Merging Seasonal Rainfall Forecasts from Multiple Statistical Models through Bayesian Model Averaging

Q. J. WANG

CSIRO Land and Water, Highett, Victoria, Australia

ANDREW SCHEPEN

Bureau of Meteorology, Brisbane, Queensland, Australia

DAVID E. ROBERTSON

CSIRO Land and Water, Highett, Victoria, Australia

(Manuscript received 3 July 2011, in final form 20 March 2012)

ABSTRACT

Merging forecasts from multiple models has the potential to combine the strengths of individual models and to better represent forecast uncertainty than the use of a single model. This study develops a Bayesian model averaging (BMA) method for merging forecasts from multiple models, giving greater weights to better performing models. The study aims for a BMA method that is capable of producing relatively stable weights in the presence of significant sampling variability, leading to robust forecasts for future events. The BMA method is applied to merge forecasts from multiple statistical models for seasonal rainfall forecasts over Australia using climate indices as predictors. It is shown that the fully merged forecasts effectively combine the best skills of the models to maximize the spatial coverage of positive skill. Overall, the skill is low for the first half of the year but more positive for the second half of the year. Models in the Pacific group contribute the most skill, and models in the Indian and extratropical groups also produce useful and sometimes distinct skills. The fully merged probabilistic forecasts are found to be reliable in representing forecast uncertainty spread. The forecast skill holds well when forecast lead time is increased from 0 to 1 month. The BMA method outperforms the approach of using a model with two fixed predictors chosen a priori and the approach of selecting the best model based on predictive performance.

1. Introduction

Seasonal climate forecasts are in high demand in Australia, and seasonal rainfall forecasts in particular are sought by farmers, water managers, and others throughout the year. The Australian Bureau of Meteorology provides probabilistic forecasts of seasonal rainfall using a statistical prediction system based on sea surface temperature (SST) anomaly patterns over the Indian and Pacific Oceans (Drosdowsky and Chambers 2001; Fawcett et al. 2005). Different forecast models are used for the 12 overlapping seasons, but the predictors are held fixed for the whole of Australia. To increase the benefit that seasonal rainfall forecasts can have to decision makers, the skill of the forecasts needs to be improved. One possible strategy to improve statistical forecasts is to allow different models to be used for different parts of Australia as well as different seasons, and to consider a range of other potentially useful predictors.

Schepen et al. (2012) investigated a range of lagged climate indices for their potential usefulness as predictors of seasonal rainfall in Australia, and they found statistical evidence for their use in some regions and seasons. Often, more than one climate index was identified as a useful predictor for a particular region and season. This is not surprising, as some of the climate indices are closely related to each other. Another reason for this is that different climate indices can represent different aspects of

Corresponding author address: Dr. Q. J. Wang, CSIRO Land and Water, P.O. Box 56, Highett VIC 3190, Australia.
E-mail: qj.wang@csiro.au

DOI: 10.1175/JCLI-D-11-00386.1

the large-scale circulation around Australia, which may act independently to influence rainfall across common regions and seasons.

One approach to seasonal rainfall forecasting is to ascertain and adopt the predictor with the highest supporting statistical evidence, based on historical data, for each season and location. A drawback to this approach is the heightened chance of choosing, because of data noise, a predictor that has no real or only a weak underlying relationship with seasonal rainfall. Robertson and Wang (2012) suggest that the best predictor is adopted only when its supporting evidence exceeds a certain threshold. However, when many climate indices show similar supporting evidence for forecasting seasonal rainfall, it can become contentious to adopt one predictor over another because models using predictors with similar supporting evidence can sometimes produce very different forecasts for a given event. This highlights the classical issue of model uncertainty in forecasting.

Various studies have shown the advantage of merging forecasts from multiple models for climate forecasting, be they statistical or dynamical models (e.g., Casey 1995; Coelho et al. 2004; Luo et al. 2007; Stephenson et al. 2005). The Bayesian model averaging (BMA) approach (Hoeting et al. 1999; Raftery et al. 1997) is highly suitable for merging forecasts from multiple models. BMA gives greater weights to better performing models while allowing for model uncertainty. BMA has been shown to outperform simple skill-based averaging in seasonal forecasting (Casanova and Ahrens 2009), and to perform well in forecasting and predictions in general (e.g., Raftery et al. 2005; Wright 2009).

In this study, we develop a BMA method for merging forecasts from multiple models. We aim for a BMA method that is capable of producing relatively stable weights in presence of significant sampling variability, leading to robust forecasts for future events. In developing the BMA weights, we adopt a mixture model approach (e.g., Raftery et al. 2005) rather than the classical model posterior probability approach (e.g., Hoeting et al. 1999). A number of empirical studies from the machine learning literature suggest that the latter approach is often outperformed in predictive performance by more ad hoc methods of model combination, such as bagging and boosting (Clarke 2003; Domingos 2000; Minka 2000; Monteith et al. 2011). The mixture model approach is likely to be more competitive because the BMA weights are inferred directly from evaluation of predictive performance of the combined model. It also has the advantage that a highly efficient expectation–maximization (EM) algorithm exists for numerical solution (Cheng et al. 2006; Zivkovic and van der Heijden 2004). In our formulation of Bayesian inference of the mixture model, we apply a prior that gives a slight preference toward more evenly distributed weights, so that the merged forecasts err on the side of model consensus when there is large uncertainty about the optimum weights. We also use the cross-validation likelihood function (e.g., Shinozaki et al. 2010) rather than the classical likelihood function, so that the weights reflect the predictive, rather than fitting, abilities of the individual models.

We apply the BMA method to merge forecasts from multiple statistical models using climate indices as predictors. We assess through cross validation the skill of the merged forecasts and examine the contributions from the Pacific, Indian, and extratropical groups of models. We compare the BMA method with two alternative methods: using models with two fixed predictors selected a priori and using the best model in terms of pseudo-Bayes factor (PsBF) as the statistical supporting evidence.

In section 2, we present the climate indices and rainfall data to be used. In section 3, we describe the BMA method in detail and outline methods for assessing the skill and reliability of probabilistic forecasts. In section 4, we present figures and maps showing the skill and reliability of the forecasts. In section 5, we present some further results and discussion. Section 6 completes the paper with a summary and conclusions.

