Seasonal Forecasts of Australian Rainfall through Calibration and Bridging of Coupled GCM Outputs

ANDREW SCHEPEN

Bureau of Meteorology, Brisbane, Australia

Q. J. WANG AND DAVID E. ROBERTSON

CSIRO Land and Water, Highett, Australia

(Manuscript received 2 August 2013, in final form 25 November 2013)

ABSTRACT

Coupled general circulation models (GCMs) are increasingly being used to forecast seasonal rainfall, but forecast skill is still low for many regions. GCM forecasts suffer from systematic biases, and forecast probabilities derived from ensemble members are often statistically unreliable. Hence, it is necessary to postprocess GCM forecasts to improve skill and statistical reliability. In this study, the authors compare three methods of statistically postprocessing GCM output—calibration, bridging, and a combination of calibration and bridging—as ways to treat these problems and make use of multiple GCM outputs to increase the skill of Australian seasonal rainfall forecasts. Three calibration models are established using ensemble mean rainfall from three variants of the Predictive Ocean Atmosphere Model for Australia (POAMA) version M2.4 as predictors. Six bridging models are established using POAMA forecasts of seasonal climate indices as predictors. The calibration and bridging forecasts are merged through Bayesian model averaging. Forecast attributes including skill, sharpness, and reliability are assessed through a rigorous leave-three-years-out cross-validation procedure for forecasts of 1-month lead time. While there are overlaps in skill, there are regions and seasons where the calibration or bridging forecasts are uniquely skillful. The calibration forecasts are more skillful for January–March (JFM) to June–August (JJA). The bridging forecasts are more skillful for July–September (JAS) to December–February (DJF). Merging calibration and bridging forecasts retains, and in some seasons expands, the spatial coverage of positive skill achieved by the better of the calibration forecasts and bridging forecasts individually. The statistically postprocessed forecasts show improved reliability compared to the raw forecasts.

1. Introduction

Organizations around the world continue to invest heavily in the development of physically based models for seasonal climate forecasting. The Australian Bureau of Meteorology and the Commonwealth Scientific and Industrial Research Organization (CSIRO) continue to develop the Predictive Ocean Atmosphere Model for Australia (POAMA), a coupled ocean–atmosphere general circulation model (GCM) used for seasonal forecasting. Coupled GCMs are generally skilled at simulating and forecasting large-scale oceanic and atmospheric climate features, such as the El Niño–Southern Oscillation (ENSO; e.g., Guilyardi et al. 2003; Jin et al. 2008; Palmer et al. 2004; Roeckner et al. 1996) and, at short lead times, the Indian Ocean dipole (Zhao and Hendon 2009). GCMs are advocated for their ability to capture the nonlinear interactions of the ocean and atmosphere and thus make spatially coherent forecasts of the climate at high temporal resolution.

Although GCMs are skilled at simulating large-scale climate features, there can be a disconnection with regional climate features such as rainfall. The rain field is greatly variable in both space and time and is influenced by factors at a much smaller scale than the typical GCM resolution (100–300 km). GCM forecasts are thus susceptible to bias arising from systematic errors, and ensemble forecasts can be statistically unreliable in terms of probabilities (Graham et al. 2005; Lim et al. 2011). At

Denotes Open Access content.

Corresponding author address: Andrew Schepen, Bureau of Meteorology, GPO Box 413, Brisbane, Queensland 4001, Australia.
E-mail: a.schepen@bom.gov.au

DOI: 10.1175/MWR-D-13-00248.1

the root of these problems are necessary parameterizations of subgrid processes, meaning the problems can only be overcome through fundamental improvements in the model physics along with increased resolution of the GCM.

Seasonal rainfall forecast users demand forecasts that are skillful, statistically reliable, and free of bias. One approach to improve the accuracy and reliability of GCM forecasts is to statistically postprocess GCM output fields (Feddersen et al. 1999). There are two general approaches taken. The first, known as statistical calibration, establishes a statistical model that relates raw GCM forecasts to observations. The model can then be used to calibrate new forecasts. This approach is potentially useful in regions and seasons where the raw GCM rainfall forecasts have underlying skill. The second approach, referred to here as bridging, establishes a statistical model to make rainfall forecasts by using GCM forecasts of large-scale climate features (indices) as predictors. This approach is potentially useful when there is an established relationship between the large-scale feature and regional rainfall and when the large-scale feature is well forecast by the GCM.

The techniques applied in calibration and/or bridging of GCM rainfall as described above are typically similar to those used in statistical downscaling, in which GCM variables are related to finescale or station rainfall through a statistical model. Zorita and von Storch (1999) and Wilby et al. (1998) summarize a variety of methods. In previous calibration and bridging studies, singular value decomposition analysis (SVDA) or canonical correlation analysis was commonly used to identify predictors for use in multiple linear regression equations (Bartman et al. 2003; Feddersen et al. 1999; Landman and Goddard 2002; Lim et al. 2011).

In Australia, Lim et al. (2011) investigated calibration and bridging for forecasting Australian September–November (SON) rainfall using version 1.5 of POAMA at a lead time of 0. For the bridging component of the study, the predictors were the leading modes of southern extratropical mean sea level pressure (MSLP) derived through SVDA. Raw POAMA 1.5 rainfall was the predictor field in calibration. The study concluded that neither bridging nor calibration significantly improved the skill of SON rainfall forecasts across the continent due to moderate skill in predicting MSLP and limited hindcast data to establish the statistical models. However, simple pooling of raw, bridging, and calibration forecasts improved skill and reliability. Lim et al. (2011) suggested that more sophisticated ensemble methods (e.g., Bayesian) of combining bridging and calibration models could improve skill further. Our view is that Bayesian methods that allow for parameter uncertainty are more likely to produce more reliable forecasts, and, more importantly, skill improvements could be attained through a better choice of predictors in bridging.

In Australia, both concurrent and lagged relationships between climate indices and monthly or seasonal rainfalls are well established (Risbey et al. 2009; Schepen et al. 2012b). Schepen et al. (2012a) found evidence that GCM skill can be augmented by using rainfall predictions from lagged observed climate indices. Possibly, a better approach to bridging is to build statistical models using, as predictors, forecasts of climate indices produced by coupled GCMs such as POAMA. Standard climate indices representing large-scale oceanic and atmospheric circulations, such as Niño-3.4, are readily calculable from POAMA fields. If the GCM forecasts of such indices are more skillful than persistence of the observed indices, then feasibly this will translate into improved rainfall forecast skill. In early exploratory work, Langford et al. (2011) mapped correlations of Australian seasonal rainfall with climate indices forecast from POAMA 1.5 and 2.4. In many seasons and locations, observed rainfall was more correlated with climate indices than with raw POAMA rainfall forecasts. Possibly, further skill improvements can be realized if the predictors are allowed to vary for different regions and seasons to reflect the different influences on regional climate.

