various sets of forecasts (raw and postprocessed) as the proportion of grid cells where certain bias, reliability, and skill score values are exceeded. Columns are for the different metrics and rows are for the different climate variables.

UBJP+WSS, MBJP, and TQM postprocessing post-processing method in Figs. 3–5, respectively. Energy score maps for UBJP+BSS are very similar to UBJP+WSS and are not shown. Maps of the variogram score (VS) skill scores for each season are shown for the UBJP + WSS, UBJP+BSS, MBJP, and TQM postprocessing methods in Figs. 6–9, respectively. Summaries of these ES and VS skill scores are shown in the top two panels in Fig. 10.

The ES has not been widely used to make intervariable comparisons. As a first check for the instructiveness of the ES skill score in this setting, we visually compare the ES and CRPS skill score maps (not shown), and we confirm that features of CRPS skill maps for individual variables are noticeable in the ES skill maps and that a sensible conjugation occurs. For example, for UBJP+WSS forecasts, Tmin and Tmax CRPS skill scores are moderately positive across northern Australia, whereas rainfall CRPS skill scores are neutral. The corresponding ES skill scores are weakly to moderately positive. As a second example, for TQM forecasts,

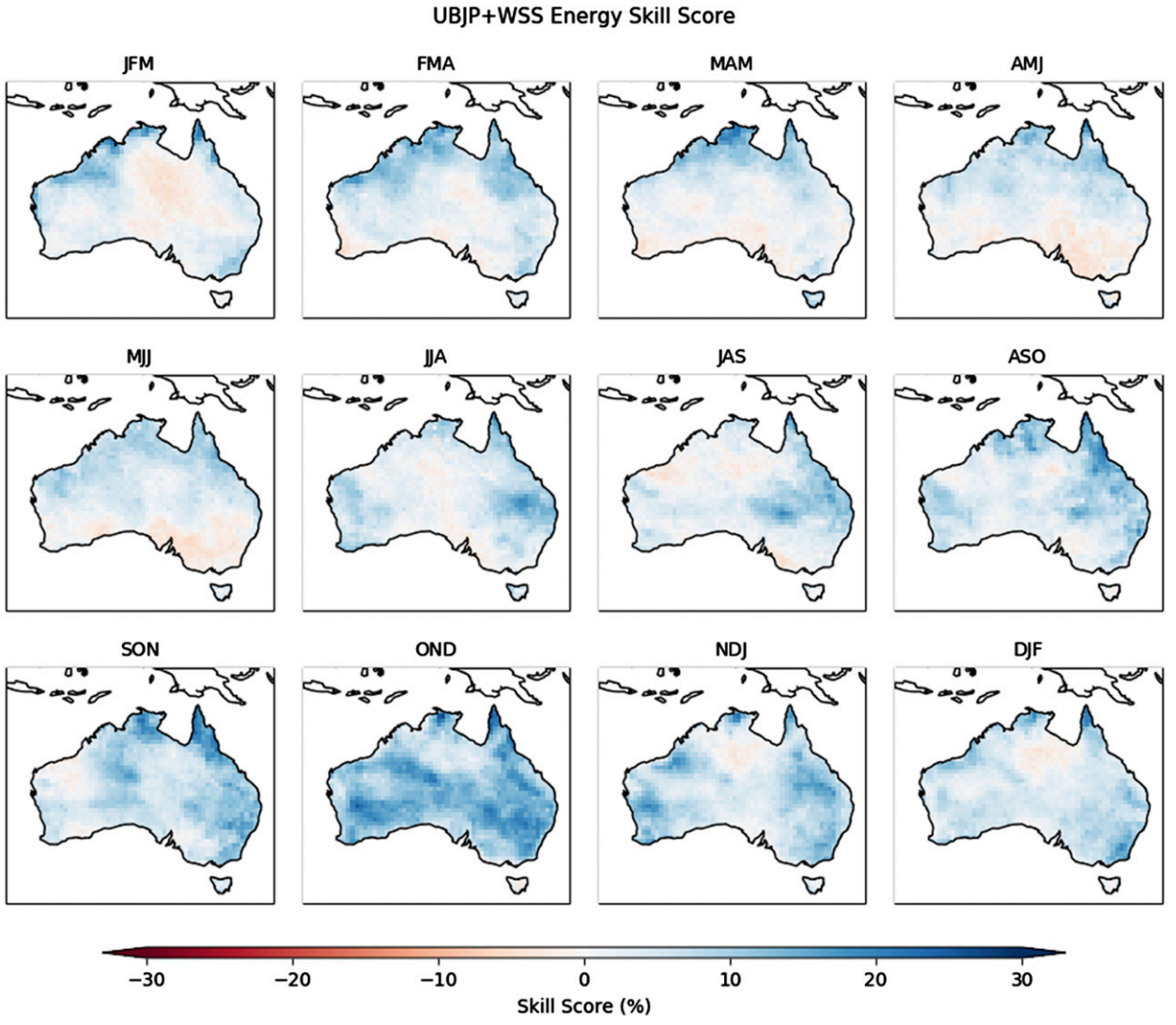

FIG. 3. Maps of energy skill scores for UBJP+WSS forecasts for the period 1981–2016. The skill scores are calculated using historical observation-based climatological reference forecasts and using leave-one-year-out cross validation. Positive skill means lower error in the UBJP+WSS forecasts compared to the reference. The skill is mapped for each target season for forecasts issued with one-month lead time.

all three variables have neutral skill in the southeast of the Australian mainland, a result that translates into the corresponding ES skill score maps.

Overall, ES skill scores are low (<20%), which is understandable given the well-known low–moderate skill of seasonal forecasts, especially with one-month lead time. Moreover, forecasts of Tmin, Tmax, and rainfall are not always similarly skillful across regions and seasons, and ES skill scores are modulated accordingly. In terms of the energy score, UBJP+WSS produces more skillful forecasts than MBJP and TQM, albeit there are broadly similar skill patterns among all three sets of forecasts.

The maps for the VS skill scores give some unique insights. Overall the VS skill scores are lower than the ES skill scores and are more frequently negative. We interpret the VS skill score maps as highlighting areas where there are remaining weaknesses in the intervariable dependence structure in the forecasts. For TQM, the intervariable relationships are largely inherited from the raw model output, and, therefore, it is expected that some regions and seasons will have imperfect intervariable correlations due to model error. Indeed, negative VS skill is observed for TQM forecasts in various regions across all seasons. We expect that either direct modeling of intervariable relationships in MBJP or ensemble reordering UBJP forecasts can deliver more realistic intervariable correlations. However, the results indicate that there are some deficiencies with both BJP