2. Data

a. *Climate indices*

Multiple forecasting models are established based on 11 monthly climate indices with lags of 1–3 months. Table 1 summarizes the 11 climate indices (including their abbreviations) and provides a brief description of each. More details on the climate indices are given in Schepen et al. (2012). The period of record used for each climate index is 1950–2009. To ease the interpretation of the results, the climate indices are grouped into Pacific, Indian, and extratropical, based on the regions from which they are derived (Table 1).

b. *Rainfall data*

The rainfall data used in this study are derived from the Australian Water Availability Project's (AWAP) 0.05° × 0.05° gridded dataset of monthly rainfall for the period 1950–2009 (Jones et al. 2009). Through simple areal averaging, the data have been upscaled to a 2.5° × 2.5° grid aligned with the grid used by the Predictive Ocean Atmosphere Model for Australia (POAMA), a dynamical prediction system used by the Australian Bureau of Meteorology. There are 122 grid cells covering Australia.

TABLE 1. Climate indices used as predictors of Australian seasonal rainfall.

Climate index	Description[a]	Group
Southern Oscillation index[b] (SOI)	Pressure difference between Tahiti and Darwin as defined by Troup (1965)	Pacific
Niño-3	Average SST anomaly over 5°N–5°S, 150°–90°W	Pacific
Niño-3.4 (Nino-34)	Average SST anomaly over 5°N–5°S, 170°–120°W	Pacific
Niño-4	Average SST anomaly over 5°N–5°S, 150°–160°E	Pacific
EMI	C – 0.5 (E + W), where the components are average SST 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	Pacific
Indian Ocean west pole index (WPI)	Average SST anomaly over 10°N–10°S, 50°E–70°E	Indian
Indian Ocean east pole index (EPI)	Average SST anomaly over 0°N–10°S, 90°–110°E	Indian
DMI	WPI − EPI	Indian
II	Average SST anomaly over 0°N–10°S, 120°–130°E	Indian
Tasman Sea index (TSI)	Average SST anomaly over 30°–40°S, 150°–160°E	Extratropical
140°E blocking index[c] (B140)	$0.5(U[25] + U[30] - U[40] - 2U[45] - U[50] + U[55] + U[60])$, where $U[X]$ is the 500-hPa zonal wind at latitude X	Extratropical

[a] Climate indices based on SST are derived from the National Center for Atmospheric Research (NCAR) Extended Reconstructed Sea Surface Temperature, version 3 (Smith et al. 2008).

[b] The Southern Oscillation index is sourced from the Australian Bureau of Meteorology.

[c] The blocking index is derived from the National Centers for Environmental Prediction (NCEP)–NCAR reanalysis zonal wind data (Kalnay et al. 1996).

3. Method

a. Statistical forecast models and predictive densities

We seek to forecast the next 3-month rainfall total at the start of each month for each of the 122 grid cells covering Australia. Statistical models are established for each season and each grid cell independently of the others. For each season and grid cell, the models included in this study are 33 single-predictor models plus one no-predictor (or climatology) model. The predictors are the 11 climate indices with lags of 1–3 months.

A Bayesian joint probability (BJP) modeling approach is used to establish the individual models. Only a brief description of the BJP modeling approach is provided here, but more details can be found in Wang et al. (2009), Wang and Robertson (2011), and Robertson and Wang (2012). For ease of reference to previous papers, we retain the description of the BJP formulation for multiple predictands, even though the models used here are all single-predictand models.

Given a predictor vector $\mathbf{y}(1)$ and a predictand vector $\mathbf{y}(2)$, Yeo–Johnson transforms (Yeo and Johnson 2000) are applied to each of the predictor and predictand variables. The transformed predictor and predictand variables are assumed to follow a joint multivariate normal distribution. A Bayesian inference of the Yeo–Johnson transform parameters and the distribution parameters is made by using historical data. The inference is numerically implemented through Markov chain Monte Carlo (MCMC) sampling. The posterior predictive density for a new event is given by

$$f[\mathbf{y}(2)\,|\,\mathbf{y}(1)] = p[\mathbf{y}(2)\,|\,\mathbf{y}(1); \mathbf{Y}_{\mathrm{OBS}}, M] = \int p[\mathbf{y}(2)\,|\,\mathbf{y}(1), \boldsymbol{\theta}, M]\,p(\boldsymbol{\theta}\,|\,\mathbf{Y}_{\mathrm{OBS}}, M)\,d\boldsymbol{\theta}, \tag{1}$$

where M denotes the model used, $\mathbf{Y}_{\mathrm{OBS}}$ contains the historical data of both predictor and predictand variables used for model inference, and $\boldsymbol{\theta}$ is the parameter vector.

Wang and Robertson (2011) and Robertson and Wang (2012) also describe the handling of variables that are bounded by zero from below, but to keep the description simple here, formulations are presented without reference to this problem.

b. BMA for merging forecasts from multiple models

BMA is a technique for combining multiple models, giving greater weights to better performing models while allowing for model uncertainty. In the introduction, we outline our broad approach to BMA and the motivation behind it. Here we describe the detailed method.

Given a group of models M_k, $k = 1, \ldots, K$, the predictive density after BMA is given by a weighted average of the individual model predictive densities, as follows:

$$f_{\mathrm{BMA}}[\mathbf{y}(2)\,|\,\mathbf{y}(1)] = \sum_{k=1}^{K} w_k f_k[\mathbf{y}(2)\,|\,\mathbf{y}(1)] = \sum_{k=1}^{K} w_k p[\mathbf{y}(2)\,|\,\mathbf{y}(1); \mathbf{Y}_{\mathrm{OBS}}, M_k]. \quad (2)$$

To determine the weights w_k, $k = 1, \ldots, K$, we write the posterior distribution of the weights given $t = 1, 2, \ldots, T$ events in $\mathbf{Y}_{\mathrm{OBS}}$ and the predictive density functions of the individual models, as follows:

$$p(w_k, k = 1, \ldots, K\,|\,\mathbf{Y}_{\mathrm{OBS}};\ f_k[\mathbf{y}(2)\,|\,\mathbf{y}(1)], k = 1, \ldots, K) \propto p(w_k,\ k = 1, \ldots, K)\prod_{t=1}^{T} f_{\mathrm{BMA}}[\mathbf{y}^t_{\mathrm{OBS}}(2)\,|\,\mathbf{y}^t_{\mathrm{OBS}}(1)]$$

$$\propto p(w_k,\ k = 1, \ldots, K)\prod_{t=1}^{T}\sum_{k=1}^{K} w_k f_k[\mathbf{y}^t_{\mathrm{OBS}}(2)\,|\,\mathbf{y}^t_{\mathrm{OBS}}(1)], \quad (3)$$

where $p(w_k, k = 1, \ldots, K)$ is a prior of the weights, the remaining term on the right-hand side is the likelihood function, and $\mathbf{y}^t_{\mathrm{OBS}}(1)$ and $\mathbf{y}^t_{\mathrm{OBS}}(2)$ are the predictor and predictand values for event t, respectively.