In this study, we extend the work of Lim et al. (2011), Langford et al. (2011), and Schepen et al. (2012a). We evaluate the merits of calibration and bridging of GCM outputs within a Bayesian multimodel framework that allows for regional and seasonal variation in the forecasting model. We test the approach for forecasts of Australian seasonal rainfall at a lead time of 1 month. A Bayesian joint probability (BJP) modeling approach (Wang and Robertson 2011; Wang et al. 2009) is employed to establish multiple calibration and bridging models, using POAMA forecasts of rainfall and sea surface temperature anomalies as predictors. Bayesian model averaging (BMA; Hoeting et al. 1999; Raftery et al. 2005; Wang et al. 2012b) is then applied to weight and merge forecasts from the multiple models. In this setup, we can merge the forecasts from many models. It is, therefore, important that we apply stringent validation and verification strategies in our assessment of the BJP–BMA approach. A common approach to verification of seasonal rainfall forecasts is leave-*n*-years-out cross validation. Usually only one year is left out. Here, we apply a rigorous leave-three-years-out cross validation to ensure that the model parameters and weights are independently inferred.

The remainder of this paper is structured as follows. In section 2, we describe the POAMA coupled GCM and

TABLE 1. Description of climate indices used as predictors of Australian seasonal rainfall.

Climate index	Description
Niño-3	Average sea surface temperature anomaly over 5°N–5°S, 150°–90°W
Niño-3.4	Average sea surface temperature anomaly over 5°N–5°S, 170°–120°W
Niño-4	Average sea surface temperature anomaly over 5°N–5°S, 160°E–150°W
ENSO Modoki index (EMI; Ashok et al. 2007)	C − 0.5 (E + W) Where the components are average sea surface temperature anomalies over C: 10°N–10°S, 165°E–140°W E: 5°N–15°S, 110°–70°W W: 20°N–10°S, 125°–145°E
Indian Ocean dipole mode index (DMI; Saji et al. 1999)	WPI − EPI Where the components are average sea surface temperature anomalies over WPI: 10°N–10°S, 50°–70°E EPI: 0°–10°S, 90°–110°E
Indonesia index (II; Verdon and Franks 2005)	Average sea surface temperature anomaly over 0°–10°S, 120°–130°E

the data used. In section 3, we describe the formulation of the calibration and bridging models using BJP, give an overview of the BMA approach, and outline methods for assessing the skill and reliability of probabilistic forecasts. In section 4, we present and discuss maps and diagrams comparing the skill and reliability of the cross-validation forecasts. Section 5 contains additional discussion points, and section 6 wraps up the paper with the main conclusions.

2. Data

a. *Description of the POAMA 2.4 GCM*

The raw forecast ensembles used in this study are extracted from the POAMA version commonly known as M2.4. POAMA M2.4 is a coupled ocean–atmosphere GCM designed to produce intraseasonal-to-seasonal climate forecasts for Australia. Since mid-2013, POAMA M2.4 is the operational seasonal forecasting model of the Australian Bureau of Meteorology. POAMA 2.4 comprises atmospheric and ocean modules, an ocean–atmosphere coupler, an atmosphere and land initialization scheme, and an ocean initialization scheme. The ocean model is the Australian Community Ocean Model version 2 (Oke et al. 2005; Schiller et al. 2002), initialized using the POAMA Ensemble Data Assimilation (PEODAS; Yin et al. 2011). The ocean model has a resolution of 2° longitude by 0.5° latitude at the equator, changing to 2° longitude by 1.5° latitude near the poles. The ocean depth is represented using 25 vertical layers. The atmospheric model is the Bureau of Meteorology Atmospheric Model version 3 (Colman et al. 2005), initialized by the Atmosphere Land Initialisation (ALI) scheme (Hudson et al. 2011). The atmospheric model has a horizontal resolution of T47 (approximately 2.5° by 2.5°) and 17 vertical levels.

Currently there are three variants of POAMA M2.4 in operation. For this reason, POAMA is sometimes referred to as a pseudo multimodel ensemble. POAMA M2.4a and M2.4b differ from POAMA M2.4c in that they use a newer atmospheric model with improved physics associated with shallow convection. POAMA M2.4b differs from M2.4a by the inclusion of a flux correction scheme that reduces biases in model climatology that appear as the forecast lead time extends. Each variant of POAMA M2.4 produces an 11-member forecast ensemble. The ensemble is generated by jointly perturbing the initial conditions of the ocean and atmosphere modules (Marshall et al. 2011a).

b. *POAMA 2.4 forecasts of rainfall and climate indices*

The set of available POAMA hindcasts are initialized on the first day of each month from January 1980 to December 2011. In this study, we analyze forecasts with 1-month lead time. The corresponding three-consecutive-months–seasons to be analyzed are February–April (FMA) 1980 through January–March (JFM) 2012, giving 32 data points for each season of the year.

Raw M2.4 seasonal rainfall forecasts are calculated by aggregating M2.4 monthly rainfall forecasts. We make use of the ensemble mean rainfall from each of the variants (a, b, and c). M2.4 seasonal climate index forecasts are derived from monthly M2.4 SST forecasts. Multiple indices are used to capture variability in the Pacific Ocean associated with ENSO: Niño-3, Niño-3.4, Niño-4, and the El Niño Modoki index (EMI). Other climate indices capture variable in the Indian Ocean region: the Indian Ocean dipole mode index (DMI) and the Indonesian SST index. Table 1 contains details on

the calculation of the six climate indices. Because the three M2.4 variants produce very similar forecasts of the climate indices, we use the ensemble mean of all models (33 members) in calculating the climate indices. This also helps to limit the number of models.

c. Observed rainfall data

Observed rainfall data are used for model parameter inference and forecast verification. The observed seasonal rainfall totals are calculated by aggregating monthly totals from Australian Water Availability Project's (AWAP) 0.05° × 0.05° gridded dataset of monthly rainfall (Jones et al. 2009). The AWAP data are upscaled to approximately 2.5° resolution by averaging within POAMA grid cells. To match the M2.4 forecasts, the period of observed rainfall data used are February 1980–March 2013.

3. Methods

a. Statistical calibration and bridging models

Statistical calibration and bridging models are established for each season, grid cell and lead time independently of the others. Each model has a single predictor and a single predictand. A Bayesian joint probability (BJP) modeling approach (Wang et al. 2009; Wang and Robertson 2011) is used to establish the individual models. The relationship between a predictor x and predictand y can be modeled as following a transformed bivariate normal distribution.

In the case of calibration models, x represents M2.4 forecast ensemble mean rainfall and y represents observed rainfall. To satisfy modeling assumptions, transformations are necessary for data normalization and variance stabilization. The variables are transformed using log–sinh transformations (Wang et al. 2012a). More specifically, x and y follow the transformations:

$$\hat{x} = \frac{1}{\beta_x}\ln[\sinh(\alpha_x + \beta_x x)], \tag{1}$$

$$\hat{y} = \frac{1}{\beta_y}\ln[\sinh(\alpha_y + \beta_y y)]. \tag{2}$$

In the case of bridging models, x represents a climate index and can take negative values. Equation (1) is no longer a suitable transformation to use. Instead, we apply the Yeo–Johnson transformation (Yeo and Johnson 2000):

$$\hat{x} = \begin{cases} [(x+1)^{\lambda_x} - 1]/\lambda_x & \lambda \neq 0, \quad x \geq 0 \\ \log(x+1) & \lambda = 0, \quad x \geq 0 \\ -[(-x+1)^{2-\lambda_x} - 1]/(2-\lambda) & \lambda \neq 2, \quad x < 0 \\ -\log(-x+1) & \lambda = 2, \quad x < 0 \end{cases}. \tag{3}$$

We still transform the observed rainfall y using Eq. (2).