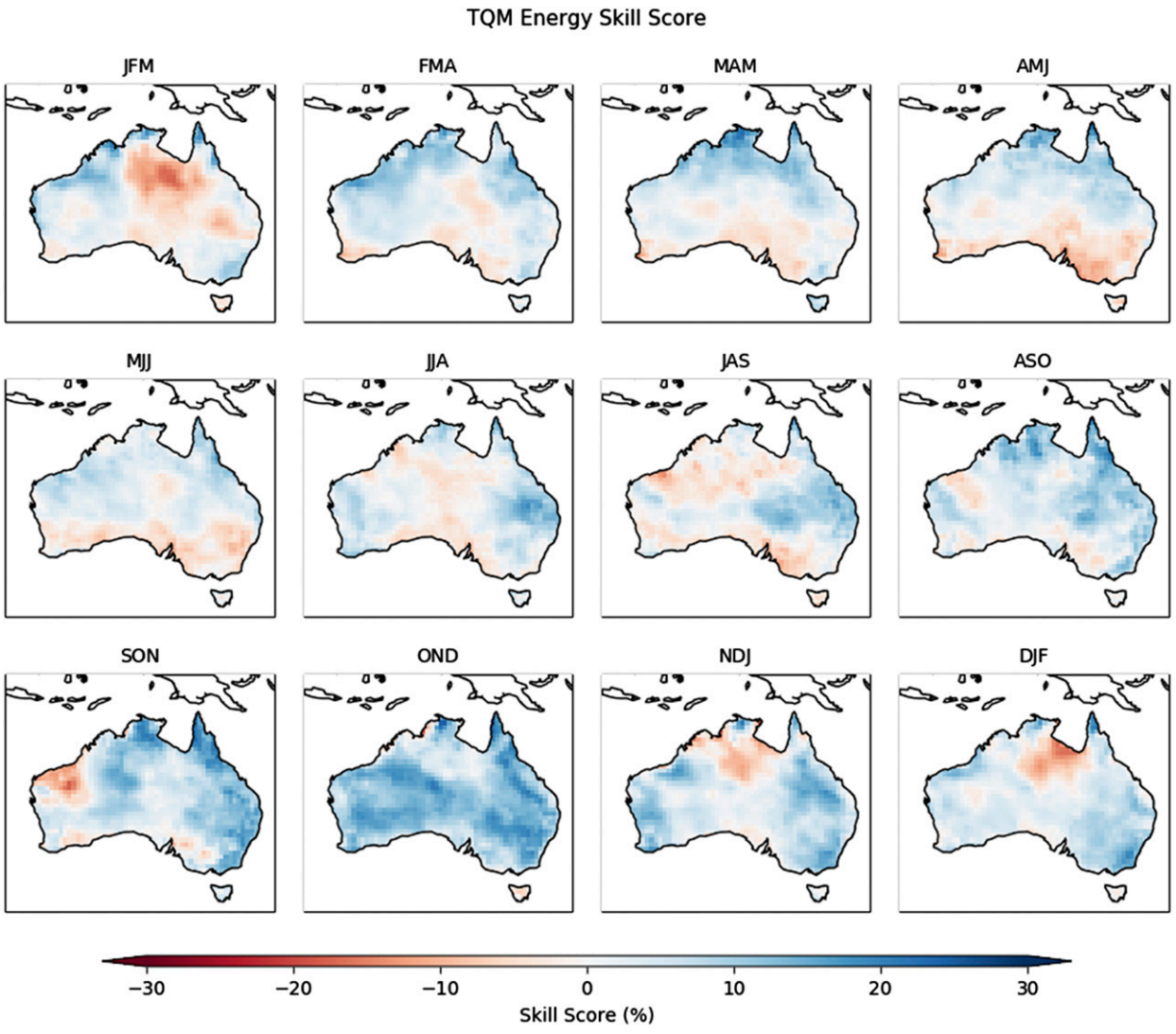

FIG. 4. As in Fig. 3, but for TQM forecasts.

approaches that require further exploration (see section 4c for further discussion).

ES and VS skill score summaries are produced by plotting the proportion of cases where a range of skill score thresholds are exceeded. Results for UBJP+WSS, UBJP+BSS, MBJP, and TQM are shown in the top row of Fig. 10. The skill score summaries support the impression given by comparing the previous skill score maps (Figs. 3–9). That is, the UBJP-WSS and UBJP+BSS forecasts exhibit the best overall performance in terms of the energy score, particularly by having fewer low or negative skill scores. MBJP and TQM perform similarly in terms of the energy score, although MBJP has marginally better performance in terms of filtering out negative skill. In terms of the variogram score, the performance of MBJP and UBJP+WSS is similar, with TQM performing overall worse, and UBJP+BSS presenting the best results. The results for the VS skill scores suggest that the calibration methods that model or enforce observed correlation structures perform better overall; however, there are factors that affect the performance of the parametric and nonparametric modeling components.

The VS skill maps for UBJP+WSS show widespread negative skill in MAM and AMJ, which is largely rectified in the the UBJP+BSS skill maps. The plausible explanation is that the construction of the Schaake shuffle dependence template using a wider window of dates is suboptimal in some regions and seasons compared to repeated use of dates more aligned with the forecast period. Certainly, the Schaake shuffle is beneficial, as skill scores calculated for UBJP forecasts

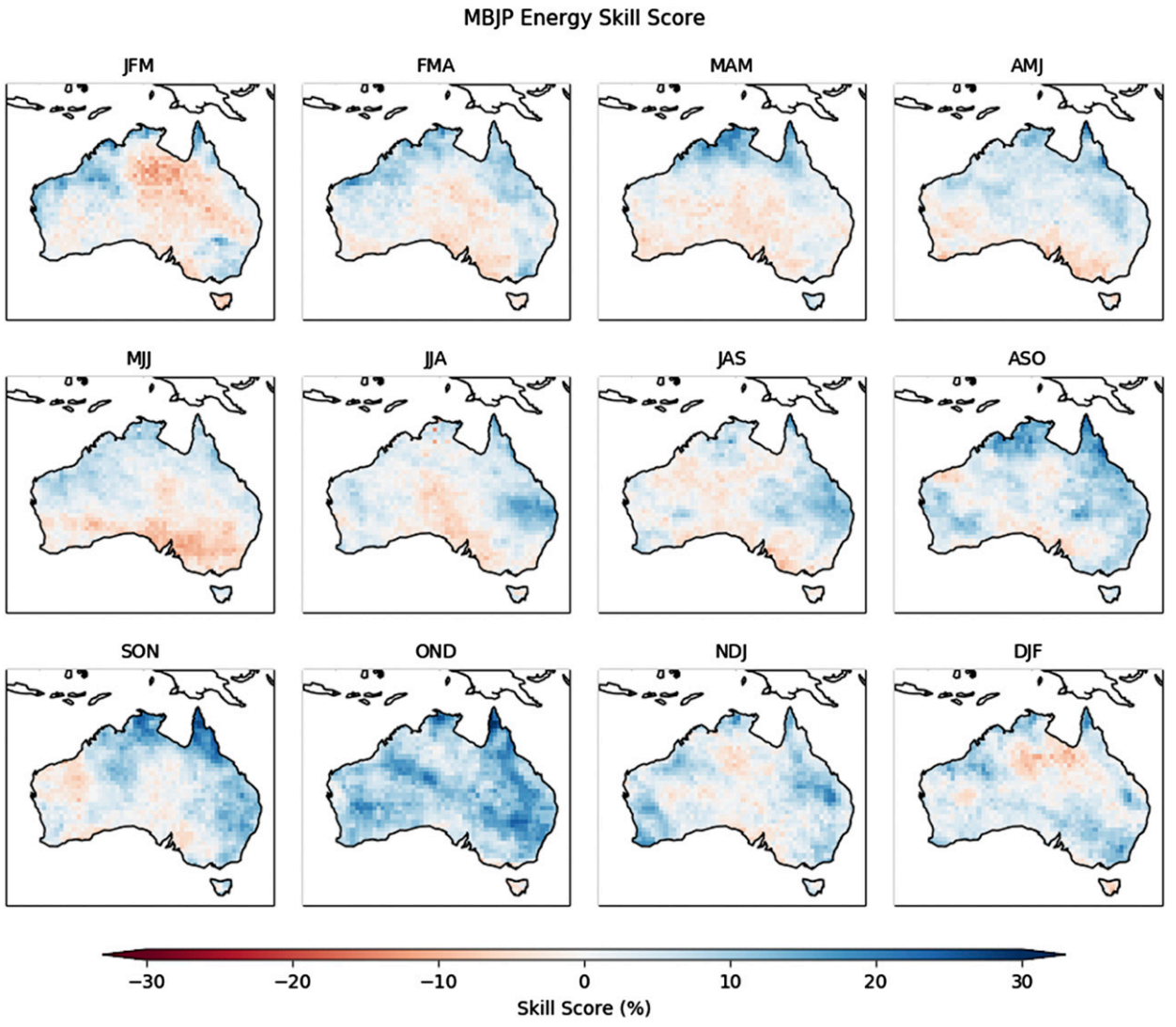

FIG. 5. As in Fig. 3, but for MBJP forecasts.

without Schaake shuffling (i.e., with random ensemble ordering) show a marked decrease in performance (not shown).

