Toward Accurate and Reliable Forecasts of Australian Seasonal Rainfall by Calibrating and Merging Multiple Coupled GCMs

ANDREW SCHEPEN

Bureau of Meteorology, Brisbane, Queensland, Australia

Q. J. WANG

CSIRO Land and Water, Melbourne, Victoria, Australia

(Manuscript received 31 August 2012, in final form 21 June 2013)

ABSTRACT

The majority of international climate modeling centers now produce seasonal rainfall forecasts from coupled general circulation models (GCMs). Seasonal rainfall forecasting is highly challenging, and GCM forecast accuracy is still poor for many regions and seasons. Additionally, forecast uncertainty tends to be underestimated meaning that forecast probabilities are statistically unreliable. A common strategy employed to improve the overall accuracy and reliability of GCM forecasts is to merge forecasts from multiple models into a multimodel ensemble (MME). The most widely used technique is to simply pool all of the forecast ensemble members from multiple GCMs into what is known as a superensemble. In Australia, seasonal rainfall forecasts are produced using the Predictive Ocean–Atmosphere Model for Australia (POAMA). In this paper, the authors demonstrate that mean corrected superensembles formed by merging forecasts from POAMA with those from three international models in the ENSEMBLES dataset remain poorly calibrated in many cases. The authors propose and evaluate a two-step process for producing MMEs. First, forecast calibration of the individual GCMs is carried out by using Bayesian joint probability models that account for parameter uncertainty. The calibration leads to satisfactory forecast reliability. Second, the individually calibrated forecasts of the GCMs are merged through Bayesian model averaging (BMA). The use of multiple GCMs results in better forecast accuracy, while maintaining reliability, than using POAMA only. Compared with using equal-weight averaging, BMA weighting produces sharper and more accurate forecasts.

1. Introduction

Many organizations around the world are now producing global seasonal rainfall forecasts from coupled ocean–atmosphere general circulation models (coupled GCMs; e.g., Lim et al. 2011; Palmer et al. 2004; Saha et al. 2006; Yasuda et al. 2007). In Australia, the Bureau of Meteorology and the Commonwealth Scientific and Industrial Research Organisation (CSIRO) have developed the Predictive Ocean–Atmosphere Model for Australia (POAMA; Wang et al. 2011), an ocean–atmosphere coupled GCM. Like most GCMs, POAMA produces ensemble forecasts that are not accurate in terms of rainfall amount. The forecasts are generally too narrow in ensemble spread, resulting in underestimated forecast uncertainty and thus statistically unreliable forecast probabilities. For the forecasts to be used in decision making and risk management, forecast probabilities need to be statistically reliable. Additionally, for many applications, it is important that rainfall amounts are accurate. In this paper, we seek ensemble seasonal forecasts of rainfall amount that are as accurate as possible and statistically reliable.

Errors in GCM forecasts arise for two main reasons. First, errors are introduced through parameterizations of subgrid processes. Parameterizations are necessary because of the coarse spatial resolutions of the models. Second, errors arise as a result of errors in the initial conditions. The impact of errors in initial conditions can be quite pronounced because coupled GCMs can be sensitive to small variations in the initial conditions (e.g., Ploshay and Anderson 2002). Although all GCMs are based on fundamental physics, they differ in physical specifications and parameterizations, in initialization

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

DOI: 10.1175/MWR-D-12-00253.1

schemes and in perturbation schemes, and they are run at different resolutions. Consequently, different GCMs can produce quite different simulations of rainfall as the large-scale ocean–atmosphere circulations are evolved and the atmosphere interacts with the topographical land surface.

For quantitative forecasts of seasonal rainfall, it is necessary to calibrate raw GCM forecasts to remove biases and to properly represent forecast uncertainty. A common strategy employed to reduce bias and to increase the forecast ensemble spread, is to combine forecasts from multiple GCMs into a multimodel ensemble (MME; e.g., Krishnamurti et al. 1999). The simplest and most common approach to MME is to pool the forecast ensemble members of multiple GCMs into what is known as a superensemble (e.g., Hagedorn et al. 2005; Palmer et al. 2004, 2000). More sophisticated approaches have been tested. For example, Luo et al. (2007) developed a Bayesian approach to MME that explicitly accounted for nonnormal distributions such as for rainfall but assumed model independence, which is an impediment to producing reliable forecasts because forecast uncertainty can be further underestimated when the assumption is not true.

Recently, Langford and Hendon (2013) investigated MME forecasts of Australian seasonal rainfall forecasts by merging forecasts from POAMA with those from three international models in the ENSEMBLES dataset. Their results focused on two-category forecasts of above or below the model climatology median. Their MME approach generally improved the reliability of the two-category forecasts as the number of models in the MME increased, although the forecasts still underestimated forecast uncertainty overall. The results motivate us to evaluate other methods to obtain calibrated ensemble seasonal forecasts of rainfall amount.

The superensemble approach to MME is an ad hoc approach, because theoretically an arbitrary number or selection of models is required to obtain calibrated forecasts. Indeed, full calibration may never be achieved. An approach to overcome this problem, as demonstrated by (Doblas-Reyes et al. 2005), is to first calibrate the forecasts of individual models and then combine all the models. However, Doblas-Reyes et al. (2005) found that a more sophisticated model combination by using multiple linear regression did not outperform a simple pooling of calibrated forecast ensembles of multiple models with equal weights. Furthermore, they pointed out that sufficiently long training data are needed for calibration to be robust. As available hindcast records of most models are too short for precise calibration, calibration methods that allow for parameter uncertainty are expected to produce more reliable forecasts.