After transforming x and y individually, we assume the joint distribution to be bivariate normal:

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

where

$$\boldsymbol{\mu} = \begin{bmatrix} \mu_{\hat{x}} \\ \mu_{\hat{y}} \end{bmatrix}$$

$$\boldsymbol{\Sigma} = \begin{bmatrix} \sigma_{\hat{x}}^2 & \rho_{\hat{x}\hat{y}}\sigma_{\hat{x}}\sigma_{\hat{y}} \\ \rho_{\hat{x}\hat{y}}\sigma_{\hat{x}}\sigma_{\hat{y}} & \sigma_{\hat{y}}^2 \end{bmatrix}.$$

The model parameters $\boldsymbol{\theta}$ include transformation coefficients (some of λ_x, α_x, α_y, β_x, and β_y), means ($\mu_{\hat{x}}$ and $\mu_{\hat{y}}$), standard deviations ($\sigma_{\hat{x}}$ and $\sigma_{\hat{y}}$), and correlation coefficient ($\rho_{\hat{x}\hat{y}}$). The model can be shown to lead to the following prediction equation:

$$p(\hat{y} \mid \hat{x}, \boldsymbol{\theta}) \sim N\left[\mu_{\hat{y}} + \rho_{\hat{x}\hat{y}}\frac{\sigma_{\hat{y}}}{\sigma_{\hat{x}}}(\hat{x} - \mu_{\hat{x}}), (1 - \rho_{\hat{x}\hat{y}}^2)\sigma_{\hat{y}}^2\right], \tag{5}$$

where $\hat{y}$ can be back-transformed to y using the inverse of Eq. (2). It is clear from Eq. (5) that the model produces probabilistic forecasts.

In this study, a Bayesian method with Markov chain Monte Carlo (MCMC) sampling is used to infer the model parameters and uncertainty (Wang et al. 2009; Wang and Robertson 2011). If $(\mathbf{x}_D, \mathbf{y}_D)$ contains training data used for model inference and we have x for a new event, the posterior predictive density for the corresponding y is

$$f(y \mid x) = p(y \mid x; \mathbf{x}_D, \mathbf{y}_D) = \int p(y \mid x; \boldsymbol{\theta}) p(\boldsymbol{\theta} \mid \mathbf{x}_D, \mathbf{y}_D)\, d\boldsymbol{\theta}. \tag{6}$$

The uncertainty of the prediction is thus influenced by parameter uncertainty as well as the strength of the correlation and natural variability. The parameter uncertainty can be important when there is only a short data record to establish the model.

When seasonal rainfall totals of zero (no rainfall) are present in the data, modifications are needed to the above formulation. As the number of zero occurrences of seasonal rainfall is limited, we simply refer to Wang

and Robertson (2011) as the source of additional information on the handling of this problem.

b. *Bayesian model averaging for merging calibration and bridging forecasts*

The BMA methodology here essentially follows that described by Wang et al. (2012b) and applied by Schepen et al. (2012a), but with some modifications. One major difference is that in the inference of model weights, we have chosen to use the predictive densities corresponding to events used to fit the model, rather than using cross-validation predictive densities. Although we believe the latter approach is preferable, the computational requirements for a completely clean cross-validation exercise are much higher. Therefore, using "fitted" predictive densities to infer model weights allows more rigorous testing in cross validation by keeping the events being verified completely independent from the inference of model parameters and model weights.

We outline the key features of the methodology here. For each event, forecasts from a set of models M_k, $k = 1, \ldots, K$, are produced using the BJP modeling approach as outlined in section 3a. A merged forecast is then given by the BMA predictive density:

$$f_{\mathrm{BMA}}(y \mid x_k, k = 1, \ldots, K) = \sum_{k=1}^{K} w_k f_k(y \mid x_k), \quad (7)$$

where x_k and y are the predictor and predictand variables, respectively, and w_k is the BMA weight for model k. The BMA weights are inferred using a finite mixture model approach (e.g., Raftery et al. 2005; Wang et al. 2012b). We write the posterior distribution of the weights, given events $t = 1, 2, \ldots, T$ as follows:

$$\begin{aligned} &p[w_k, k = 1, \ldots, K \mid \mathbf{y}_D, \mathbf{x}_{D,k}, f_k(y \mid x_k), k = 1, \ldots, K] \\ &\propto p(w_k, k = 1, \ldots, K) \prod_{t=1}^{T} f_{\mathrm{BMA}}(y_D^t \mid x_{D,k}^t, k = 1, \ldots, K) \\ &\propto p(w_k, k = 1, \ldots, K) \prod_{t=1}^{T} \sum_{k=1}^{K} w_k f_k(y_D^t \mid x_{D,k}^t), \end{aligned} \quad (8)$$

where $p(w_k, k = 1, \ldots, K)$ is a prior of the weights, the remaining term on the rhs is the likelihood function and $x_{D,k}^t$ and y_D^t are the observed values for the predictor and predictand variables, respectively.

The prior is used here to constrain the effect of sampling error that arises due to the short historical data period and thus stabilize the weights. Following Wang et al. (2012b), we specify a symmetric Dirichlet prior:

$$p(w_k, k = 1, \ldots, K) \propto \prod_{k=1}^{K} (w_k)^{\alpha - 1}. \quad (9)$$

The parameter α is known as the concentration parameter. In this study we set it to $\alpha = 1.0 + \alpha_0/K$ with $\alpha_0 = 1.0$. This gives a slight preference toward more evenly distributed weights. The rhs of Eq. (8) then reduces to

$$A = \prod_{k=1}^{K} (w_k)^{\alpha - 1} \prod_{t=1}^{T} \sum_{k=1}^{K} w_k f_k(y_D^t \mid x_{D,k}^t). \quad (10)$$

We find a point estimate of the weights by maximizing A using an efficient expectation-maximization (EM) algorithm (Cheng et al. 2006; Wang et al. 2012b; Zivkovic and van der Heijden 2004). All weights are initially set to be equal ($1/K$), and the EM algorithm is then iterated until convergence of $\ln(A)$ is achieved. At each iteration j, the EM algorithm has two steps. In the first step, ownerships are calculated for all t and k:

$$O_k^{t,(j+1)} = \frac{w_k^{(j)} f_k(y_D^t \mid x_{D,k}^t)}{\sum_{m=1}^{K} w^{(j)} f_m(y_D^t \mid x_{D,m}^t)}. \quad (11)$$

Then, new weights are calculated by

$$w_k^{(j+1)} = \frac{(1/T) \sum_{t=1}^{T} O_k^{t,(j+1)} + (\alpha - 1)/T}{1 + K(\alpha - 1)/T}. \quad (12)$$

To form the BMA forecasts, we randomly sample a number of ensemble members from each model's ensemble, with the number drawn being proportional to the model weight.

c. *Forecast assessment*

In this study, we assess the performance of leave-three-years-out cross-validation forecasts for the period FMA 1980 to JFM 2012. Separate models are established for each season and grid cell independently. For each historical forecast event to be tested, the data points for the year to be forecast, the year before and the year after are omitted so as not to influence the model parameter and weight inferences. This procedure is repeated for each event in the historical record. The limited data available does not allow for verification in an independent period, but by leaving three years out forecast events are considered to be independent of the training data.