We use a symmetric Dirichlet distribution prior, given as

$$p(w_k, k = 1, \ldots, K) \propto \prod_{k=1}^{K}(w_k)^{\alpha-1}, \quad (4)$$

where α is the concentration parameter. More evenly distributed weights among the models are encouraged when $\alpha > 1$, and the opposite is true when $\alpha < 1$. We choose to have α slightly over 1 and more specifically, $\alpha = 1 + a/K$ with $a = 0.5$. Such a prior helps stabilizing the weights in presence of significant sampling variability.

A point estimate of the weights may then be obtained by maximizing the posterior distribution of the weights. This is the maximum a posterior (MAP) solution. In this study, we make a change to the MAP method, following the approach of Shinozaki et al. (2010). We replace the likelihood function in Eq. (3) with a cross-validation likelihood function (Rust and Schmittlein 1985; Shinozaki et al. 2010; Smyth 1996, 2000; Stone 1977). More specifically, the posterior predictive densities $f_k[\mathbf{y}^t_{\mathrm{OBS}}(2)\,|\,\mathbf{y}^t_{\mathrm{OBS}}(1)]$ in Eq. (3) are replaced by the corresponding cross-validation predictive densities

$$f_k^{(t)}[\mathbf{y}^t_{\mathrm{OBS}}(2)\,|\,\mathbf{y}^t_{\mathrm{OBS}}(1)] = p[\mathbf{y}^t_{\mathrm{OBS}}(2)\,|\,\mathbf{y}^t_{\mathrm{OBS}}(1); \mathbf{Y}^{(t)}_{\mathrm{OBS}}, M_k], \quad (5)$$

where $\mathbf{Y}^{(t)}_{\mathrm{OBS}}$ is a matrix containing observed values of predictor and predictand variables for all the events except event t. The right-hand side of Eq. (3) becomes

$$A = \prod_{k=1}^{K}(w_k)^{\alpha-1}\prod_{t=1}^{T}\sum_{k=1}^{K} w_k f_k^{(t)}[\mathbf{y}^t_{\mathrm{OBS}}(2)\,|\,\mathbf{y}^t_{\mathrm{OBS}}(1)]. \quad (6)$$

A point estimate of the weights is then obtained by maximizing A.

In using the cross-validation likelihood function instead of the classical likelihood function, the weights are assigned according to the model predictive abilities rather than fitting abilities. Indeed, there is much literature in support of using predictive performance measures for model choice and combination based on the idea that a model is only as good as its predictions (e.g., Eklund and Karlsson 2007; Geweke and Whiteman 2006; Jackson et al. 2009).

Maximization solution of the weights can be found by using an iterative EM algorithm (Cheng et al. 2006; Zivkovic and van der Heijden 2004). The algorithm is as follows. Given weights $w_k^{\{j\}}, k = 1, \ldots, K$ at iteration j, first, ownerships are calculated by

$$O_k^{t,\{j+1\}} = \frac{w_k^{\{j\}} f_k^{(t)}[\mathbf{y}^t_{\mathrm{OBS}}(2)\,|\,\mathbf{y}^t_{\mathrm{OBS}}(1)]}{\sum_{m=1}^{K} w_m^{\{j\}} f_m^{(t)}[\mathbf{y}^t_{\mathrm{OBS}}(2)\,|\,\mathbf{y}^t_{\mathrm{OBS}}(1)]} \quad (7)$$

for all t and k. Then, new weights are calculated by

$$w_k^{\{j+1\}} = \frac{(1/T)\sum_{t=1}^{T} O_k^{t,\{j+1\}} + (\alpha - 1)/T}{1 + K(\alpha - 1)/T}. \quad (8)$$

Equations (7) and (8) are iteratively applied until the value of $\ln(A)$ converges. In our study, we use equal weights for $w_k^{\{0\}}$ to start the first iteration.

c. *Forecast assessment*

In this study, we assess the performance of forecasts using data from 1950 to 2009. For each of the forecast events, a leave-one-out cross-validation forecast is obtained from each of the models. The BMA merged forecast for a particular event is then produced using weights derived