In this paper, we evaluate Bayesian approaches for obtaining calibrated ensemble forecasts of Australian seasonal rainfall amount from POAMA and three other models in the ENSEMBLES dataset. Improvements to reliability will be mainly achieved through individual calibration of the GCMs using Bayesian joint probability models that account for parameter uncertainty (Wang and Robertson 2011; Wang et al. 2009). A Bayesian model averaging (BMA) approach is then applied to weight and merge the forecasts of the multiple models. We show that including the three international models improves the regional and seasonal coverage of skill compared to using POAMA alone. In comparison to equally weighted forecasts, the BMA-weighted forecasts have better forecast sharpness and lead to higher skill scores. As part of our evaluation, we also compare our Bayesian approach to the superensemble approach to highlight the advantages of a more sophisticated approach in obtaining well-calibrated forecasts.

2. Methods and data

a. Bayesian joint probability calibration models

We calibrate each raw GCM forecast using a Bayesian joint probability (BJP) modeling and forecasting approach (Wang and Robertson 2011; Wang et al. 2009). A joint probability distribution model describes the relationship between raw GCM forecast ensemble means x and observed rainfall amounts y as following a bivariate normal distribution after Yeo–Johnson transformations (Yeo and Johnson 2000). More specifically, the variables x and y are transformed by

$$\hat{x} = \begin{cases} [(x+1)^{\lambda_x} - 1]/\lambda_x & \text{if} \quad \lambda_x \neq 0 \\ \log(x+1) & \text{if} \quad \lambda_x = 0 \end{cases} \tag{1}$$

$$\hat{y} = \begin{cases} [(y+1)^{\lambda_y} - 1]/\lambda_y & \text{if} \quad \lambda_y \neq 0 \\ \log(y+1) & \text{if} \quad \lambda_y = 0 \end{cases}. \tag{2}$$

The transformed variables are then assumed to be jointly normal:

$$p(\hat{x}, \hat{y}) \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma}), \tag{3}$$

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 (λ_x, λ_y), means ($\mu_{\hat{x}}, \mu_{\hat{y}}$), standard deviations ($\sigma_{\hat{x}}, \sigma_{\hat{y}}$), and correlation ($\rho_{\hat{x}\hat{y}}$). The transformations are for modeling typically right-skewed rainfall distributions. The model can be shown to lead to the following prediction (calibration) equation:

$$p(\hat{y} \mid \hat{x}, \boldsymbol{\theta}) \sim N\left[\mu_{\hat{y}} + \rho_{\hat{x}\hat{y}}\frac{\sigma_{\hat{y}}}{\sigma_{\hat{x}}}(\hat{x} - \mu_{\hat{x}}), (1 - \rho_{\hat{x}\hat{y}}^2)\sigma_{\hat{y}}^2\right], \quad (4)$$

where $\hat{y}$ can be back-transformed to y using the inverse of Eq. (2). It is clear from Eq. (4) that when there is little relationship between the predictor–predictand variables in the calibration (i.e., the correlation is near to zero), the prediction will revert to the marginal distribution of the predictand (natural variability or climatology). If the correlation is high, the prediction range will be considerably narrower (sharper forecasts).

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

$$f(y \mid x) = p(y \mid x; \mathbf{x}_D, \mathbf{y}_D) = \int p(y \mid x; \boldsymbol{\theta}) p(\boldsymbol{\theta} \mid \mathbf{x}_D, \mathbf{y}_D) d\boldsymbol{\theta}. \quad (5)$$

The uncertainty of the prediction is thus influenced by parameter uncertainty as well as the strength of the correlation and natural variability. Parameter uncertainty can be important when there is only a short data record to establish the model. By allowing for the strength of the correlation, natural variability, and model parameter uncertainty, forecast uncertainty should be better quantified.

b. Bayesian model averaging for forecast merging

We apply BMA to construct weighted MMEs of BJP calibrated forecasts. The BMA methodology used here is described in full by Wang et al. (2012) and has been applied by Schepen et al. (2012) for a statistical-dynamical modeling approach. For ease of understanding, we outline the key features of the methodology here. For each event, raw forecasts from K different models are calibrated using BJP as outlined in section 2a. An MME forecast is then given by the BMA predictive density:

$$f_{\mathrm{BMA}}(y \mid x_1, \ldots, x_K) = \sum_{k=1}^{K} w_k f_k(y \mid x_k), \quad (6)$$

where x_k is the predictor variable and w_k is the weight of model k.

The models' weights are inferred from the performance of merged forecasts using a finite mixture model approach, initially used by Raftery et al. (2005) and adapted by Wang et al. (2012). The mixture model approach is different to early approaches to BMA whereby weights were calculated from the posterior probabilities of individual models (Hoeting et al. 1999). Following the methodology of Wang et al. (2012), we constrain the effect of sampling error that arises due to the short historical data period (26 yr) by specifying a prior probability distribution for the weights. We specify a symmetric Dirichlet prior:

$$p(w_k, k = 1, \ldots K) \propto \prod_{k=1}^{K} (w_k)^{\alpha - 1}. \quad (7)$$

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 and helps to stabilize the weights. Together with the prior, a Bayesian inference of the weights is made using the performance of leave-one-year-out cross-validation predictive densities, such that the posterior distribution of the weights is proportional to

$$A = \prod_{k=1}^{K} (w_k)^{\alpha - 1} \prod_{t=1}^{T} \sum_{k=1}^{K} w_k f_k^{(t)}(y^t \mid x_k^t), \quad (8)$$

where $f_k^{(t)}(y^t \mid x_k^t)$ is the cross-validation predictive density for event t, $t = 1, 2, \ldots, T$. This setup is different to that used by Raftery et al. (2005) and others, as the weights are based on forecasting performance, rather than from fitting performance, as would be the case if we used posterior predictive densities. This choice to use cross-validation predictive densities is justified as it is likely to be more robust, in the sense that it reduces the risk of overfitting.