Forecasts from the BMA method are probabilistic. Important attributes of probabilistic forecasts include accuracy, sharpness, resolution, and reliability. To assess forecast accuracy, we use the continuous ranked probability score (CRPS; Matheson and Winkler 1976). CRPS is used to assess full forecast probability distributions, and, therefore, it scores forecast sharpness as well as forecast accuracy. Mathematically, CRPS is given by

$$\mathrm{CRPS} = \frac{1}{T}\sum_{t=1}^{T}\int [F(y^t) - H(y^t - y_D^t)]^2\, dy^t, \quad (13)$$

where F is the forecast CDF and H is the Heaviside step function such that

$$H(y^t - y_D^t) = \begin{cases} 0 & y^t < y_D^t \\ 1 & y^t \geq y_D^t \end{cases}. \quad (14)$$

To assess forecast value, the CRPS score is converted to a skill score, to measure the relative improvement of the forecasts over the corresponding leave-three-years-out cross-validation climatology (used as reference forecasts):

$$\mathrm{CRPS}_{\mathrm{SkillScore}} = \frac{\mathrm{CRPS}_{\mathrm{ref}} - \mathrm{CRPS}}{\mathrm{CRPS}_{\mathrm{ref}}} \times 100. \quad (15)$$

A skill score of 100 means perfect forecasts, while a skill score of 0 means that the model forecasts are no better than using climatology, and thus considered of no skill. A negative skill score means that the model forecasts are worse than using climatology, and the model may have been unduly influenced by data noise.

Reliability, sharpness, and resolution are assessed for binary expressions of the probabilistic forecasts by plotting an attributes diagram (Hsu and Murphy 1986). Here, we express the forecasts as the probability of exceeding the climatological median as is done for official forecasts in Australia. Reliability and resolution are checked by plotting the forecast probabilities of events against their observed relative frequencies. Sharpness is checked by plotting the number of forecasts in bins of the forecast probability.

4. Results

a. Skill of calibration forecasts

We first assess calibration forecasts, which are those based on POAMA rainfall forecasts. We calibrate the forecasts using BJP and then merge the forecasts using BMA. The leave-three-years-out cross-validation CRPS skill scores for FMA 1980 to JFM 2012 are mapped in Fig. 1. There is one panel for each three-consecutive-months–season, and the average skill score for the season is written in the bottom-left corner. The three seasons with the lowest average skill scores are, in increasing order, as follows: DJF, NDJ, and AMJ. The seasons with the highest average skill scores are, in decreasing order, as follows: MAM, SON, and JAS. The calibration forecasts are skillful in parts of eastern and northern Australia from JJA to OND, suggesting some ability to forecast ENSO modulated rainfall. However, we note that the skill scores are overall modest. In many regions there is little or no skill. It is, therefore, desirable to try to increase the overall skill of the forecasts.

b. Skill of bridging forecasts

We now assess bridging forecasts, which are those based on POAMA forecasts of Pacific and Indian Ocean SSTs. We produce forecasts of rainfall using separate BJP models with Niño-3, Niño-4, Niño-3.4, EMI, DMI, and the Indonesia index (II) as predictors. We merge the forecasts of all bridging models using BMA. The leave-three-years-out cross-validation CRPS skill scores for FMA 1980 to JFM 2012 are mapped in Fig. 2. Again, there is one panel for each season, and the average skill score for the season is written in the bottom-left corner. The three seasons with the lowest average skill scores are, in increasing order, as follows: MJJ, JFM, and MAM. The seasons with the highest average skill scores are, in decreasing order, as follows: OND, SON, and JAS.

Compared to the calibration forecasts, the average skill scores for the JAS–DJF bridging forecasts are higher than for the corresponding calibration forecasts, with the exception of ASO. In contrast, the average skill scores for the JFM–JJA bridging forecasts are lower than for the corresponding calibration forecasts, with the exception of FMA. To generalize, the calibration forecasts are more skillful in the first half of the year, whereas the bridging forecasts are more skillful in the second half of year.

A comparison of Figs. 1 and 2 also reveals finer details. For SON, we see that the calibration and bridging forecasts are similarly skillful in southeastern Australia; however, the bridging forecasts are more skillful across northern Australia, leading to a higher average seasonal skill score. For OND, we see that the bridging forecasts have a superior spatial coverage of skill. In contrast, for MAM, we see that the calibration forecasts have a superior coverage of skill. What is clear is that the calibration and bridging forecasts each have some unique ability to forecast Australian seasonal rainfall, and the overall most skillful forecasts may be achieved by merging all calibration and bridging forecasts.

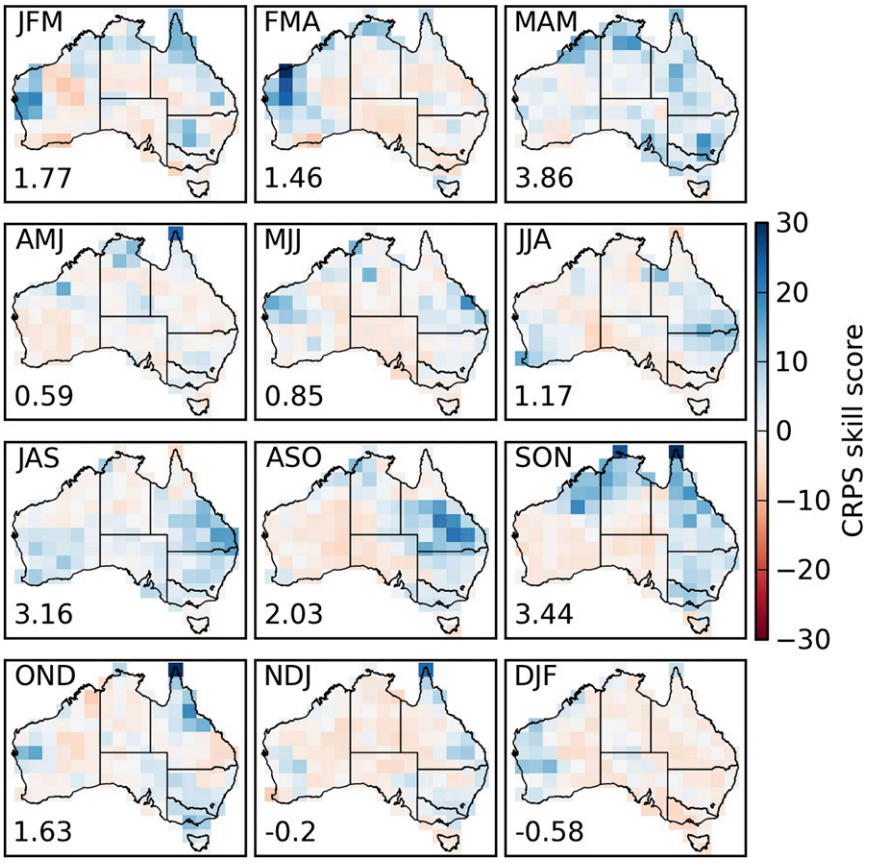

FIG. 1. Three-consecutive-months–seasonal CRPS skill scores for FMA 1980 to JFM 2012 calibration forecasts at a lead time of 1 month.

c. *Skill of merged calibration and bridging forecasts*

We now assess merged calibration and bridging forecasts, which are based on both POAMA rainfall and POAMA SST forecasts. The forecasts of all calibration and bridging models are merged using BMA. The leave-three-years-out cross-validation CRPS skill scores for FMA 1980 to JFM 2012 are mapped in Fig. 3. Again, there is one panel for each season, and the average skill score for the season is written in the bottom-left corner. The three seasons with the lowest average skill scores are, in increasing order, as follows: DJF, MJJ, and AMJ. The seasons with the highest average skill scores are, in decreasing order, as follows: OND, SON, and JAS (i.e., the same as for the bridging forecasts).