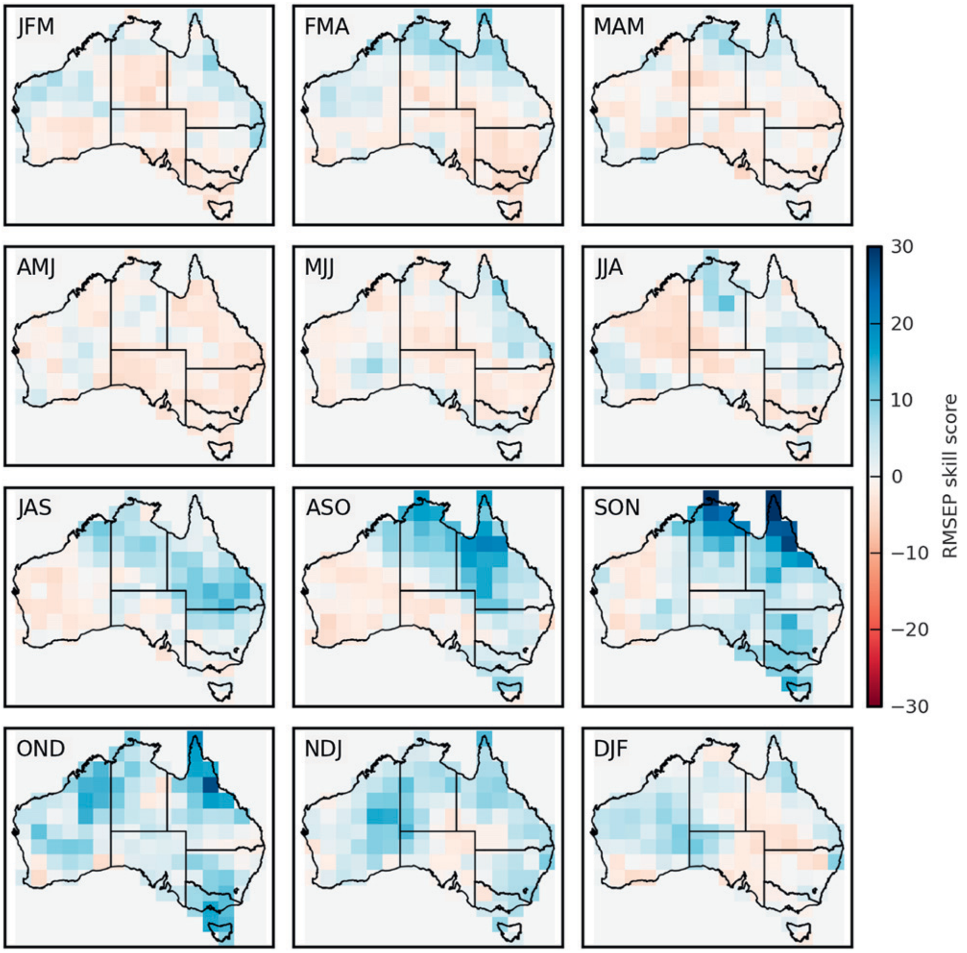

FIG. 1. Skill of fully merged BMA forecasts of the next season rainfall.

from the cross-validation predictive densities of individual models for all the events except the event being forecast. The procedure is repeated to produce BMA merged forecasts for all the events from 1950 to 2009. The forecasts so obtained are then assessed to give an indication of the likely forecast performance for future independent events.

Forecasts from the BMA method are probabilistic. Two of the most important attributes of probabilistic forecasts are accuracy and reliability. Forecasters aim to produce forecasts that are as accurate as possible. However for a given level of accuracy achievable, the forecast probability distributions should reliably reflect the uncertainty of the forecasts. We assess both the accuracy and reliability of the merged forecasts.

For forecast accuracy, we use the root-mean-squared error in probability (RMSEP) skill score to assess forecast medians, using the climatological median as a reference forecast (Wang and Robertson 2011). The reference forecasts used are the corresponding cross-validation climatologies. A skill score of 100% means perfect forecasts, while a skill score of 0 means that the forecasts are no better than using the climatological median, and thus considered of no skill. We also assess forecast accuracy in terms of mean-squared error (MSE) of forecast mean and of continuous ranked probability score (CRPS) of forecast distributions. Conclusions drawn from the results of MSE- and CRPS-based skill scores are consistent with RMSEP, and therefore are not presented in this paper.

For forecast reliability, we use reliability diagrams by plotting the probability of a threshold forecast against the observed relative frequency of such an event (Wilks 1995). Reliability diagrams show how well the predicted probabilities of events correspond to their observed frequencies. In this study, we report the reliability of forecast probabilities of events not exceeding the 33.3, 50, and 66.7 percentiles of climatology. As with skill assessment, cross-validation climatologies are used here too.

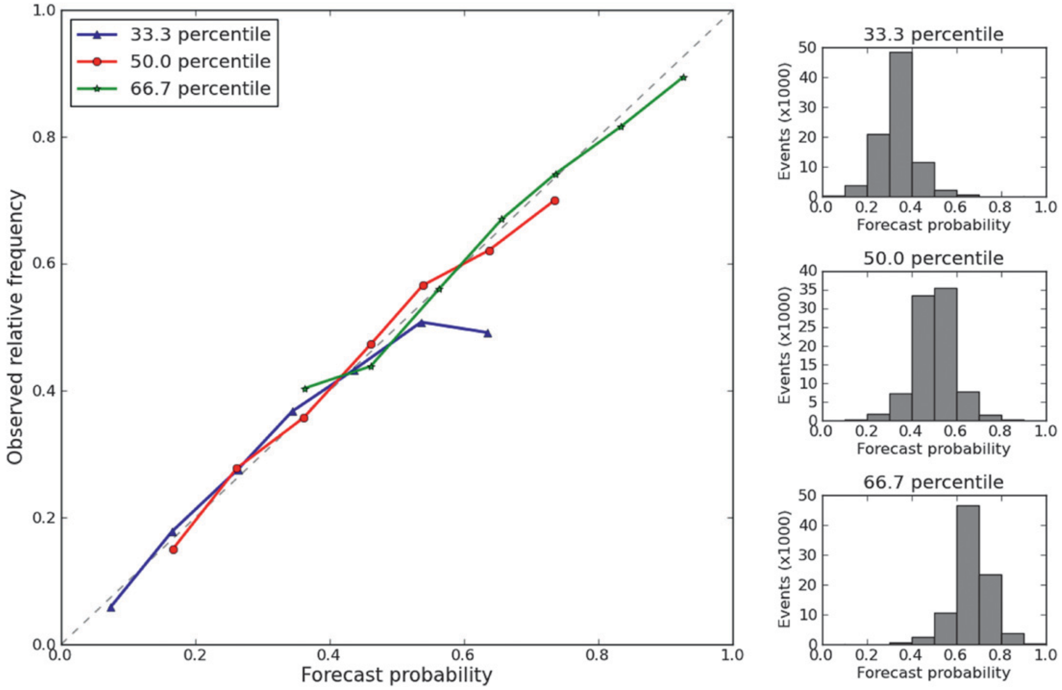

FIG. 2. (left) Reliability diagram and (right) resolution of fully merged BMA forecasts of the next season rainfall, for probabilities of events not exceeding 33.3, 50.0, and 66.7 percentiles of climatology.

Merged forecasts are also produced for each of the three groups of models—Pacific, Indian, and extratropical. The group forecast skills are compared with the fully merged forecast skill to understand the contributions from individual groups.

d. Comparison with two alternative approaches

We compare the BMA approach with two alternative forecasting approaches. The first is to select predictors a priori based on physical understanding of empirical relationships. The current operational model used by the Australian Bureau of Meteorology for forecasting seasonal rainfall is a statistical model that uses four predictors: the Niño-3.4 index and the second EOF of Pacific–Indian SST anomalies with lags of 2 and 4 months (for 1-month lead-time forecasts). Presumably, these predictors were chosen because they represent El Niño–Southern Oscillation (ENSO) and the state of the Indian Ocean, which are known to link with Australian rainfall. In this study, we construct a two-predictor model that uses the 1-month lagged Niño-3.4 index and the Indian Ocean dipole mode index (DMI) as predictors.