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. 2012; 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.

c. Verification of ensemble forecasts of rainfall amount

In this study, we assess leave-one-year-out cross-validation forecasts. This means that the predictor and predictand data points corresponding to the year being

forecast are removed when inferring the parameters of the BJP models, and the cross-validation predictive density for the year being forecast is left out when calculating the BMA weights. Although the limited data available does not allow for proper verification in an independent period, by leaving one year out we can be reasonably confident that the forecast events are independent of the training data. We can therefore draw appropriate conclusions from the results, but with the caveat that the results will be subject to sampling variability.

The ensemble forecasts analyzed in this study are essentially probabilistic. Important attributes of probabilistic forecasts include accuracy, sharpness, and reliability. Accuracy refers to the agreement between the forecasts and the observations, reliability refers to the statistical consistency of forecast probabilities with observed events, and sharpness refers to the tendency to forecast more extreme probabilities. We assess these attributes in the paper by calculating scoring measures and visually assessing diagnostic graphs. Visual assessment of graphs is an effective way to identify any systemic calibration problems.

To assess forecast accuracy and sharpness, we calculate the continuous ranked probability score (CRPS) (Matheson and Winkler 1976). CRPS is given by

$$\mathrm{CRPS} = \frac{1}{T}\sum_{t=1}^{T}\int [F(y^t) - H(y^t - y_D^t)]^2\, dx, \qquad (9)$$

where $F(y^t)$ is the forecast CDF and H is the Heaviside step function such that

$$H(y^t - y_D^t) = \begin{cases} 0 & y^t < y_D^t \\ 1 & y^t \geq y_D^t \end{cases}. \qquad (10)$$

To better assess forecast value, CRPS skill scores are formulated as generalized skill scores, to measure the relative improvement of a set of forecasts over the corresponding climatology reference forecasts:

$$\mathrm{CRPS}_{\mathrm{SkillScore}} = \frac{\mathrm{CRPS}_{\mathrm{ref}} - \mathrm{CRPS}_{\mathrm{fcst}}}{\mathrm{CRPS}_{\mathrm{ref}}} \times 100. \qquad (11)$$

A skill score of 100 means perfect forecasts, while a skill score of 0 means that the forecasts are no better than using climatology, and thus considered of no skill. A negative skill score means that the forecasts are worse than using climatology. We note that both the CRPS score and the generalized skill score can be sensitive to small sample sizes.

Attributes diagrams (Hsu and Murphy 1986) are used to assess forecast reliability, sharpness, and resolution. The diagrams are more suitable for assessing large samples and, therefore, forecasts from multiple grid cells and seasons are pooled in their construction. Reliability is 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. To remove the effect of any bias, we analyze forecasts of the probability of exceeding the model's climatological median.

d. Data and description of four MME methods for comparison

We use six GCMs in this study. There are three variants of POAMA, versions 24A, B, and C (Wang et al. 2011). For this reason, POAMA is sometimes referred to as a pseudo-MME. We use all three variants in this study and treat them as separate models. POAMA hindcasts are sourced from the Climate-system Historical Forecast Project (CHFP) project (Kirtman and Pirani 2009). We also use three international models from the United Kingdom, the European Centre for Medium-Range Weather Forecasts (ECMWF) (EC), and Météo-France (MF) modeling centers. The three international models are sourced from the ENSEMBLES project. Details including model descriptions are given by Weisheimer et al. (2009). The convenience of sourcing data from the ENSEMBLES and CHFP projects is that the data have been consistently interpolated onto a regular 2.5° × 2.5° (latitude–longitude) grid by the data providers. The period of hindcasts overlap 1980–2005; therefore, we restrict our analysis to this period.

Observed seasonal rainfall totals are derived from the Australian Water Availability Project's (AWAP) 0.05° × 0.05° gridded dataset of monthly rainfall (Jones et al. 2009). The monthly data are averaged within 2.5° × 2.5° grid cells, corresponding to the GCM grid cells, and then aggregated to seasonal totals.

To demonstrate the need for forecast calibration, we analyze a selected grid cell located in the northeastern region of Australia and compare the climatologies of the forecasts from individual uncalibrated models with the climatology of observed September–November (SON) rainfall (Fig. 1). The model climatologies are created by pooling all ensemble members of all 1-month lead-time forecasts 1980–2005. The bars show the [0.25, 0.75] and [0.1, 0.9] quantile ranges of the climatologies and the dot shows the mean. It is clear that all of the raw GCM forecasts exhibit a negative bias for forecasting SON rainfall in the selected grid cell and season. There are also obvious differences in the ensemble spread between the GCMs; however, the individual GCMs all underestimate forecast uncertainty.

We evaluate our Bayesian calibration and MME approach for ensemble forecasts of Australian seasonal rainfalls at a lead time of 1 month for each 2.5° grid cell.