To assess whether it is overall beneficial to merge the calibration and bridging forecasts, we plot the count of grid cells where CRPS skill score thresholds are exceeded and compare the results for calibration, bridging, and merged calibration and bridging forecasts (Fig. 4). We consider all seasons at once, therefore, higher counts represent a higher regional and seasonal coverage of skill. The merged calibration and bridging forecasts have the overall highest coverage of skill, and the calibration forecasts have the overall lowest coverage of skill. For CRPS skill score thresholds of 15 and above, the bridging forecasts and the merged calibration and bridging forecasts have an overall similar coverage of skill.

d. *Reliability and sharpness of merged calibration and bridging forecasts*

The overall reliability and sharpness of the merged calibration and bridging forecasts are visually assessed using an attributes diagram (Fig. 5). We check the consistency of forecast probabilities of exceeding the observed climatological median with observed relative frequencies of occurrence. For this analysis, all forecasts for all grid cells and seasons have been pooled. Referring to Fig. 5, the forecast probabilities have been binned into bins of width 0.1. The plotting points for the x axis are the average forecast probabilities within each bin. The plotting points on the y axis are the observed relative frequencies. The size of the dots represents the

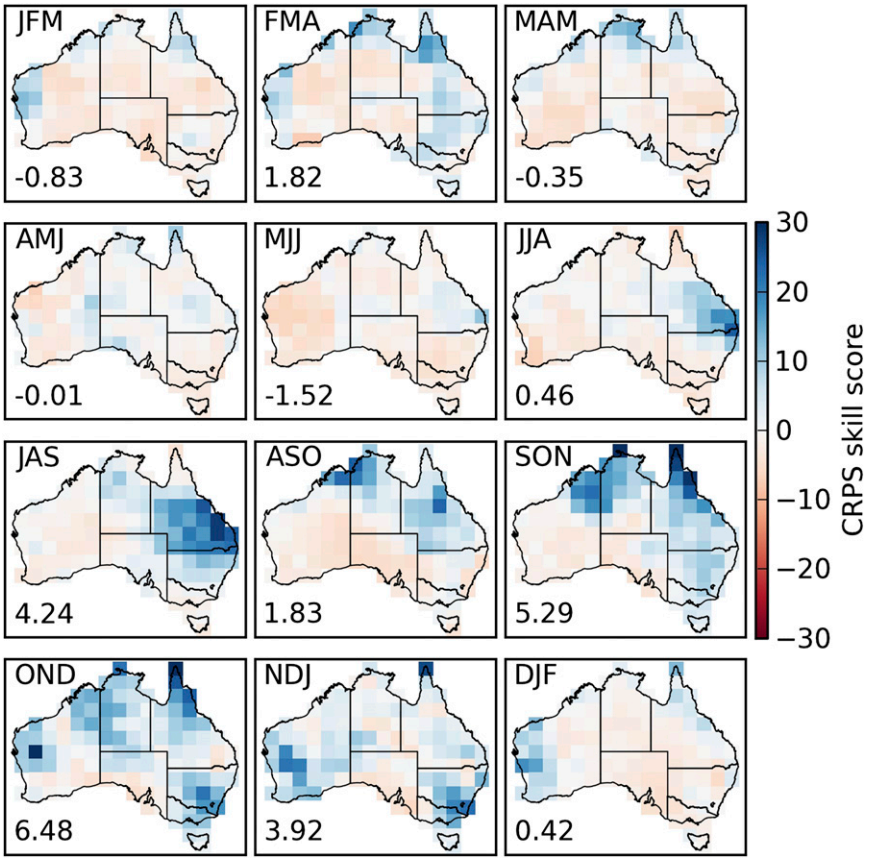

FIG. 2. As in Fig. 1, but for bridging forecasts.

proportion of forecasts in the bin and, therefore, forecast sharpness. For perfect forecasts, that is if the forecasts predicted the correct category for every event with absolute certainty, then we would have two points on the chart; one at the top-right corner and one at the bottom-left corner. However, reliable forecasts are still possible when there is considerable uncertainty. The forecast probabilities in the four central bins are reasonably consistent with observed relative frequencies of occurrence, as indicated by the centers of the dots being positioned close to the 1:1 line, suggesting that the forecast probabilities are reliable. For more emphatic forecasts, the centers of the dots deviate slightly from the 1:1 line, suggesting that the forecast probabilities are a little overly emphatic compared to the observed frequencies. However, all of the points lie within the shaded area.

Focusing on forecast sharpness, we observe that the forecast probabilities are clustered toward the climatological probability and there are relatively few events with emphatic probabilities. Given the overall low-to-moderate skill of seasonal forecasting (Figs. 1–3), the lack of forecast sharpness is expected.

e. Comparison to the approach of correcting the ensemble mean

The Bayesian calibration and forecast merging methodologies employed in this study are sophisticated. It is, therefore, reasonable to ask how well this approach compares to a simpler approach. We now present the results for the approach of applying a mean correction to the raw forecast ensembles. Following a leave-three-years-out cross-validation sequence, we adjust each M2.4 raw forecast ensemble by subtracting the error between the M2.4 climatological mean (from hindcasts) and the observed climatological mean. We apply the same correction to all ensemble members, after pooling the 33 ensemble members of M2.4a, b, and c. The aim of the mean correction is to correct systematic biases only. It does not adjust ensemble spread.

One disadvantage of the mean-correction approach is that is does not allow us to incorporate bridging, which has already been shown to improve the skill over calibration. It is a correction of POAMA rainfall only. The leave-three-years-out cross-validation CRPS skill scores for FMA 1980 to JFM 2012 are mapped in Fig. 6. Again,