The second alternative approach is to select the single ‘‘best’’ model for each location and season using the PsBF as the selection criterion (Robertson and Wang 2012). The PsBF for a competing model M_k against a climatology model M_1 can be formulated in logarithmic form by

$$\ln(\mathrm{PsBF}_k) = \ln \prod_{t=1}^{T} \frac{f_k^{(t)}[\mathbf{y}_{\mathrm{OBS}}^t(2) \mid \mathbf{y}_{\mathrm{OBS}}^t(1)]}{f_1^{(t)}[\mathbf{y}_{\mathrm{OBS}}^t(2) \mid \mathbf{y}_{\mathrm{OBS}}^t(1)]}. \qquad (9)$$

The model with the highest ln(PsBF) that exceeds 4 is selected as the best model, and if none of the models has ln(PsBF)>4, then the climatology model is selected (Robertson and Wang 2012). The application of a threshold is to reduce the influence of data noise on model selection. For forecast assessment based on historical data, the forecast for a particular event is produced by using the best model selected based on the ln(PsBF) calculated from the cross-validation predictive densities of individual models for all the events except the event being forecast. This treatment is the same as for the BMA weight determination described earlier.

4. Results and discussion

a. Skill and reliability of merged forecasts

RMSEP skill scores for the fully merged BMA forecasts are presented by season (Fig. 1). Overall, the patterns of the skill scores in each season are consistent with the known phases of climate that are linked with Australian rainfall variability. In the first half of the year, noise tends to dominate the signal in climate indices and limits the ability to differentiate climate from one year

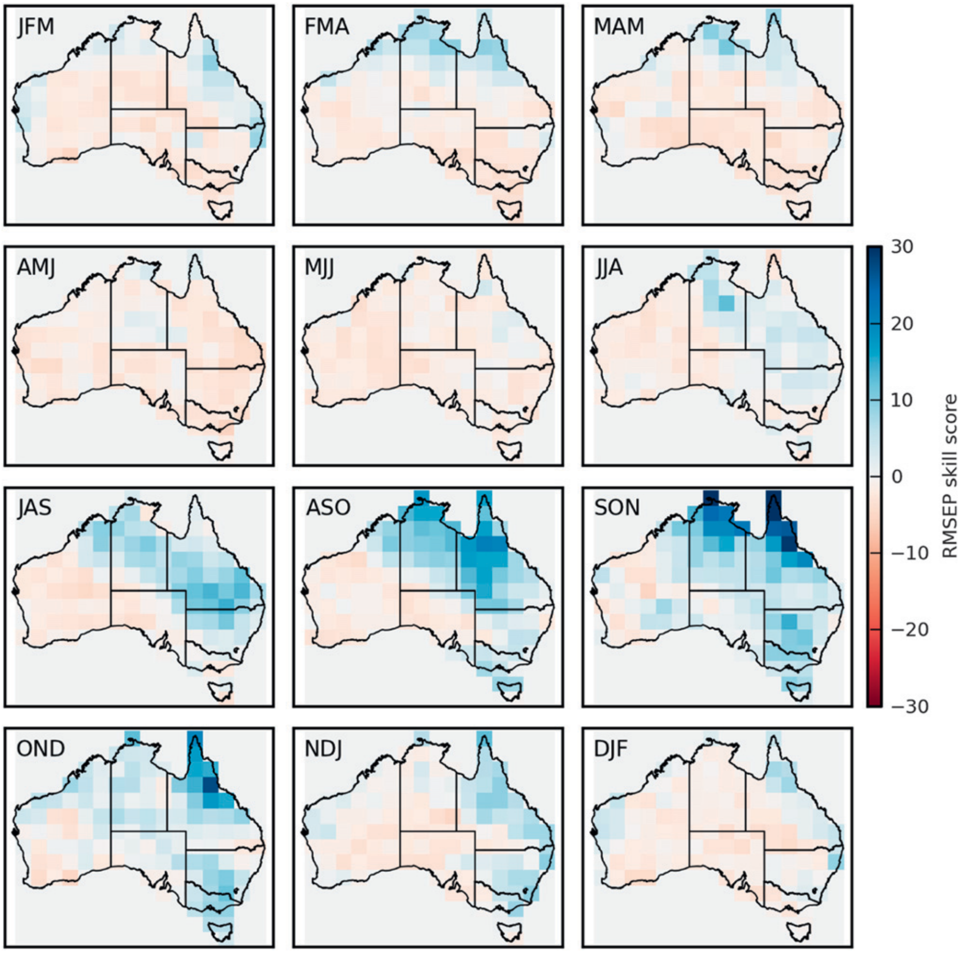

FIG. 3. Skill of merged BMA forecasts of the next season rainfall of the Pacific group of models.

to the next. This is reflected in the skill scores for January–March (JFM) to June–August (JJA), which are generally close to zero, apart from small patches of weak positive skill scores across northern Australia.

In the latter half of the year, the signals in climate indices become stronger, increasing the ability to forecast seasonal rainfall. Positive skill scores are found across large areas of northern, eastern, and central Australia from July to September (JAS) to September to November (SON). There are positive skill scores across most of the continent in October–December (OND). In November–January (NDJ) and December–February (DJF), the skill scores decrease, first in the east, followed by the west.

The overall reliability of the fully merged BMA forecasts is assessed using the reliability diagrams for forecast probabilities of events not exceeding the 33.3, 50, and 66.7 percentiles of climatology (Fig. 2). Nearly all the points in the reliability diagrams are aligned well to the 1:1 lines, suggesting that the forecast nonexceedance probabilities are consistent with the observed frequencies. In other words, the forecast distributions are reliable in representing forecast uncertainty.

b. *Contributions from the Pacific, Indian, and extratropical groups of models*

To demonstrate the advantage of BMA in effectively merging forecasts of different models, BMA is applied separately to models in the Pacific, Indian, and extratropical groups and skill scores are assessed (Figs. 3–5).