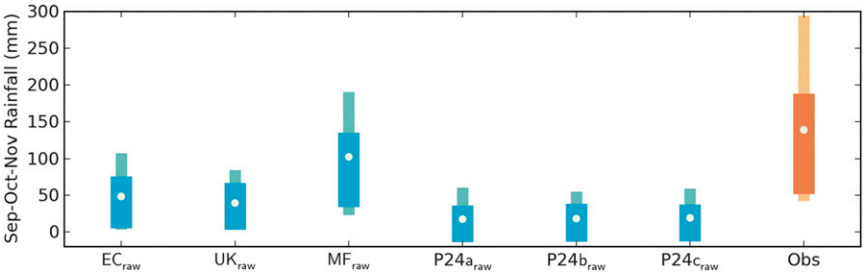

FIG. 1. Climatologies of raw GCM forecasts compared to observed climatology of SON rainfall for a grid cell in northeast Australia, derived from 1-month lead-time 1980–2005 forecasts. Box plots show [0.25, 0.75] and [0.1, 0.9] quantile ranges. Dot shows the mean.

The seasons studied are March–May (MAM), June–August (JJA), SON, and December–February (DJF) with each GCM initialized at the beginning of the prior month. As a baseline for comparison, we also evaluate a mean corrected superensemble method. The mean correction is to exclude the effect of overall bias and make the results more comparable with the Bayesian calibration and MME approach. In cross validation, the mean bias is calculated by excluding the data for the year being forecast.

To evaluate the advantage of using additional international models over using only POAMA, we compare results for the MME formed using all six models, calibrated with BJP and merged with BMA, with the results for the MME formed using the three POAMA models only, calibrated with BJP and merged with BMA. To evaluate the benefit of using unequal weights in BMA over using equal weights, we also include results for the MME formed using all six models, calibrated with BJP and weights set to equal. In total, we compare four MME methods, refer to Table 1.

3. Results with discussion

a. Efficacy of calibration for a single grid cell

In section 2d we demonstrated that individual GCMs were biased and underestimated forecast uncertainty for a single grid cell and season (northeastern Australia; SON). We analyze the same grid cell and season once more and compare the climatologies of the BJP-calibrated MME forecasts with the observed climatology (Fig. 2). The quantile ranges of the climatology of the mean corrected superensemble forecasts (method 1) are still considerably narrower than the observed climatology. Method 1 underestimates forecast uncertainty as a consequence of the individual GCMs all underestimating forecast uncertainty. Therein lays the motivation for a more sophisticated calibration approach. A simple mean corrected superensemble is not guaranteed to provide well calibrated forecasts. In contrast, the quantile ranges of the climatologies of the BJP calibrated forecasts (methods 2, 3, and 4) are far better aligned with the quantile ranges of the observed climatology. The BJP-calibrated MMEs provide more reliable forecast uncertainty estimates compared to the mean corrected superensemble approach.

The results found for this grid cell and season are typical. Our analysis of other locations and different seasons found that GCM ensemble spreads are most frequently too narrow; indicating that forecast uncertainty is usually underestimated. However, there are cases where ensemble spreads are too wide; indicating that forecast uncertainty can be overestimated (results not shown). The BJP calibration approach is applicable in each case.

b. Skill scores for all locations and seasons

We now view maps of the CRPS scores for each calibration and MME method (Fig. 3). The two numbers in each panel show the number of grid cells where CRPS skill scores of 5 and 10 are exceeded, respectively. The purpose of including the counts is to help discern the

TABLE 1. Description of the four MMEs studied. Each MME is given an alias for easy reference in the text.

Alias	Included models	Description
Method 1	P24A, P24B, P24C, UK, ECMWF, MF	Superensemble with mean correction; all models have equal weight
Method 2	P24A, P24B, P24C	Each model calibrated with BJP; model weights derived through BMA
Method 3	P24A, P24B, P24C, UK, ECMWF, MF	Each model calibrated with BJP; model weights derived through BMA
Method 4	P24A, P24B, P24C, UK, ECMWF, MF	Each model calibrated with BJP; all models assigned equal weights

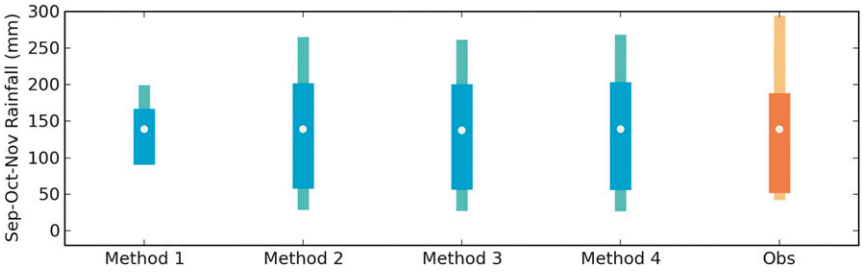

FIG. 2. Climatologies of calibrated MME forecasts compared to observed climatology of SON rainfall for a grid cell in northeast Australia, derived from 1-month lead-time 1980–2005 forecasts. Box plots show [0.25, 0.75] and [0.1, 0.9] quantile ranges. Dot shows the mean. Refer to Table 1 for a description of calibration methods 1–4.

coverage of higher skill, which may be difficult to discern by eye. We note that seasonal forecasting is highly challenging and there are many grid cells with little or no skill.