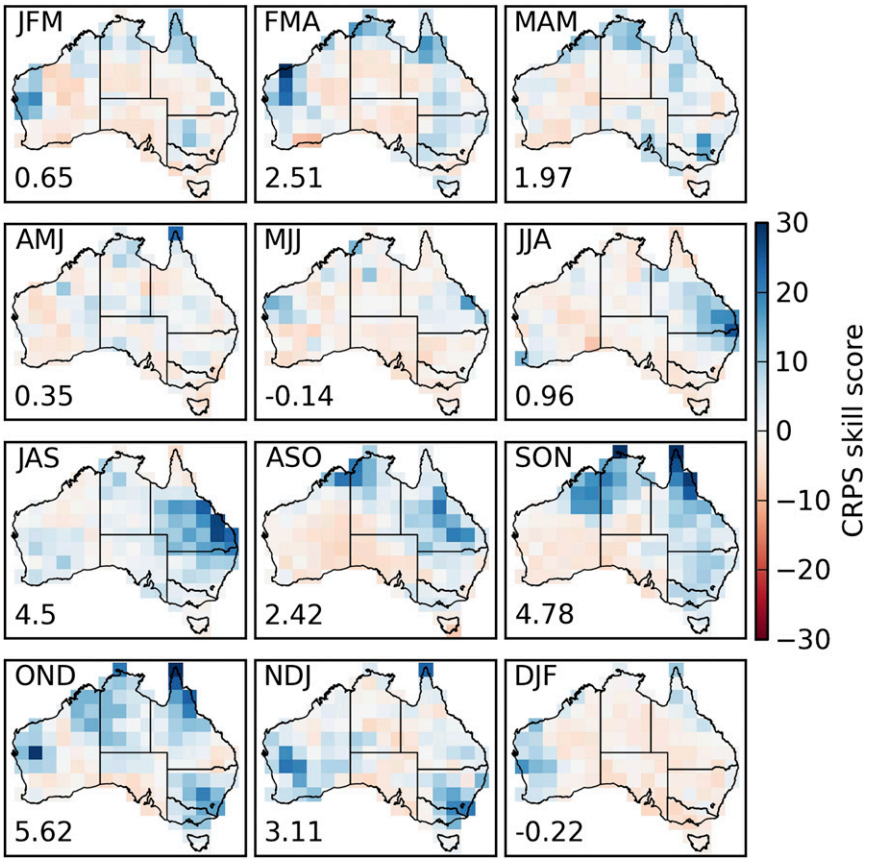

FIG. 3. As in Fig. 1, but for merged calibration and bridging forecasts.

there is one panel for each season, and the average skill score for the season is written in the bottom-left corner. In many regions and seasons, the CRPS skill scores of the mean-corrected forecasts are sharply negative, less than −10. For all seasons except JFM and MAM, the average skill score is negative. The skill maps for the BJP calibration and bridging forecasts (Figs. 1–3) are not afflicted by severe negative skill scores as is the case with the mean corrected forecasts. One of the reasons for this is that the BJP modeling approach is designed to produce climatological forecasts in the absence of forecasting skill.

The overall reliability and sharpness of the mean-corrected forecasts are visually assessed using an attributes diagram (Fig. 7). The forecast probabilities are not consistent with observed outcomes, as indicated by the centers of the dots deviating from the 1:1 line, suggesting that the forecast probabilities are not reliable. In many cases, the centers of the dots lie outside the shaded skill area. When comparing with the reliability of the merged calibration and bridging forecasts (Fig. 5), the reliability of the mean-corrected forecasts is poor. Even though the mean-correction approach leads to a sharper forecasts and a higher proportion of more emphatic forecast probabilities, reliability is compromised. This is evident by the probabilities not being consistent with the observed relative frequencies of occurrence. Through this comparison, the calibration and bridging forecasts are seen to represent a significant improvement over the mean corrected forecasts.

5. Discussion

To determine exactly why the calibration models tend to be more skillful in the first half of the year and why bridging models tend to be more skillful in the second half of the year would require detailed investigations of POAMA's ability to simulate well-known physical mechanisms that have an impact on Australian rainfall. Research in this area is ongoing (Lim et al. 2009; Marshall et al. 2011a,b, 2012; Zhao and Hendon 2009), but to elaborate by way of example, consider SON–OND rainfall in southeastern Australia, which is known to have a physical link to the Indian Ocean circulation through a sequence of Rossby wave trains (Cai et al. 2011). The statistical relationship between the IOD and

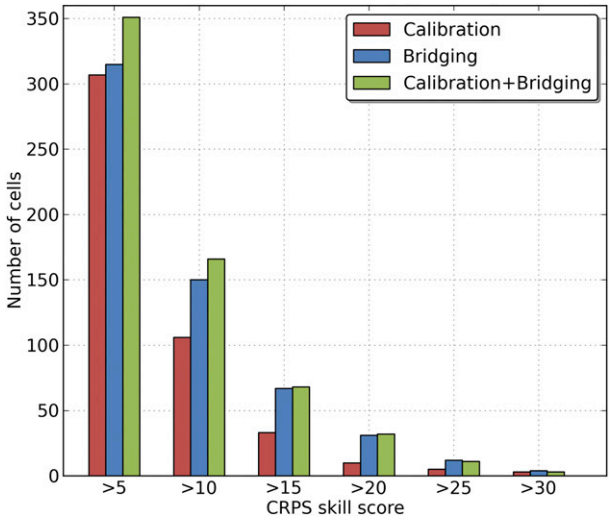

FIG. 4. Counts of grid cells where a range of CRPS skill score levels are exceeded for calibration forecasts (red), bridging models (blue), and merged calibration and bridging forecasts (green). Higher counts indicate a greater regional and seasonal coverage of skill.

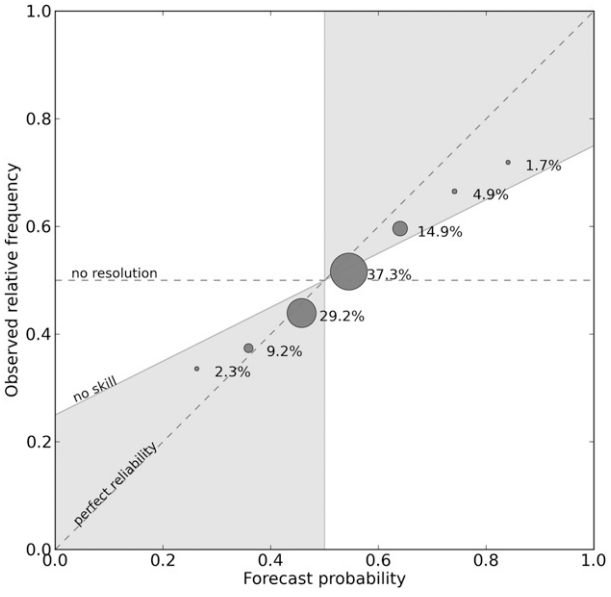

FIG. 5. Attributes diagram for merged calibration and bridging forecasts of probability of exceeding the observed climatological median at a lead time of 1 month. Forecasts for all grid cells and seasons FMA 1980 to JFM 2012 have been pooled. Forecast probabilities have been binned with width 0.1. The dot size and corresponding written percentages depict forecast sharpness as the proportion of forecasts in each bin.

southeastern Australian spring rainfall is well known (e.g., Schepen et al. 2012b; Risbey et al. 2009). If POAMA fails to correctly simulate Indian Ocean variability and/or fails to correctly propagate the Rossby wave trains, then the atmospheric convection over southeastern Australia may not be modeled properly. Consequently, calibration forecasts could have limited skill due to limited skill of the underlying raw POAMA forecasts. Further to the point, if the IOD is well forecast by POAMA, then bridging forecasts could be expected to have skill by virtue of the simple statistical relationship. We do not delve deeper into understanding the physics here; suffice to say, much of the skill in bridging models in northern and eastern Australia in the second half of the year can be attributed to strong statistical relationships of the rainfall with the ENSO (e.g., Risbey et al. 2009; Schepen et al. 2012b). The expected future improvements to POAMA's physics, initialization scheme, and parameterizations can only improve the skill of the calibration models, whether through spatial correction of large-scale features or improved modeling of localized rainfall effects.