The benefit of the Schaake shuffle can also be evaluated in terms of its ability to improve the TQM forecasts. To test this idea, we run an additional experiment whereby TQM forecasts are Schaake shuffled using forecast dates aligned with the start of the forecast. Block resampling is not required since the number of ensemble members is less than the data available, so we call the combination TQM+SS. The evaluation of TQM+SS forecasts is in the middle row of Fig. 10. Similar to previous results, the Schaake shuffle provides limited benefit in terms of energy score evaluation. However, there is a marked improvement in the variogram score, suggesting that the Schaake shuffle with observations can improve upon the TQM intervariable correlations in many instances. Nevertheless, TQM+SS is unable to outperform UBJP+BSS overall. This is because quantile mapping has more serious shortcomings as a forecast calibration method (Zhao et al. 2017) that cannot be overcome by ensemble reordering.

The worse overall performance of MBJP relative to UBJP+WSS and UBJP+BSS could be surprising, except that the forecast verification is being done within a cross-validation framework and MBJP is known to have more parameters (see section 2c); therefore, overfitting is a real risk. To test whether overfitting is indeed a problem causing lower performance of MBJP forecasts, we repeat several of the forecast calibration and verification experiments without applying cross validation.

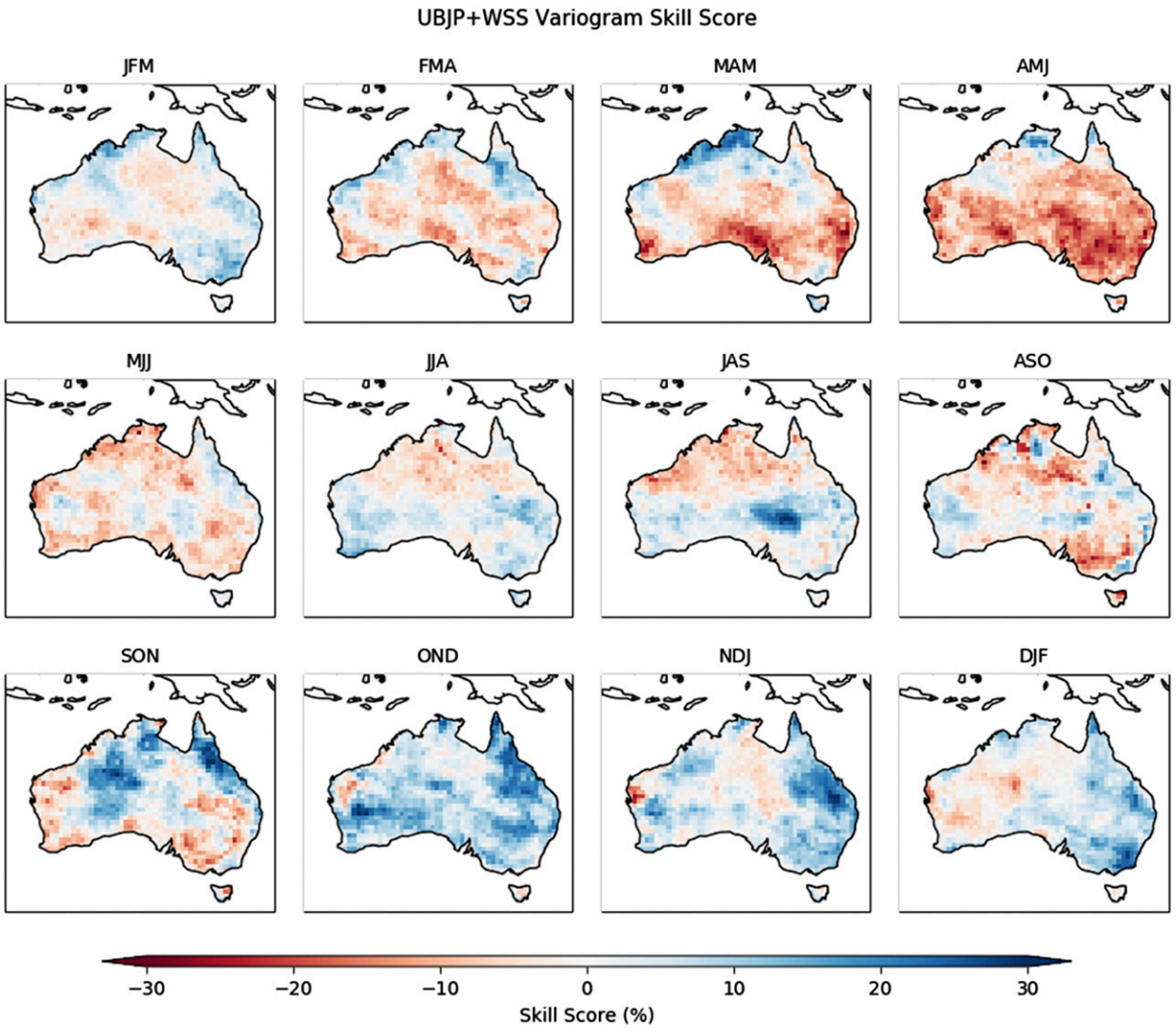

FIG. 6. Maps of variogram skill scores for UBJP+WSS forecasts for the period 1981–2016. The skill scores are calculated using historical observation-based climatological reference forecasts and using leave-one-year-out cross validation. Positive skill means lower error in the UBJP+WSS forecasts compared to the reference. The skill is mapped for each target season for forecasts issued with one-month lead time.

The ES and VS skill score summaries for all grid cells are reproduced for the no cross-validation (no xv) experiments and compared with the originals (bottom row, Fig. 10). We refer to these results as in-sample results whereas the original results are out-of-sample. It is clear that UBJP+BSS and MBJP provide better in-sample than out-of-sample predictive performance, although this boost in predictive performance can be attributed artificial skill. It is also seen that MBJP moves from being inferior to UBJP+BSS to superior to it. This result hints that more sophisticated calibration approaches could be beneficial where sufficient data exists. However, it appears in the current study that there is insufficient data to robustly infer the MBJP model parameters and realize a predictive performance benefit over UBJP+BSS for calibrating independent (out-of-sample) forecasts.