What is immediately clear from the skill maps is that the mean corrected superensemble approach (method 1) results in large areas of strongly negative skill. The mean corrected superensemble approach may perform poorly

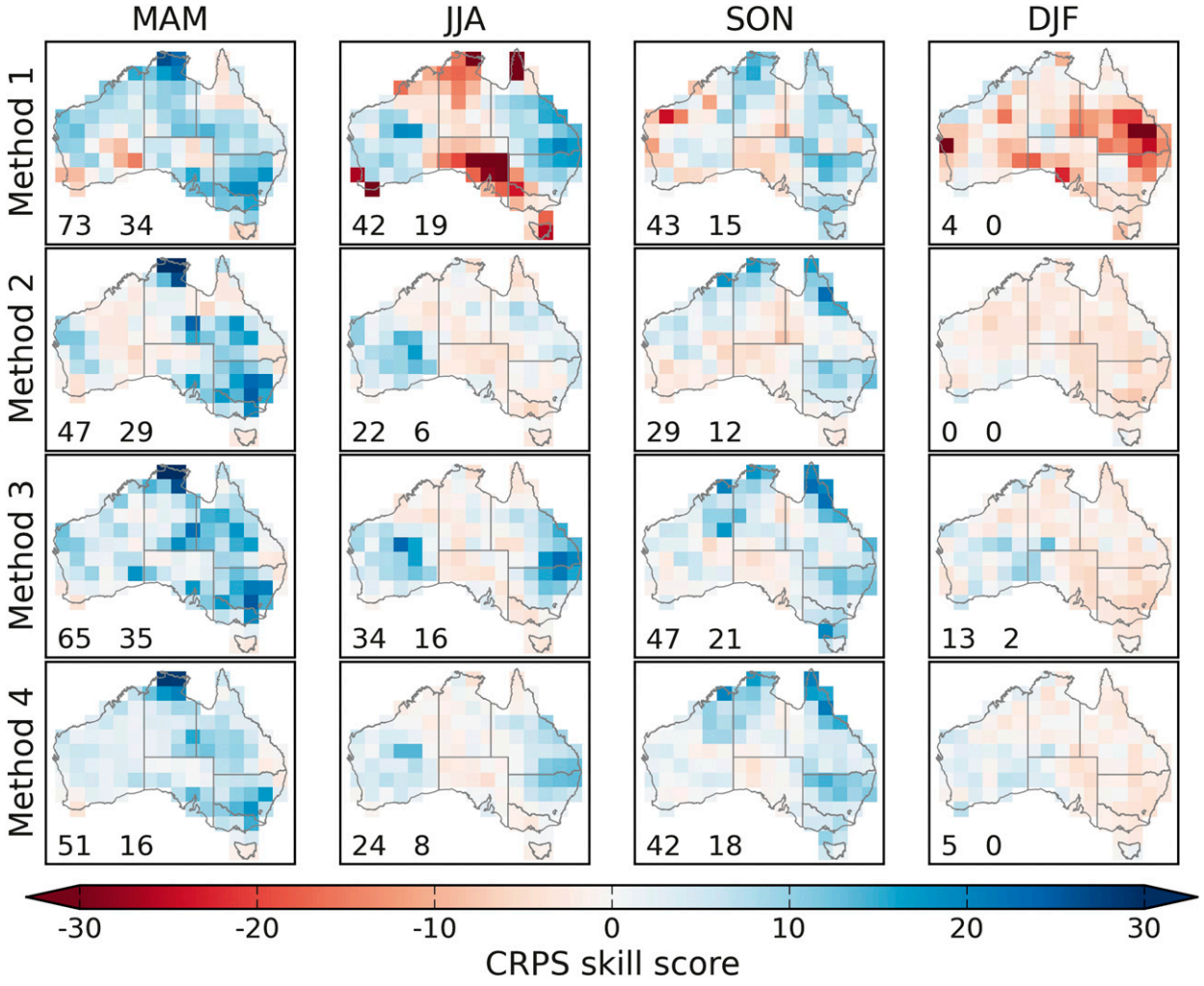

FIG. 3. Maps of cross-validated (1980–2005) continuous ranked probability score (CRPS) skill scores for forecasting Australian seasonal rainfalls at a lead time of 1 month using four different calibration and forecast merging methods. Refer to Table 1 for a description of calibration methods 1–4.

when there is no real relationship between the forecast ensemble members and observed rainfall. In areas where there is no real relationship, a climatological forecast would be more appropriate. We do not observe large areas of strongly negative skill scores in the results for the BJP-calibrated MME forecasts (methods 2, 3, and 4). In this respect, the BJP-calibrated MME forecasts represent a significant improvement over the mean corrected superensemble approach. As described in the methods, the BJP-calibrated models are designed to produce climatological forecasts in the absence of a relationship between the predictor and predictand. We do expect small negative skill scores in some cases due to cross validating with a small number of samples. It should be noted that results (not shown here) for superensembles without mean corrections are much worse than the mean corrected superensembles.

By comparing method 2 and method 3 in Fig. 3, we analyze the benefit of including the three international GCMs in the MME on top of using the locally developed POAMA GCM alone. In all seasons, method 3 has a higher number of grid cells exceeding CRPS skill scores of 5 and 10 than method 2. Merging POAMA forecasts with the forecasts from the three international models results in improvements in the coverage of skill compared to using the locally developed model alone. For an obvious example, we compare the skill of JJA forecasts in eastern Australia. Method 3 shows much improved skill for forecasting JJA rainfall in eastern Australia compared to method 2.