The results show that overall more skillful forecasts can be obtained by choosing a combination of calibration and bridging as a forecast strategy. While the merged calibration and bridging forecasts are overall more skillful than the calibration or bridging forecasts, when considering all seasons, it is noted that overall higher skill scores would be achieved if, say, calibration models are used for the first half of the year and bridging models are used for the second half of the year. The downside to this strategy is that the reasoning for such a strategy may be difficult to communicate. Also, a more consistent, objective approach is expected to lead to better results in the long run. In this respect, merging calibration and bridging forecasts for all seasons is a better strategy, as the BMA will automatically favor the historically better performing models in each region and season.

The results shown here are based on leave-three-years-out cross validation. To ensure independent events were reserved for verification, we inferred model weights based on how well the models reforecast the events used to fit the models. For forecasting new events, cross-validation predictive densities could be used to infer the weights as in Wang et al. (2012b) instead of the fitted predictive densities, so that the weights reflect more the predictive abilities of the models.

The BMA approach is used in this study for merging forecasts from multiple calibration and bridging models and is shown to be effective. The approach can be easily used to incorporate other international GCMs (e.g., from the European Centre for Medium-Range Weather Forecasts), which may perform better in certain seasons and/or regions. This will be studied in our future research.

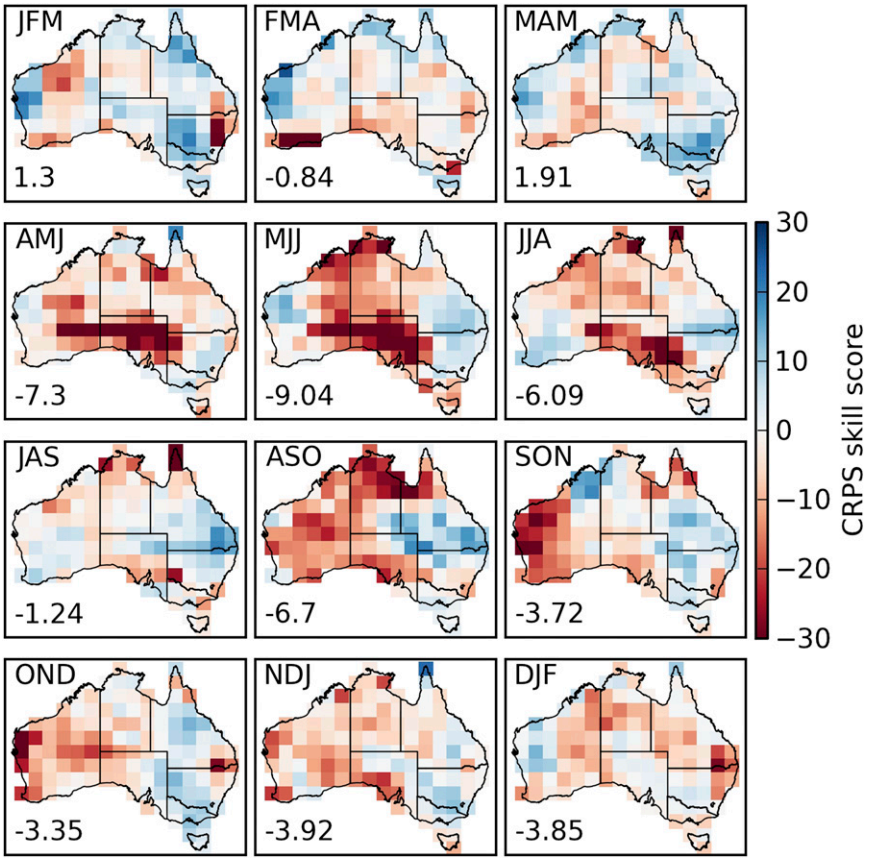

FIG. 6. As in Fig. 1, but for mean-corrected forecasts.

6. Conclusions

We have applied Bayesian joint probability modeling and Bayesian model averaging (BJP–BMA) to improve seasonal forecasts of Australian seasonal rainfall through statistical calibration and bridging of POAMA (version M2.4) coupled GCM outputs. Calibration of seasonal rainfall amount accounts for systematic errors that cause raw forecast ensembles to be biased and statistically unreliable. When compared to simple bias correction methods, such as mean correction, BJP–BMA calibration has been shown to markedly improve forecast accuracy as measured by the CRPS skill score. The skill scores obtained through calibration alone, however, are modest, and in many regions there is little or no skill. Including bridging models in the BMA has been shown to improve the regional and seasonal coverage of higher CRPS skill scores.

In general, the calibration forecasts are more skillful in the first half of the year, whereas the bridging forecasts are more skillful in the second half of year, although there are seasons and regions where calibration and bridging perform similarly. Because the calibration and bridging forecasts each have some unique ability to forecast Australian seasonal rainfall, overall the most skillful forecasts can be achieved by merging all calibration and bridging forecasts.

For the merged calibration and bridging forecasts, forecast probabilities of exceeding the observed

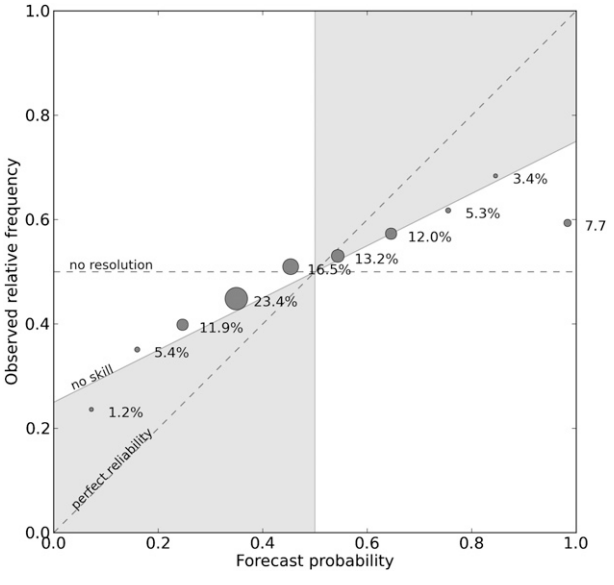

FIG. 7. As in Fig. 5, but for mean-corrected forecasts.

climatological median are mostly consistent with observed outcomes although extreme forecast probabilities exhibit compromised forecast reliability. When compared to the reliability of mean-corrected forecasts, the BJP–BMA forecasts show significantly improved reliability.

Acknowledgments. This work was completed as part of the Water Information Research and Development Alliance between CSIRO and the Bureau of Meteorology. We thank our colleagues from the Centre for Australian Weather and Climate Research for providing access to hindcast and real-time data of the Predictive Ocean Atmosphere Model for Australia.

REFERENCES

Ashok, K., S. Behera, S. Rao, H. Weng, and T. Yamagata, 2007: El Niño Modoki and its possible teleconnection. *J. Geophys. Res.,* **112,** C11007, doi:10.1029/2006JC003798.

Bartman, A., W. Landman, and C. J. D. E. Rautenbach, 2003: Recalibration of general circulation model output to austral summer rainfall over southern Africa. *Int. J. Climatol.,* **23,** 1407–1419, doi:10.1002/joc.954.

Cai, W., P. van Rensch, T. Cowan, and H. H. Hendon, 2011: Teleconnection pathways of ENSO and the IOD and the mechanisms for impacts on Australian rainfall. *J. Climate,* **24,** 3910–3923, doi:10.1175/2011JCLI4129.1.

Cheng, J., J. Yang, Y. Zhou, and Y. Cui, 2006: Flexible background mixture models for foreground segmentation. *Image Vis. Comput.,* **24,** 473–482, doi:10.1016/j.imavis.2006.01.018.