Figure 2 shows that positive biases in the range of 5%–10% can sometimes arise in UBJP and MBJP rainfall forecasts. Tmin and Tmax forecasts are unaffected. Mapping of the seasonal and spatial distribution of the biases in UBJP forecasts (Fig. S1 in the online supplemental material) reveals that these biases are by-and-large contained to very dry grid cells, particularly in northern Australia during the seasons MJJ–JAS when monthly rainfall totals are mostly near zero. In such cases, a small absolute bias can manifest as a large percentage bias. Moreover, BJP adds parameter uncertainty,

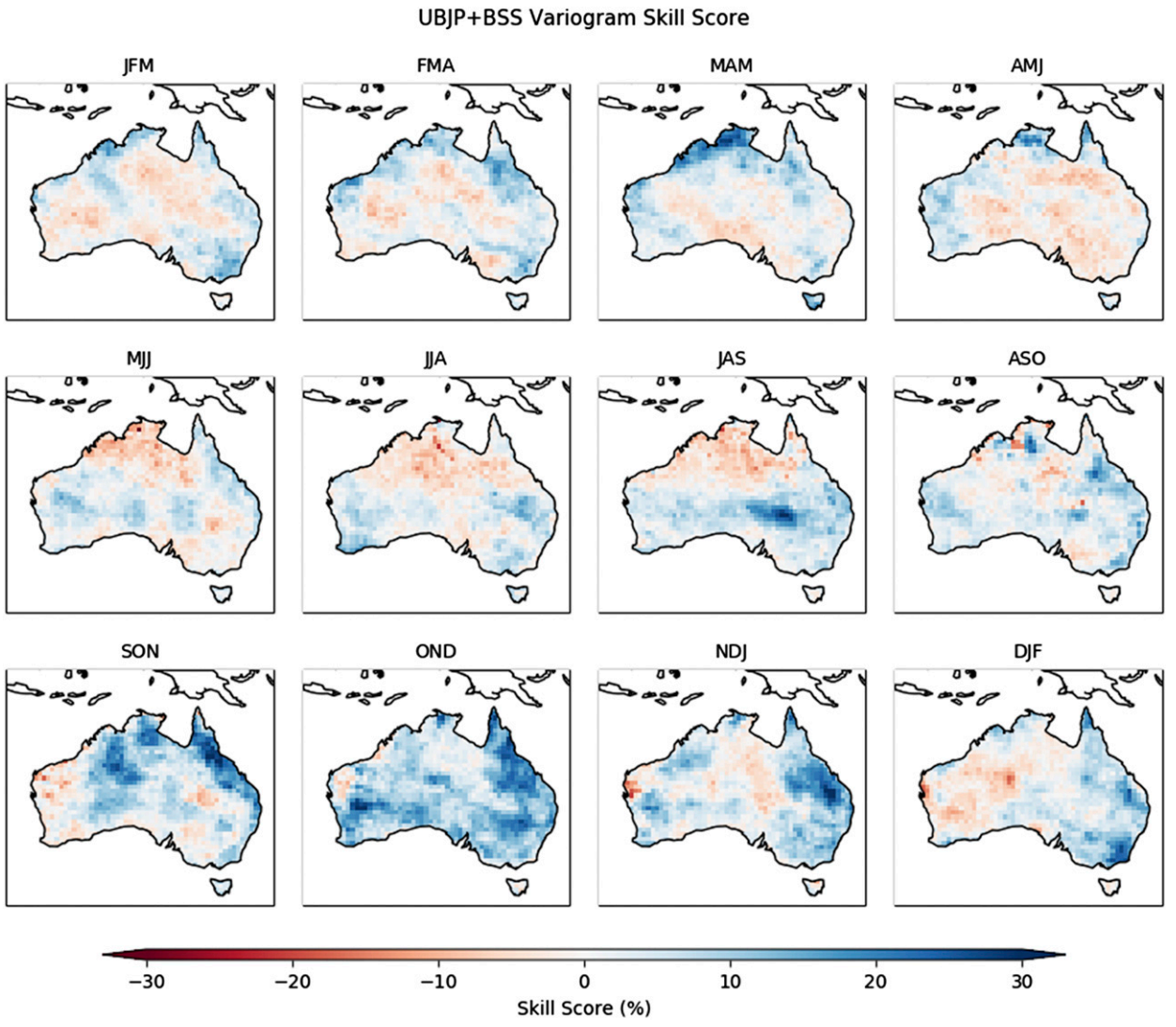

FIG. 7. As in Fig. 6, but for UBJP+BSS forecasts.

which we suspect can lead to some extreme values being generated in the back-transformation procedure, causing noticeably higher means in very dry grid cells. Although not shown in these results, we find that BJP models fitted to observed data generate samples with the same biases, so it is not strictly a problem related to the calibration of GCM forecasts, but rather to do with the challenges of modeling highly skewed distributions.

c. *Extension opportunities*

In this study we only considered postprocessing of variables at the local scale. An alternative approach that remains untested, which may add skill while reducing overfitting, is to set up single predictor–multiple predictand models where the predictor represents a relevant large-scale climate feature (i.e., an ENSO climate index). Furthermore, multiple forecasts may be combined using Bayesian model averaging or another combination method to improve skill in different regions and seasons (e.g., Schepen et al. 2014; Wang et al. 2012a).

The results show that flexible modeling of Tmin, Tmax, and rainfall marginal distributions permits multivariate postprocessing using joint probability models and alternative implementations of extant methods like quantile mapping. While we used the flexible Yeo–Johnson transformation and the hydrologically specific log–sinh transformation, any appropriate normalizing transformation could be substituted into the workflows (e.g., a Box–Cox transformation). We expect that the strategies employed here could be tested more widely, including to other variables including pressure, wind speed, solar radiation and evaporation. A broader understanding of