The five climate indices in the Pacific group (Table 1) all capture ENSO signals but have the potential to capture different ‘‘flavors’’ of El Niño and La Niña events, which may have variable impact on regions of Australia (Wang and Hendon 2007). The BMA forecasts of the group are mainly skillful in eastern and northern Australia from JAS to NDJ (Fig. 3). There are patches of positive skill scores across northern Australia from JFM to March–May (MAM), which can be associated with an ENSO Modoki index (EMI) type of events (Schepen et al. 2012;

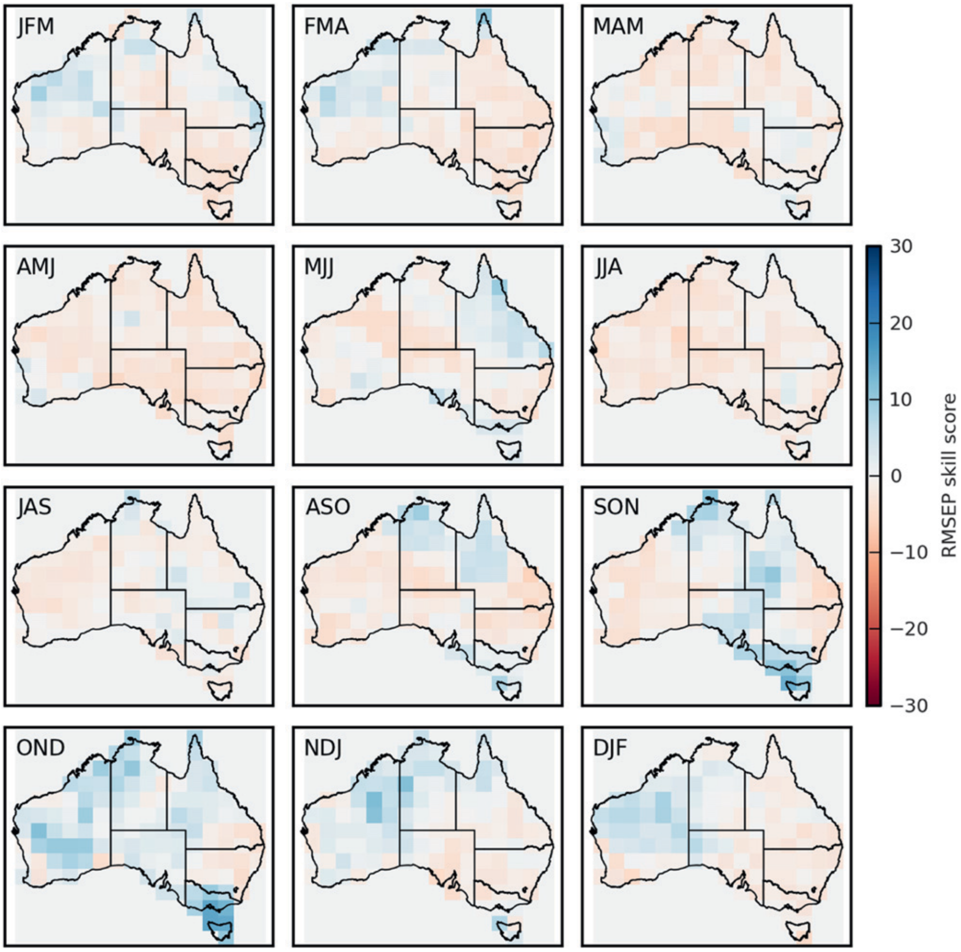

FIG. 4. As in Fig. 3, but for the Indian group of models.

Taschetto et al. 2009). By comparing Fig. 3 with Fig. 1, it is clear that the Pacific models contribute a significant amount of skill to seasonal rainfall forecasts in Australia. Still, the fully merged BMA forecasts have skill in more regions.

The models in the Indian group are included to represent the Indian Ocean dipole (IOD) as well as SST anomalies near Indonesia. The IOD is commonly linked to Australian rainfall; however, the Indonesia index (II) has also been shown to be a useful predictor in some seasons (Schepen et al. 2012; Verdon and Franks 2005). The BMA forecasts of the Indian group are skillful mainly in western parts of Australia from OND to JFM (Fig. 4). Skill arising from the Indian Ocean reaches the southeast region of Australia in SON and OND when the amplitude of the IOD tends to be at a maximum (Zhao and Hendon 2009). Overall, the BMA forecasts of the Indian group of models are less skillful than those of the Pacific group. However, the models in the Indian group contribute exclusively to the skill in some regions and seasons, and are therefore valuable inclusions in the full BMA.

The models in the extratropical group are included as they have been shown to have useful predictive relationships with Australian rainfall (Schepen et al. 2012). The BMA forecasts of models in the extratropical group are most skillful in Western Australia in OND and NDJ. It is noted that the regions where the extratropical group BMA forecasts are skillful overlap with regions where the Indian or Pacific group BMA forecasts are also skillful. However, inclusion of this group of models in the fully merged BMA is justified because the magnitude of the skill scores can be higher than the other two groups.

These results demonstrate the effectiveness of BMA in merging forecasts. The fully merged BMA forecasts have greater spatial coverage of positive skill than any one group. In general, BMA works to enhance predictive performance through the addition of models with distinct skills, but there is no significant penalty for applying BMA