By comparing methods 3 and 4 in Fig. 3, we now analyze the benefit of determining BMA weights over assigning equal weights to the BJP-calibrated models. Although the weighted forecasts (method 3) and equal-weights forecasts (method 4) exhibit similar spatial patterns of skill scores for each season, method 3 tends to produce marginally higher skill scores; in each season we observe that method 3 produces a higher number of grid cells that exceed CRPS skill scores of 5% and 10% than method 4. Using BMA to weight and merge forecasts results in CRPS skill score improvements of up to 10%. Because we are merging forecasts from GCMs that have similar fundamentals, the differences in skill scores between methods 3 and 4 are overall small.

c. Overall reliability of above-median forecasts

Attributes diagrams for all methods are contained in Fig. 4. These diagrams have been formulated for forecasts of the probability of exceeding the model climatological median. In the construction of the attributes diagram, all forecasts (i.e., for all grid cells and all four seasons), have been pooled. The relative sizes of the dots show the proportion of forecasts in each forecast probability bin and are therefore representative of forecast sharpness. Reliability is visually assessed by fidelity of the points to the 1:1 line. The forecasts are generally considered to be reliable if the center of the points lie in the shaded region.

Consider reliability first. For all Bayesian calibration and MME methods (methods 2, 3, and 4), all of the points lie within the shaded area. In the case of method 3, the points lie almost perfectly along the 1:1 line, indicating that overall the forecast probabilities are very reliable. In the case of method 4, although the points lie acceptably close to the 1:1 line, the forecasts may marginally overestimate forecast uncertainty because of each GCM being calibrated independently. The BJP-calibrated MME forecasts show much improved reliability compared to the mean corrected superensemble forecasts (method 1), which tend to be too emphatic, as highlighted by points falling outside the shaded area.

Consider now, sharpness. Although method 1 forecasts are the sharpest, we have already deduced that they are not sufficiently reliable (and accurate). Method 3 forecasts are marginally sharper than the method 2 forecasts, although significant differences are difficult to detect. Method 3 forecasts are noticeably sharper than method 4 forecasts. In the case of method 3, approximately 71% of forecast probabilities are in the 0.4–0.6 range. In the case of method 4, 87% of forecast probabilities are in the 0.4–0.6 range. More emphatic forecast probabilities are arguably of a higher importance to seasonal forecasts users as there are enhanced risks and rewards attached to the use of such forecasts. By this reasoning, a higher concentration of forecasts in the more extreme ranges is desirable, provided the forecasts are reliable. From the point of view of forecast sharpness, the method 3 forecasts are potentially more valuable to users. To summarize the attributes diagrams, the weighted forecasts (method 3) have the overall most appealing reliability and sharpness attributes.

4. Supplementary results and further discussion

As described in the methodology section, our BMA approach to MME implements a prior that constrains the weights to be more equal. Additionally, the algorithm is started from an equal-weights position. It is by careful design that the method will not disproportionately heap weight on one particular model. The approach will shift weights in favor of one or more models when the data support it, but be constrained by the prior. To demonstrate this, we show in Fig. 5 the BMA weights of each model (method 3), for a single grid cell and season, and for each year in the cross validation. We select the same grid cell in northeastern Australia as analyzed in