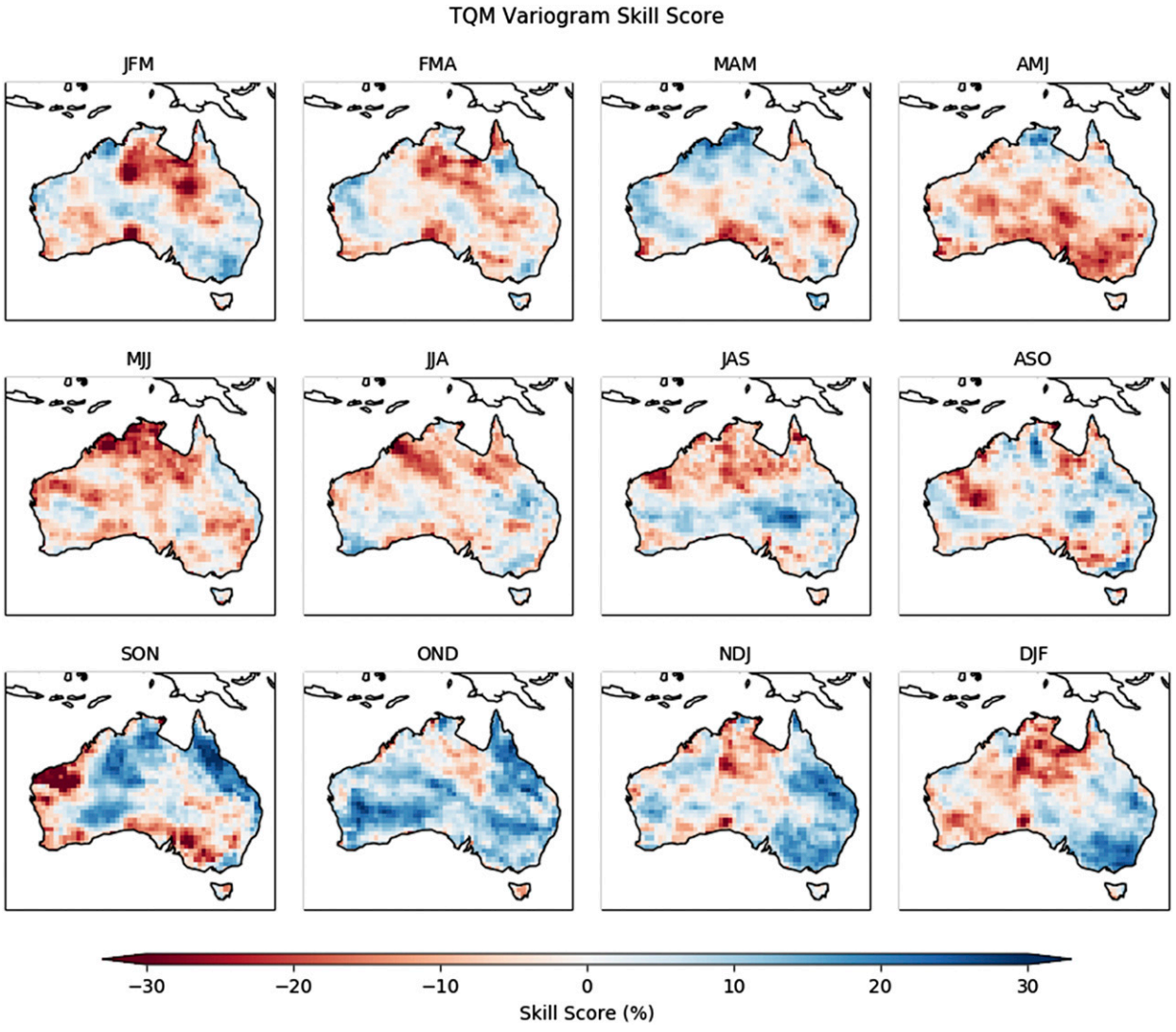

FIG. 8. As in Fig. 6, but for TQM forecasts.

multivariate forecasting skill can benefit applications beyond agriculture and natural resources management, including in energy, mining and insurance.

It was found that the choice of the unconditional Schaake shuffle using a window of starting dates led to subpar forecast performance in terms of the variogram score, which can be related to the imperfect modeling of intervariable correlations. Scheuerer et al. (2017) detected improved results after applying a variation of the Schaake shuffle in which the dependence template was constructed by the preferential selection of dates such that the chosen sequences were more representative of the forecast distribution. Such a method could improve the results of UBJP+WSS in certain seasons and bring the results closer to or improve upon UBJP+BSS. As an aside, Scheuerer et al. (2017) also remarked on the enhanced possibility of variogram skill scores being negative compared to the energy score due to it offering less reward for correctly predicting magnitude, a feature that we see in these results. Other studies have highlighted the partial ineffectiveness of the Schaake shuffle (Verkade et al. 2013) or proposed selective variants that yield improvements. For example, Bellier et al. (2017) evaluated analog-based methods for selecting Schaake shuffle dates and found it outperformed the unconditional Schaake shuffle for short-term rainfall forecasts, especially in impact on subsequent streamflow forecasts. Wu et al. (2018) point out how ties in data ranks can impact the effectiveness of rank reordering schemes, which will be pertinent in daily or subdaily studies; however, we expect it would only have a very minor impact in this seasonal study (e.g., multiple zeros in

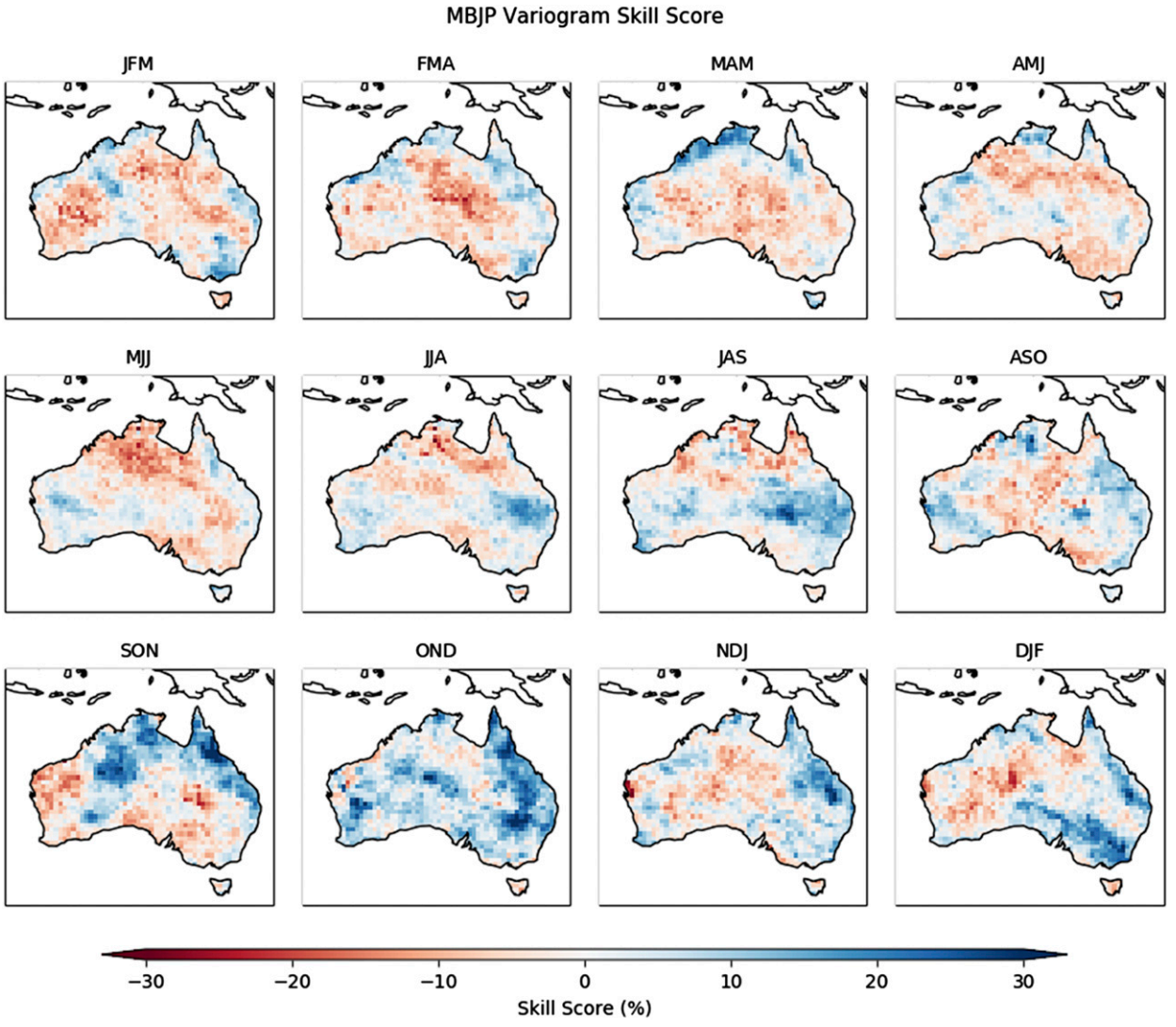

FIG. 9. As in Fig. 6, but for MBJP forecasts.

rainfall records may occur in exceptionally dry areas). Evidence is building around the shortcomings in ensemble reordering methods and thus further work is needed to identify the most efficient and effective options to use these to restore multivariate dependence structures.

Overall, the results in this study point to plenty of challenges to address in integrating robust low-dimensional postprocessing approaches in high-dimensional application domains (e.g., multiple variables, subcatchments, lead times, and so forth). There may be gains made by alternative avenues, such as by establishing models of covariance that require fewer parameters, particularly in combination with other dimension-reduction techniques. For the foreseeable future, both parametric calibration and empirical ensemble reordering methods are going to play a role in seasonal forecast postprocessing, while much more research is needed to find balanced solutions that improve multivariate forecasting skill for independent predictions.