Colman, R., and Coauthors, 2005: BMRC Atmospheric Model (BAM) version 3.0: Comparison with mean climatology. BMRC Research Rep. 108, 60 pp.

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, doi:10.1175/1520-0442(1999)012<1974:ROMSEB>2.0.CO;2.

Graham, R. J., M. Gordon, P. J. McLean, S. Ineson, M. R. Huddleston, M. K. Davey, A. Brookshaw, and R. T. H. Barnes, 2005: A performance comparison of coupled and uncoupled versions of the Met Office seasonal prediction general circulation model. *Tellus,* **57,** 320–339, doi:10.1111/j.1600-0870.2005.00116.x.

Guilyardi, E., P. Delecluse, S. Gualdi, and A. Navarra, 2003: Mechanisms for ENSO phase change in a coupled GCM. *J. Climate,* **16,** 1141–1158, doi:10.1175/1520-0442(2003)16<1141:MFEPCI>2.0.CO;2.

Hoeting, J. A., D. Madigan, A. E. Raftery, and C. T. Volinsky, 1999: Bayesian model averaging: A tutorial. *Stat. Sci.,* **14,** 382–417, doi:10.1214/ss/1009212519.

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

Hudson, D., O. Alves, H. Hendon, and G. Wang, 2011: The impact of atmospheric initialisation on seasonal prediction of tropical Pacific SST. *Climate Dyn.,* **36,** 1155–1171, doi:10.1007/s00382-010-0763-9.

Jin, E. K., and Coauthors, 2008: Current status of ENSO prediction skill in coupled ocean–atmosphere models. *Climate Dyn.,* **31,** 647–664, doi:10.1007/s00382-008-0397-3.

Jones, D. A., W. Wang, and R. Fawcett, 2009: High-quality spatial climate data-sets for Australia. *Aust. Meteor. Oceanogr. J.,* **58,** 233–248.

Landman, W. A., and L. Goddard, 2002: Statistical recalibration of GCM forecasts over southern Africa using model output statistics. *J. Climate,* **15,** 2038–2055, doi:10.1175/1520-0442(2002)015<2038:SROGFO>2.0.CO;2.

Langford, S., H. Hendon, and E. P. Lim, 2011: Assessment of POAMA's predictions of some climate indices for use as predictors of Australian Rainfall. CAWCR Tech. Rep. 31, 60 pp.

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, doi:10.1175/2009MWR2904.1.

——, ——, D. L. T. Anderson, A. Charles, and O. Alves, 2011: Dynamical, statistical-dynamical, and multimodel ensemble forecasts of Australian spring season rainfall. *Mon. Wea. Rev.,* **139,** 958–975, doi:10.1175/2010MWR3399.1.

Marshall, A. G., D. Hudson, M. C. Wheeler, H. Hendon, and O. Alves, 2011a: Evaluating key drivers of Australian intra-seasonal climate variability in POAMA-2: A progress report. *CAWCR Research Letters,* No. 7, Center for Australian Weather and Climate Research, Melbourne, Australia, 10–16.

——, ——, ——, H. H. Hendon, and O. Alves, 2011b: Assessing the simulation and prediction of rainfall associated with the MJO in the POAMA seasonal forecast system. *Climate Dyn.,* **37,** 2129–2141, doi:10.1007/s00382-010-0948-2.

——, ——, ——, ——, and ——, 2012: Simulation and prediction of the Southern Annular Mode and its influence on Australian intra-seasonal climate in POAMA. *Climate Dyn.,* **38,** 2483–2502, doi:10.1007/s00382-011-1140-z.

Matheson, J. E., and R. L. Winkler, 1976: Scoring rules for continuous probability distributions. *Manage. Sci.,* **22,** 1087–1096, doi:10.1287/mnsc.22.10.1087.

Oke, P., A. Schiller, D. Griffin, and G. Brassington, 2005: Ensemble data assimilation for an eddy resolving ocean model of the Australian region. *Quart. J. Roy. Meteor. Soc.,* **131,** 3301–3311, doi:10.1256/qj.05.95.

Palmer, T., and Coauthors, 2004: Development of a European Multimodel Ensemble System for Seasonal-to-Interannual Prediction (DEMETER). *Bull. Amer. Meteor. Soc.,* **85,** 853–872, doi:10.1175/BAMS-85-6-853.

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

Risbey, J. S., M. J. Pook, P. C. McIntosh, M. C. Wheeler, and H. H. Hendon, 2009: On the remote drivers of rainfall variability in Australia. *Mon. Wea. Rev.,* **137,** 3233–3253, doi:10.1175/2009MWR2861.1.

Roeckner, E., J. M. Oberhuber, A. Bacher, M. Christoph, and I. Kirchner, 1996: ENSO variability and atmospheric response in a global coupled atmosphere-ocean GCM. *Climate Dyn.,* **12,** 737–754, doi:10.1007/s003820050140.

Saji, N. H., B. N. Goswami, P. N. Vinayachandran, and T. Yamagata, 1999: A dipole mode in the tropical Indian Ocean. *Nature,* **401,** 360–363.

Schepen, A., Q. J. Wang, and D. E. Robertson, 2012a: Combining the strengths of statistical and dynamical modeling approaches for forecasting Australian seasonal rainfall. *J. Geophys. Res.,* **117,** D20107, doi:10.1029/2012JD018011.

——, ——, and D. Robertson, 2012b: Evidence for using lagged climate indices to forecast Australian seasonal rainfall. *J. Climate,* **25,** 1230–1246, doi:10.1175/JCLI-D-11-00156.1.

Schiller, A., J. S. Godfrey, P. C. McIntosh, G. Meyers, N. R. Smith, O. Alves, G. Wang, and R. Fiedler, 2002: A new version of the Australian Community Ocean Model for seasonal climate prediction. CSIRO Marine Laboratories Rep. 240, 82 pp. [Available online at http://www.cmar.csiro.au/publications/cmrreports/240/abs240.html.]

Verdon, D. C., and S. W. Franks, 2005: Indian Ocean sea surface temperature variability and winter rainfall: Eastern Australia. *Water Resour. Res.,* **41,** W09413, doi:10.1029/2004WR003845.

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.

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

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

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

Wilby, R. L., T. Wigley, D. Conway, P. Jones, B. Hewitson, J. Main, and D. Wilks, 1998: Statistical downscaling of general circulation model output: A comparison of methods. *Water Resour. Res.,* **34,** 2995–3008, doi:10.1029/98WR02577.

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

Yin, Y., O. Alves, and P. R. Oke, 2011: An ensemble ocean data assimilation system for seasonal prediction. *Mon. Wea. Rev.,* **139,** 786–808, doi:10.1175/2010MWR3419.1.

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, doi:10.1002/qj.370.

Zivkovic, Z., and F. van der Heijden, 2004: Recursive unsupervised learning of finite mixture models. *IEEE Trans. Pattern Anal. Mach. Intell.,* **26,** 651–656, doi:10.1109/TPAMI.2004.1273970.

Zorita, E., and H. von Storch, 1999: The Analog method as a simple statistical downscaling technique: Comparison with more complicated methods. *J. Climate,* **12,** 2474–2489, doi:10.1175/1520-0442(1999)012<2474:TAMAAS>2.0.CO;2.