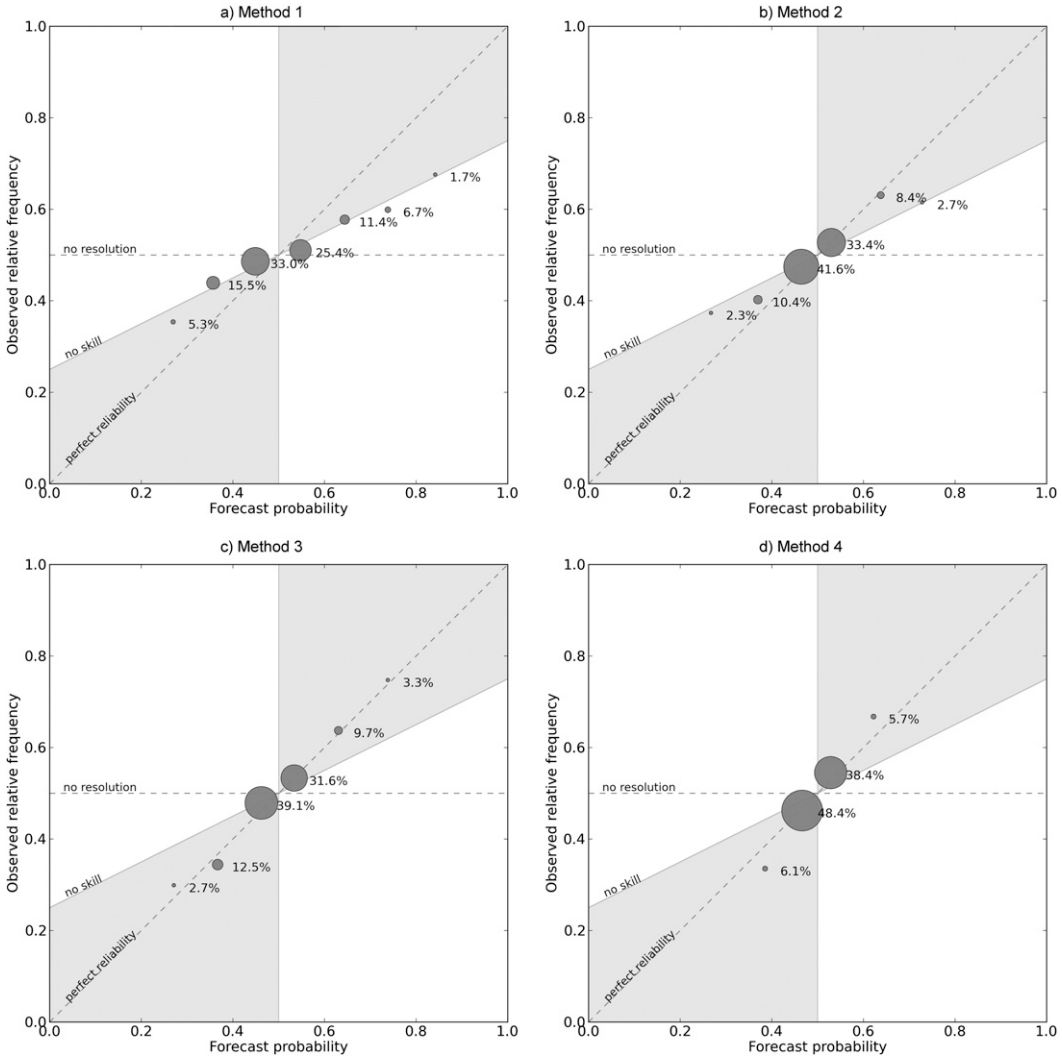

FIG. 4. Attributes diagrams for forecasts of the probability of exceeding of the climatological median of Australian seasonal rainfalls at a lead time of 1 month using four different calibration and forecast merging methods. Forecast sharpness is depicted by the size of the dots, which shows the proportion of forecasts in each forecast probability bin. The proportions are also shown as labels near the dots. Reliability is assessed by fidelity the 1:1 line. Refer to Table 1 for a description of calibration methods 1–4.

sections 2d and 3a. In this case, the P24C model is assigned about 40% of the weight, the U.K. model is assigned about 30% of the weight, and the remaining 30% is split fairly evenly among the remaining models. When applying our BMA method to merge forecasts, we chose an α_0 value of 1.0 (see section 2b). Indeed, if a modeler has a strong prior belief that the model weights should be more equal, the approach caters to this belief by allowing the modeler to increase the value of α_0 to increase the pressure on the weights to be more equal. Conversely, a modeler could reduce α_0 to relax the pressure on the weights to be more equal. We speculate that there will be applications where unequal weights are desirable. For example in large ensembles it can save the modeler the effort of weeding out poor models. Additionally, if the modeler desires to include models with a fundamentally different basis (e.g., statistical) in the mix, then unequal weighs may be more appropriate.

As stated in the introduction, our aim is to obtain ensemble seasonal rainfall forecasts of rainfall amount that are as accurate as possible and statistically reliable. Our Bayesian calibration and MME approach shows promise for achieving this goal. Since our approach is effectively a postprocessing approach, we note that the GCM development community continues to tackle the fundamental problems that cause over or underestimation

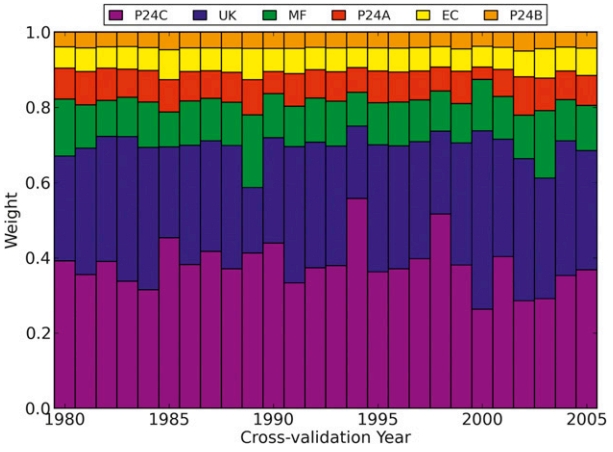

FIG. 5. Weights of each model derived through BMA (method 3) for a grid cell in northeastern Australia; SON forecasts. Each stacked bar represents a year in the leave-one-year-out cross validation (1980–2005).

of forecast uncertainty by, for example, improving the model initial conditions (e.g., Marshall et al. 2011) and using stochastic parameterizations (Weisheimer et al. 2011).

5. Summary and conclusions

Coupled ocean–atmosphere general circulation models (coupled GCMs) are increasingly being used in seasonal forecasting centers around the world. In this study, we set out to use GCMs to obtain ensemble seasonal rainfall forecasts of Australian seasonal rainfall amount that are as accurate as possible and statistically reliable. Raw GCM rainfall forecasts are typically not accurate in terms of rainfall amount and tend to be too narrow in ensemble spread, resulting in underestimated forecast uncertainty and thus statistically unreliable forecast probabilities. We identify forecast calibration and forecast merging as techniques to address these problems.

In this study, we analyze a two-step Bayesian approach to forecast calibration and forecast merging. First, forecast calibration of the individual GCMs is carried out by using Bayesian joint probability models that account for parameter uncertainty. Second, the individually calibrated forecasts of the GCMs are merged through Bayesian model averaging (BMA). As a baseline comparison, we also evaluate a mean corrected superensemble method. We also compare results of BJP-calibrated forecasts merged using BMA weights with forecasts merged using equal weights. We analyze forecasts of MAM, JJA, SON, and DJF rainfalls at a lead time of 1 month on a 2.5° grid.

At the gridcell scale, the climatologies of the BJP-calibrated forecasts tend to be well aligned with the observed climatologies. In contrast, the mean corrected superensemble approach can easily result in forecast climatologies that are still too wide or too narrow. As a result, the BJP-calibrated MMEs provide consistently more reliable forecast uncertainty estimates compared to the mean corrected superensemble approach.

At the continental scale, the mean corrected superensemble approach results in large areas of strongly negative CRPS skill scores. We do not observe large areas of strongly negative skill scores in the results for the BJP-calibrated MME forecasts. In this respect, the BJP-calibrated MME forecasts are a significant improvement over the mean corrected superensemble forecasts. The BJP-calibration models are designed to produce climatological forecasts in the absence of a relationship between the predictor and predictand, a design feature that is absent in the mean corrected superensemble approach.