In this study, we have addressed only seasonal (three-month) forecasts. However, many operational models that could receive climate forecast information (e.g., hydrological and biophysical models) require data at daily time steps and at subgrid locations. More research is needed to spatially and temporally downscale multivariate seasonal climate forecasts.

d. Conclusions

GCM forecasts are increasingly in demand to support the expansion of natural resource management initiatives, which require coherent multivariate seasonal climate forecasts. Raw GCM forecasts are readily

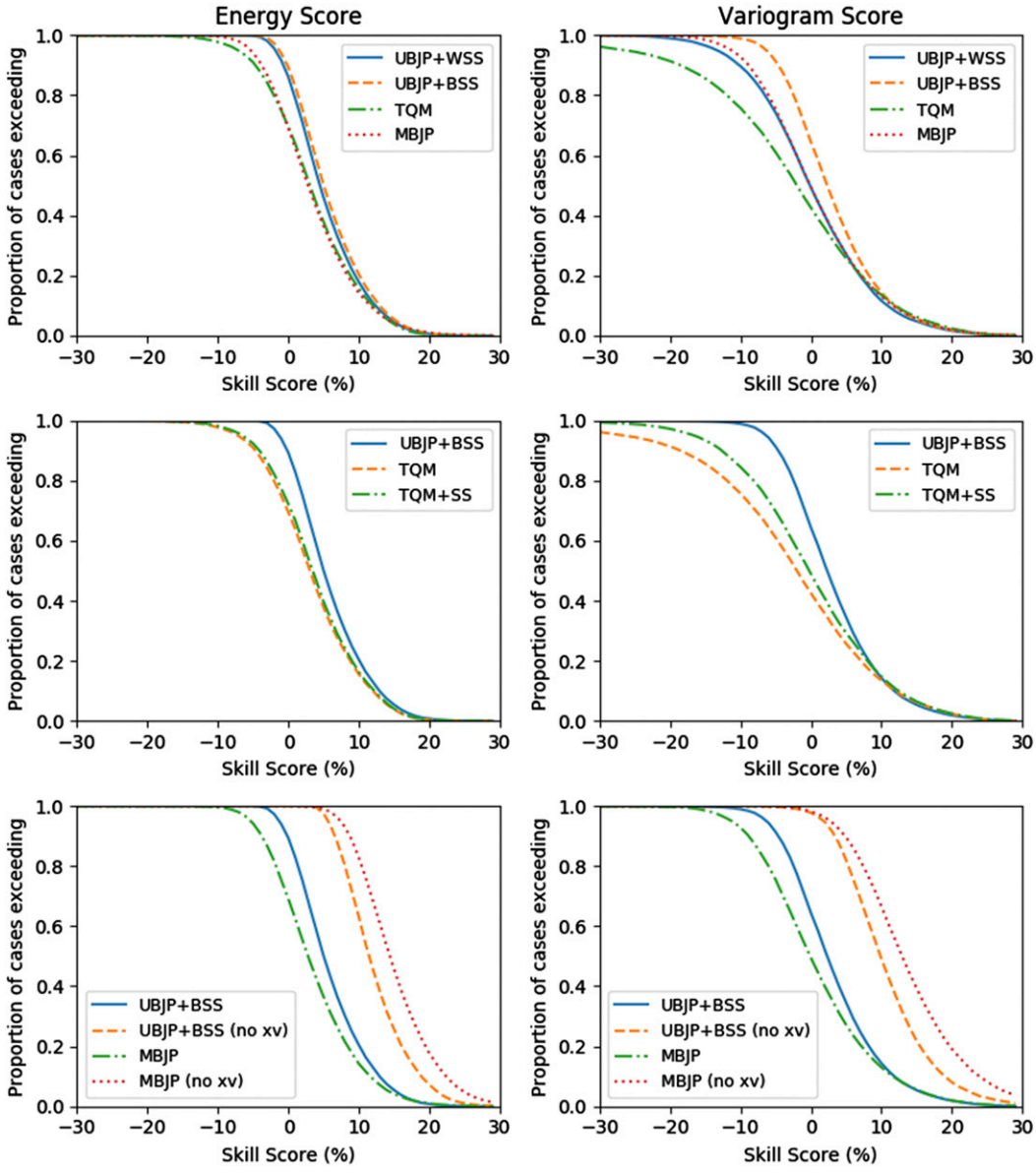

FIG. 10. Summary of multivariate forecast performance across all grid cells and seasons and a comparison of the results for various postprocessing methods. The curves plot the proportion of cases where ES and VS skill score values are exceeded. The multivariate skill scores consider all three climate variables (Tmin, Tmax, and rainfall) in their calculation. The VS is more sensitive to the calibration of the dependencies between the variables. (top) Comparison of the core postprocessing methods; (middle) additional analysis evaluating the benefit of applying the Schaake shuffle to TQM forecasts; and (bottom) additional analysis testing the effect cross validation has on forecast performance.

available but they require calibration to remove biases and reliably quantify forecast uncertainty. While multivariate postprocessing has been considered previously in the very specific problem of short-term temperature and wind speed forecasting, very little attention has been paid to the multivariate calibration of seasonal GCM outputs. Usually, any bias correction or calibration in seasonal forecasting is done on variables independently. In this study, we develop and test three strategies for calibrating multivariate forecasts of Tmin, Tmax, and rainfall, finding each approach has unique strengths and weaknesses.

UBJP+WSS and UBJP+BSS apply a univariate BJP calibration to each variable and subsequently establishes the intervariable correlation structure from observations using the Schaake shuffle. The UBJP+BSS approach performs best in terms of univariate skill and reliability scores and multivariate skill scores. This provides evidence that the unconditional sampling of

historical trajectories for the Schaake shuffle is suboptimal in some instances, especially when the template data are not representative of the forecast period.

MBJP simultaneously calibrates each variable by modeling the full joint distribution of all relevant predictor and predictand variables. In in-sample testing MBJP presents itself as the far superior approach; however, in cross validation with out-of-sample testing, MBJP generally performs worse than UBJP+BSS, apparently due to the lack of sufficient data to robustly infer the more numerous model parameters. That said, MBJP may remain feasible for problems with more data available.

TQM is a quantile-mapping approach that uses the same marginal transformations as BJP. We find that while it offers substantial improvements over raw forecasts and has fewer parameters, its fundamental weakness of not modeling correlations between forecasts and observations or between variables means that it performs overall the worst in terms of univariate and multivariate verification metrics. Ensemble reordering is unable to improve TQM forecasts enough to outperform the BJP-based approaches.

Continued research efforts are likely to optimize the calibration of seasonal forecasts for complex application domains requiring multivariate climate inputs. We suggest that further research should investigate the robust modeling of covariances, dimension-reduction techniques, and resolution of emerging challenges in ensemble reordering techniques (including handling ties and more efficient construction of conditional dependence templates).

Acknowledgments. We thank the Queensland government and the Australian Bureau of Meteorology for the Silo meteorological data used in this study. We thank the European Centre for Medium-Range Weather Forecasts for the System4 seasonal forecast data used in this study. We are appreciative of the thoughtful discussions had with Dr. David Robertson regarding multivariate calibration and ensemble reordering. Our manuscript was much improved thanks to feedback on an early draft by Dr. Ming Li followed by reviews and suggestions from three anonymous peer reviewers.