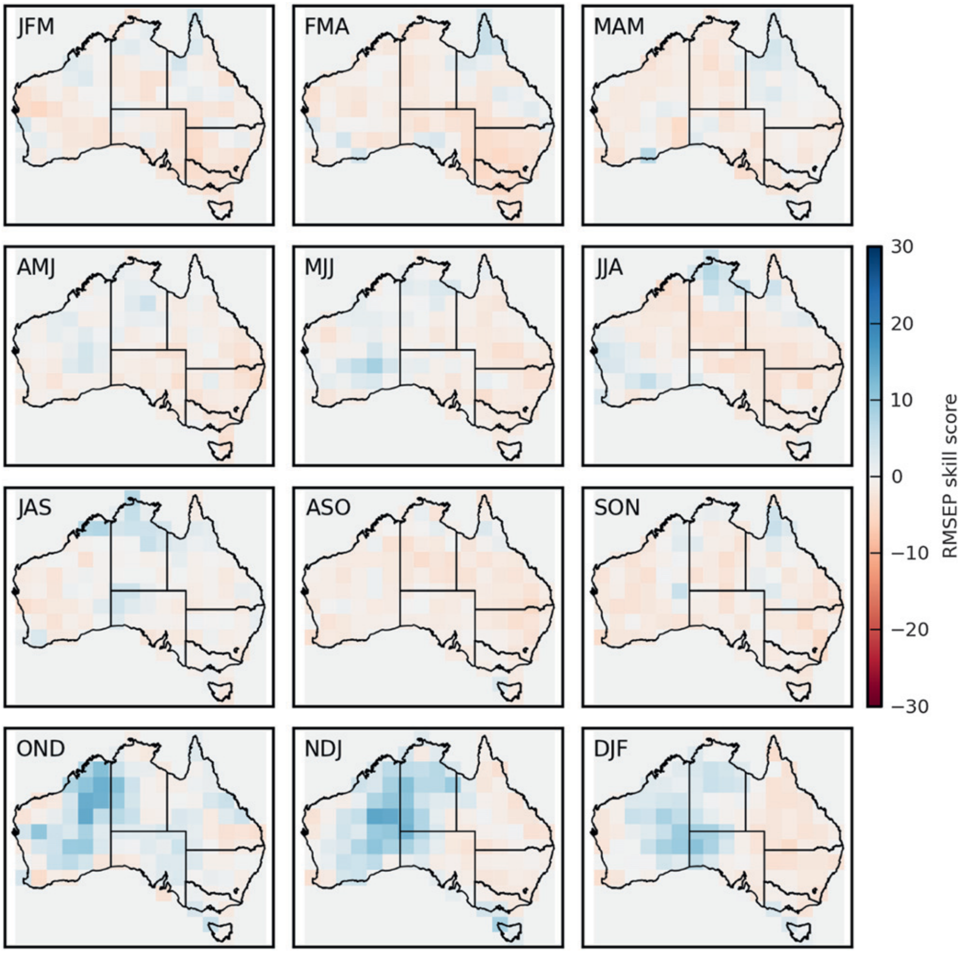

FIG. 5. As in Fig. 3, but for the extratropical group of models.

when the models have overlapping skills. Naturally, where there is no skill in any model, the fully merged forecasts also have no skill.

c. Comparison with alternative approaches

The first alternative forecasting approach assessed is to use a model with two fixed predictors chosen a priori. In this study, a model is constructed with the 1-month lagged Niño-3.4 index and the DMI as predictors. The resulting spatial and temporal coverage of RMSEP skill scores (Fig. 6) is not as great as the fully merged BMA forecasts (Fig. 1), with less skill in Western Australia from OND to DJF particularly evident. One theoretical advantage of the two-predictor model is that it allows for possible interactions between the predictors. Here, both the Niño-3.4 index and the DMI have been linked to rainfall, for example, over southeast Australia in SON and OND (Risbey et al. 2009; Schepen et al. 2012), and ENSO and the IOD are also known to interact (Meyers et al. 2007). However, the two-predictor model does not show any perceptible gain in skill over the BMA of the multiple single-predictor models (Fig. 6 compared with Fig. 1). This will be further investigated in the next section. For now, we should point out that it is undesirable to use models with too many predictors when only limited data are available to calibrate the models, especially if the predictors are correlated with each other. Overly fitted models tend to perform poorly for independent events. In this regard, BMA is a more attractive approach to include more predictors.

The second alternative forecasting approach assessed is to select the best model based on the ln(PsBF) with a threshold of 4. The resulting skill scores (Fig. 7) immediately identify this approach as inferior to the BMA and fixed-predictor approaches. Where the underlying skill is high, the best model approach generally produces skillful forecasts. However, the skill scores are sharply negative in many regions and seasons. This is caused by model switching in cross validation and giving full weight to the different ‘‘best’’ models, whose selection has been

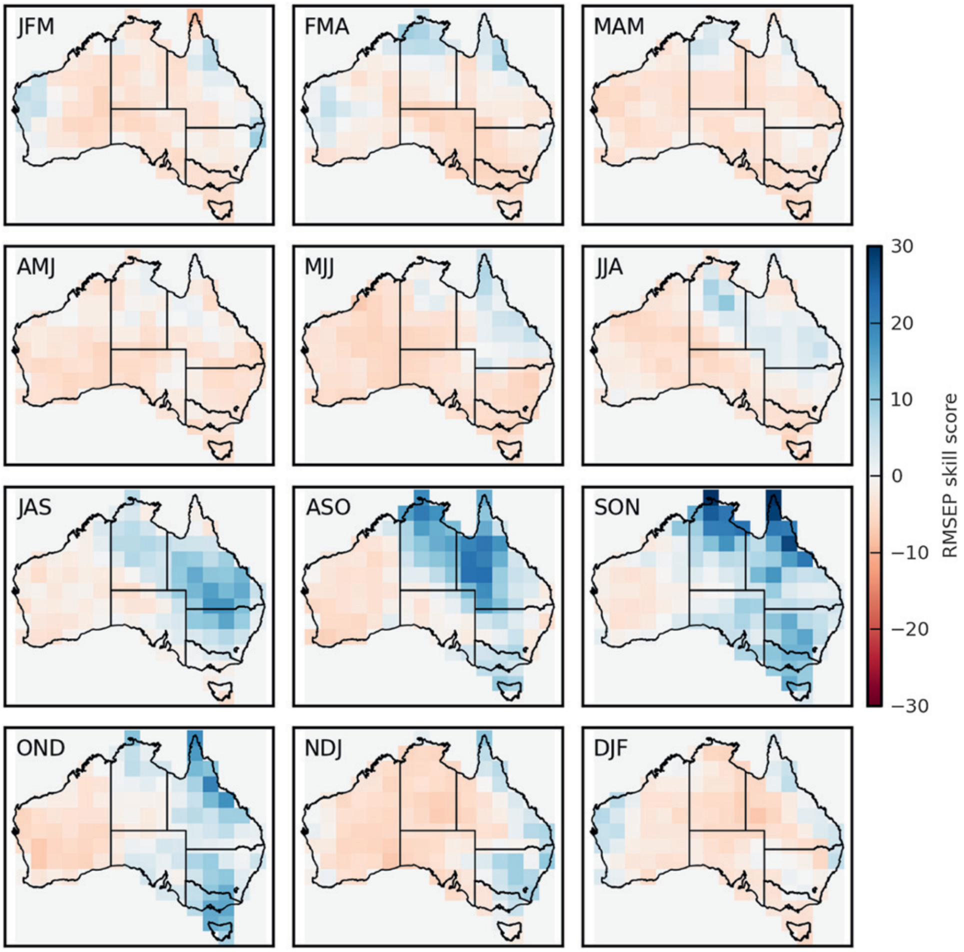

FIG. 6. Skill of forecasts of the next season rainfall of a model with two fixed predictors (Niño-3.4 and DMI).

influenced by data noise despite the use of the selection threshold to limit the problem. In contrast, the BMA method is shown to be robust in cross validation, despite the fact that it combines a large number (34) of models.