Merging POAMA forecasts with the forecasts from the three internationally developed models results in improvements in the regional and seasonal coverage of positive CRPS skill scores compared to using the locally developed model alone. Additionally, using BMA weights to merge the forecasts results in marginal improvements in skill compared to using equal weights to merge the forecasts. The use of BMA weights is likely to have more impact when merging forecasts from models with a different fundamental basis.

In this study, the attributes of reliability and forecast sharpness are assessed for all forecasts (all grid cells and seasons) expressed in terms of the probability of exceeding the climatological median. The BJP-calibrated MME forecasts show much improved reliability compared to the mean corrected superensemble forecasts. The BMA-weighted MME forecasts and the equal-weighted MME forecasts are similarly reliable. However, the BMA-weighted forecasts are noticeably sharper compared to the equal weights forecasts. Therefore, the weighted forecasts are likely to be of higher value to decision makers.

Acknowledgments. The ENSEMBLES data used in this work were funded by the EU FP6 Integrated Project ENSEMBLES (Contract 505539) whose support is gratefully acknowledged. We acknowledge the WCRP/CLIVAR Working Group on Seasonal-to-Interannual Prediction (WGSIP) for establishing the Climate-system Historical Forecast Project (CHFP; see Kirtman and Pirani 2009) and the Centro de Investigaciones del Mar y la Atmosfera (CIMA) for providing the model output (http://chfps.cima.fcen.uba.ar/). We also thank the data providers for making the model output available through CHFP. We appreciate the constructive and thoughtful

comments and suggestions from two anonymous reviewers that helped improve this manuscript.

REFERENCES

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

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.

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

Hoeting, J. A., D. Madigan, A. E. Raftery, and C. T. Volinsky, 1999: Bayesian model averaging: A tutorial. *Stat. Sci.,* **14,** 382–401.

Hsu, W., and A. H. Murphy, 1986: The attributes diagram: A geometrical framework for assessing the quality of probability forecasts. *Int. J. Forecasting,* **2,** 285–293.

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

Kirtman, B., and A. Pirani, 2009: The state of the art of seasonal prediction: Outcomes and recommendations from the first World Climate Research Program workshop on seasonal prediction. *Bull. Amer. Meteor. Soc.,* **90,** 455–458.

Krishnamurti, T. N., C. M. Kishtawal, T. E. LaRow, D. R. Bachiochi, Z. Zhang, C. E. Williford, S. Gadgil, and S. Surendran, 1999: Improved weather and seasonal climate forecasts from multimodel superensemble. *Science,* **285,** 1548–1550.

Langford, S., and H. H. Hendon, 2013: Improving reliability of coupled model forecasts of Australian seasonal rainfall. *Mon. Wea. Rev.,* **141,** 728–741.

Lim, E.-P., H. H. Hendon, 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.

Luo, L., E. F. Wood, and M. Pan, 2007: Bayesian merging of multiple climate model forecasts for seasonal hydrological predictions. *J. Geophys. Res.,* **112,** D10102, doi:10.1029/2006JD007655.

Marshall, A. G., D. Hudson, M. C. Wheeler, H. H. Hendon, and O. Alves, 2011: Evaluating key drivers of Australian intraseasonal climate variability in POAMA-2: A progress report. *CAWCR Res. Lett.,* **7,** 10–16.

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

Palmer, T. N., Č. Branković, and D. S. Richardson, 2000: A probability and decision-model analysis of PROVOST seasonal multi-model ensemble integrations. *Quart. J. Roy. Meteor. Soc.,* **126,** 2013–2033.

——, and Coauthors, 2004: Development of a European Multimodel Ensemble System for Seasonal-to-Interannual Prediction (DEMETER). *Bull. Amer. Meteor. Soc.,* **85,** 853–872.

Ploshay, J., and J. Anderson, 2002: Large sensitivity to initial conditions in seasonal predictions with a coupled ocean-atmosphere general circulation model. *Geophys. Res. Lett.,* **29,** 1262, doi:10.1029/2000GL012710.

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.

Saha, S., and Coauthors, 2006: The NCEP Climate Forecast System. *J. Climate,* **19,** 3483–3517.

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

Wang, G., D. Hudson, Y. Ying, O. Alves, H. Hendon, S. Langford, G. Liu, and F. Tseitkin, 2011: POAMA-2 SST skill assessment and beyond. *CAWCR Res. Lett.,* **6,** 40–46.

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.

——, A. Schepen, and D. E. Robertson, 2012: Merging seasonal rainfall forecasts from multiple statistical models through Bayesian model averaging. *J. Climate,* **25,** 5524–5537.

Weisheimer, A., and Coauthors, 2009: ENSEMBLES: A new multi-model ensemble for seasonal-to-annual predictions—Skill and progress beyond DEMETER in forecasting tropical Pacific SSTs. *Geophys. Res. Lett.,* **36,** L21711, doi:10.1029/2009GL040896.

——, T. N. Palmer, and F. J. Doblas-Reyes, 2011: Assessment of representations of model uncertainty in monthly and seasonal forecast ensembles. *Geophys. Res. Lett.,* **38,** L16703, doi:10.1029/2011GL048123.

Yasuda, T., Y. Takaya, C. Kobayashi, M. Kamachi, H. Kamahori, and T. Ose, 2007: Asian monsoon predictability in JMA/MRI seasonal forecast system. *CLIVAR Exchanges,* No. 43, International CLIVAR Project Office, Southampton, United Kingdom, 18–24.

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

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