APPENDIX

Transformed Quantile Mapping (TQM)

TQM is described as follows in two parts:

1) Model the marginal distributions of the forecasts and observations
   (i) Collect all the historical forecast ensemble members.
   (ii) Fit a transformed-normal distribution to the forecasts using either the log–sinh or Yeo–Johnson transformation. Save the estimated normal distribution parameters μ_F and σ_F and the transformation τ_F.
   (iii) Collect all the observations corresponding to the forecasts from step (i). There will be fewer observation data points than forecast data points because the forecasts are ensembles.
   (iv) Fit a transformed-normal distribution to the observations using either the log–sinh or Yeo–Johnson transformation. Save the estimated normal distribution parameters μ_O and σ_O and the transformation τ_O.
2) Postprocess a new ensemble forecast
   (i) Transform the ith ensemble member $y_{F,i}$ to $z_{F,i} = \tau_F(y_{F,i})$.
   (ii) Convert $z_{F,i}$ to a dimensionless z score: $z^*_{F,i} = (z_{F,i} - \mu_F)/\sigma_F$.
   (iii) Invert $z^*_{F,i}$ using μ_O and σ_O to get $z_{O,i} = (z^*_{F,i} \times \sigma_O) + \mu_O$.
   (iv) Back transform $z_{O,i}$ to $y_{O,i} = \tau_O^{-1}(z_{O,i})$.
   (v) Repeat steps (i)–(iv) for all ensemble members, $k = 1, \ldots, M$.

The procedure is a fully parametric implementation of quantile mapping. It differs substantially from any other implementation in the literature because it makes use of the log–sinh and Yeo–Johnson transformations that are used with BJP. In addition, the new method handles the mixed discrete-continuous nature of variables like rainfall using a censored data approach, which is quite different to the more common split-model approach, whereby intensity and frequency are modeled using separate distributions (e.g., Volosciuk et al. 2017).

REFERENCES

Baran, S., and A. Möller, 2015: Joint probabilistic forecasting of wind speed and temperature using Bayesian model averaging. *Environmetrics*, **26**, 120–132, https://doi.org/10.1002/env.2316.

——, and ——, 2017: Bivariate ensemble model output statistics approach for joint forecasting of wind speed and temperature. *Meteor. Atmos. Phys.*, **129**, 99–112, https://doi.org/10.1007/s00703-016-0467-8.

Barnston, A. G., and M. K. Tippett, 2013: Predictions of Nino3. 4 SST in CFSv1 and CFSv2: A diagnostic comparison. *Climate Dyn.*, **41**, 1615–1633, https://doi.org/10.1007/s00382-013-1845-2.

——, ——, H. M. van den Dool, and D. A. Unger, 2015: Toward an improved multimodel ENSO prediction. *J. Appl. Meteor. Climatol.*, **54**, 1579–1595, https://doi.org/10.1175/JAMC-D-14-0188.1.

Bellier, J., G. Bontron, and I. Zin, 2017: Using meteorological analogues for reordering postprocessed precipitation ensembles in hydrological forecasting. *Water Resour. Res.*, **53**, 10 085–10 107, https://doi.org/10.1002/2017WR021245.

Bennett, J. C., 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, https://doi.org/10.1002/2016WR019193.

Box, G. E., and D. R. Cox, 1964: An analysis of transformations. *J. Roy. Stat. Soc.*, **26B**, 211–252, https://doi.org/10.1111/j.2517-6161.1964.tb00553.x.

Brown, J. N., Z. Hochman, D. Holzworth, and H. Horan, 2018: Seasonal climate forecasts provide more definitive and accurate crop yield predictions. *Agric. For. Meteor.*, **260–261**, 247–254, https://doi.org/10.1016/j.agrformet.2018.06.001.

Clark, M., S. Gangopadhyay, L. Hay, B. Rajagopalan, and R. Wilby, 2004: The Schaake shuffle: A method for reconstructing space–time variability in forecasted precipitation and temperature fields. *J. Hydrometeor.*, **5**, 243–262, https://doi.org/10.1175/1525-7541(2004)005<0243:TSSAMF>2.0.CO;2.

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, https://doi.org/10.3402/tellusa.v57i3.14658.

Feddersen, H., A. Navarra, and M. N. Ward, 1999: Reduction of model systematic error by statistical correction for dynamical seasonal predictions. *J. Climate*, **12**, 1974–1989, https://doi.org/10.1175/1520-0442(1999)012<1974:ROMSEB>2.0.CO;2.

Del Giudice, D., M. Honti, A. Scheidegger, C. Albert, P. Reichert, and J. Rieckermann, 2013: Improving uncertainty estimation in urban hydrological modeling by statistically describing bias. *Hydrol. Earth Syst. Sci.*, **17**, 4209–4225, https://doi.org/10.5194/hess-17-4209-2013.

Gneiting, T., and A. E. Raftery, 2007: Strictly proper scoring rules, prediction, and estimation. *J. Amer. Stat. Assoc.*, **102**, 359–378, https://doi.org/10.1198/016214506000001437.

——, ——, 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, https://doi.org/10.1175/MWR2904.1.

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

Hawthorne, S., Q. 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, https://doi.org/10.1002/wrcr.20453.

Hudson, D., O. Alves, H. H. Hendon, and A. G. Marshall, 2011: Bridging the gap between weather and seasonal forecasting: Intraseasonal forecasting for Australia. *Quart. J. Roy. Meteor. Soc.*, **137**, 673–689, https://doi.org/10.1002/qj.769.

Jeffrey, S. J., J. O. Carter, K. B. Moodie, and A. R. Beswick, 2001: Using spatial interpolation to construct a comprehensive archive of Australian climate data. *Environ. Modell. Software*, **16**, 309–330, https://doi.org/10.1016/S1364-8152(01)00008-1.

Kim, H.-M., P. J. Webster, and J. A. Curry, 2012: Seasonal prediction skill of ECMWF System 4 and NCEP CFSv2 retrospective forecast for the Northern Hemisphere Winter. *Climate Dyn.*, **39**, 2957–2973, https://doi.org/10.1007/s00382-012-1364-6.

Lim, E.-P., H. H. Hendon, D. Hudson, G. Wang, and O. Alves, 2009: Dynamical forecast of inter–El Niño variations of tropical SST and Australian spring rainfall. *Mon. Wea. Rev.*, **137**, 3796–3810, https://doi.org/10.1175/2009MWR2904.1.

Luo, L., and E. F. Wood, 2008: Use of Bayesian merging techniques in a multimodel seasonal hydrologic ensemble prediction system for the eastern United States. *J. Hydrometeor.*, **9**, 866–884, https://doi.org/10.1175/2008JHM980.1.

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

Marshall, A., D. Hudson, H. Hendon, M. Pook, O. Alves, and M. Wheeler, 2014a: Simulation and prediction of blocking in the Australian region and its influence on intra-seasonal rainfall in POAMA-2. *Climate Dyn.*, **42**, 3271–3288, https://doi.org/10.1007/s00382-013-1974-7.

——, ——, M. Wheeler, O. Alves, H. Hendon, M. Pook, and J. Risbey, 2014b: Intra-seasonal drivers of extreme heat over Australia in observations and POAMA-2. *Climate Dyn.*, **43**, 1915–1937, https://doi.org/10.1007/s00382-013-2016-1.

McLean Sloughter, J., T. Gneiting, and A. E. Raftery, 2013: Probabilistic wind vector forecasting using ensembles and Bayesian model averaging. *Mon. Wea. Rev.*, **141**, 2107–2119, https://doi.org/10.1175/MWR-D-12-00002.1.

Möller, A., A. Lenkoski, and T. L. Thorarinsdottir, 2013: Multivariate probabilistic forecasting using ensemble Bayesian model averaging and copulas. *Quart. J. Roy. Meteor. Soc.*, **139**, 982–991, https://doi.org/10.1002/qj.2009.

Pegion, K., T. DelSole, E. Becker, and T. Cicerone, 2019: Assessing the fidelity of predictability estimates. *Climate Dyn.*, **53**, 7251–7265, https://doi.org/10.1007/s00382-017-3903-7.

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

Pinson, P., 2012: Adaptive calibration of (u, v)-wind ensemble forecasts. *Quart. J. Roy. Meteor. Soc.*, **138**, 1273–1284, https://doi.org/10.1002/qj.1873.

Renard, B., D. Kavetski, G. Kuczera, M. Thyer, and S. W. Franks, 2010: Understanding predictive uncertainty in hydrologic modeling: The challenge of identifying input and structural errors. *Water Resour. Res.*, **46**, W05521, https://doi.org/10.1029/2009WR008328.

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, https://doi.org/10.5194/hess-17-3587-2013.

Schefzik, R., 2016: Combining parametric low-dimensional ensemble postprocessing with reordering methods. *Quart. J. Roy. Meteor. Soc.*, **142**, 2463–2477, https://doi.org/10.1002/qj.2839.

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

Schepen, A., and Q. Wang, 2013: Toward accurate and reliable forecasts of Australian seasonal rainfall by calibrating and merging multiple coupled GCMS. *Mon. Wea. Rev.*, **141**, 4554–4563, https://doi.org/10.1175/MWR-D-12-00253.1.

——, ——, and D. E. Robertson, 2014: Seasonal forecasts of Australian rainfall through calibration and bridging of coupled GCM outputs. *Mon. Wea. Rev.*, **142**, 1758–1770, https://doi.org/10.1175/MWR-D-13-00248.1.

——, ——, and Y. Everingham, 2016: Calibration, bridging, and merging to improve GCM seasonal temperature forecasts in Australia. *Mon. Wea. Rev.*, **144**, 2421–2441, https://doi.org/10.1175/MWR-D-15-0384.1.

Scheuerer, M., and T. M. Hamill, 2015: Variogram-based proper scoring rules for probabilistic forecasts of multivariate

quantities. *Mon. Wea. Rev.*, **143**, 1321–1334, https://doi.org/10.1175/MWR-D-14-00269.1.

——, ——, B. Whitin, M. He, and A. Henkel, 2017: A method for preferential selection of dates in the Schaake shuffle approach to constructing spatiotemporal forecast fields of temperature and precipitation. *Water Resour. Res.*, **53**, 3029–3046, https://doi.org/10.1002/2016WR020133.

Schuhen, N., T. L. Thorarinsdottir, and T. Gneiting, 2012: Ensemble model output statistics for wind vectors. *Mon. Wea. Rev.*, **140**, 3204–3219, https://doi.org/10.1175/MWR-D-12-00028.1.

Shi, L., H. H. Hendon, O. Alves, J.-J. Luo, M. Balmaseda, and D. Anderson, 2012: How predictable is the Indian Ocean dipole? *Mon. Wea. Rev.*, **140**, 3867–3884, https://doi.org/10.1175/MWR-D-12-00001.1.

Strazzo, S., D. C. Collins, A. Schepen, Q. J. Wang, E. Becker, and L. Jia, 2019: Application of a hybrid statistical–dynamical system to seasonal prediction of North American temperature and precipitation. *Mon. Wea. Rev.*, **147**, 607–625, https://doi.org/10.1175/MWR-D-18-0156.1.

Vannitsem, S., D. S. Wilks, and J. Messner, 2018: *Statistical Postprocessing of Ensemble Forecasts*. Elsevier Science, 362 pp.

Verkade, J. S., J. D. Brown, P. Reggiani, and A. H. Weerts, 2013: Post-processing ECMWF precipitation and temperature ensemble reforecasts for operational hydrologic forecasting at various spatial scales. *J. Hydrol.*, **501**, 73–91, https://doi.org/10.1016/j.jhydrol.2013.07.039.

Volosciuk, C. D., D. Maraun, M. Vrac, and M. Widmann, 2017: A combined statistical bias correction and stochastic downscaling method for precipitation. *Hydrol. Earth Syst. Sci.*, **21**, 1693–1719, https://doi.org/10.5194/hess-21-1693-2017.

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

——, ——, and F. Chiew, 2009: A Bayesian joint probability modeling approach for seasonal forecasting of streamflows at multiple sites. *Water Resour. Res.*, **45**, W05407, https://doi.org/10.1029/2008WR007355.

——, A. Schepen, and D. E. Robertson, 2012a: Merging seasonal rainfall forecasts from multiple statistical models through Bayesian model averaging. *J. Climate*, **25**, 5524–5537, https://doi.org/10.1175/JCLI-D-11-00386.1.

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

——, Y. Shao, Y. Song, A. Schepen, D. E. Robertson, D. Ryu, and F. Pappenberger, 2019: An evaluation of ECMWF SEAS5 seasonal climate forecasts for Australia using a new forecast calibration algorithm. *Environ. Modell. Software*, **122**, 104550, https://doi.org/10.1016/j.envsoft.2019.104550.

Weisheimer, A., and T. Palmer, 2014: On the reliability of seasonal climate forecasts. *J. Roy. Soc. Interface*, **11**, 20131162, https://doi.org/10.1098/rsif.2013.1162.

Western, A. W., K. B. Dassanayake, K. C. Perera, R. M. Argent, O. Alves, G. Young, and D. Ryu, 2018: An evaluation of a methodology for seasonal soil water forecasting for Australian dry land cropping systems. *Agric. For. Meteor.*, **253–254**, 161–175, https://doi.org/10.1016/j.agrformet.2018.02.012.

White, C. J., D. Hudson, and O. Alves, 2014: ENSO, the IOD and the intraseasonal prediction of heat extremes across Australia using POAMA-2. *Climate Dyn.*, **43**, 1791–1810, https://doi.org/10.1007/s00382-013-2007-2.

Wu, L., Y. Zhang, T. Adams, H. Lee, Y. Liu, and J. Schaake, 2018: Comparative evaluation of three Schaake Shuffle schemes in postprocessing GEFS precipitation ensemble forecasts. *J. Hydrometeor.*, **19**, 575–598, https://doi.org/10.1175/JHM-D-17-0054.1.

Yeo, I. K., and R. A. Johnson, 2000: A new family of power transformations to improve normality or symmetry. *Biometrika*, **87**, 954–959, https://doi.org/10.1093/biomet/87.4.954.

Yuan, X., and E. F. Wood, 2012: Downscaling precipitation or bias-correcting streamflow? Some implications for coupled general circulation model (CGCM)-based ensemble seasonal hydrologic forecast. *Water Resour. Res.*, **48**, W12519, https://doi.org/10.1029/2012WR012256.

Zhao, M., and H. H. Hendon, 2009: Representation and prediction of the Indian Ocean dipole in the POAMA seasonal forecast model. *Quart. J. Roy. Meteor. Soc.*, **135**, 337–352, https://doi.org/10.1002/qj.370.

Zhao, T., A. Schepen, and Q. Wang, 2016: Ensemble forecasting of sub-seasonal to seasonal streamflow by a Bayesian joint probability modelling approach. *J. Hydrol.*, **541**, 839–849, https://doi.org/10.1016/j.jhydrol.2016.07.040.

——, J. Bennett, Q. J. Wang, A. Schepen, A. Wood, D. Robertson, and M.-H. Ramos, 2017: How suitable is quantile mapping for postprocessing GCM precipitation forecasts? *J. Climate*, **30**, 3185–3196, https://doi.org/10.1175/JCLI-D-16-0652.1.