5. Further results and discussion

In presenting the results of the fixed-predictor approach in the last section, we discussed briefly possible interactions between different climate variables in influencing Australian rainfall. To investigate further if better forecast skills can be obtained by allowing for such interactions, we add a number of two-predictor models into the pool of models for BMA. These two-predictor models allow for interactions between the Pacific and Indian groups, the Pacific and extratropical groups, and the Indian and extratropical groups. Our results (not shown here) indicate that the inclusion of the additional two-predictor models does not result in any obvious improvement in forecast skill.

The Australian Bureau of Meteorology currently provides 1-month lead-time seasonal rainfall forecasts. The results presented so far have been for 0 lead-time forecasts. We apply here the BMA method to produce 1-month lead-time forecasts by including only the models that use climate indices with lags of 2 and 3 months. There are a total of 22 single-predictor models and one no-predictor (climatology) model. The skill scores for the 1-month lead-time forecasts (Fig. 8) are surprisingly similar to the 0 lead-time forecasts (Fig. 1), with only a slight drop overall. One possible reason for this is that many of the climate indices show strong monthly persistence.

As stated in the introduction, we aim to develop a BMA method that is capable of producing relatively stable weights in presence of significant sampling variability, leading to robust forecasts for future events. For this purpose,we apply a number of techniques, including

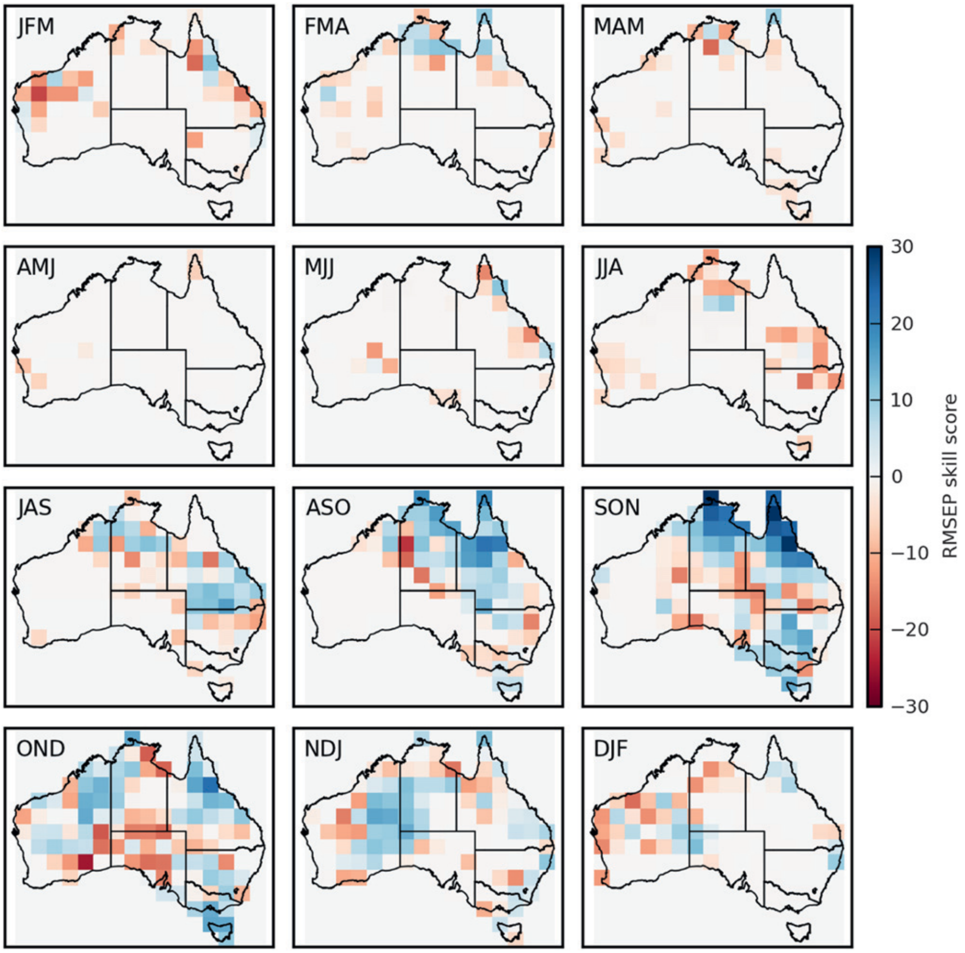

FIG. 7. Skill of the best predictor model forecasts of the next season rainfall.

the mixture model approach rather than the classical model posterior probability approach, the use of a prior of BMA weights that gives a slight preference toward more evenly distributed weights, and the use of the cross-validation likelihood function rather than the classical likelihood function in Bayesian inference. We conduct extensive analyses and find that our BMA method yields more stable weights and better forecast skill in cross validation than the more classical approaches (results not shown).

In this study, we merge forecasts from all models when applying the BMA method. Averaging over all models can give the best predictive ability (Madigan and Raftery 1994). However, to reduce data requirement and computational burden, it may be desirable for operational applications to include only a subset of the models to derive the final merged forecasts. One way to achieve this is to use OCCAM's window (Madigan and Raftery 1994), whereby models with weights smaller than c (say, 0.05) times the weight of the best model are omitted.

Although this paper has focused on merging forecasts of statistical models, the BMA method developed can be easily applied to merging forecasts of multiple dynamical climate models and to merging forecasts of both dynamical and statistical models. Indeed, the results from this study have clearly demonstrated that the forecast skill from statistical models alone is still low for many seasons and locations despite forecast merging. An approach that combines the strengths of both statistical and dynamical models is likely to produce more skillful forecasts. We will report on our study on this approach in a future paper.

6. Conclusions

Seasonal rainfall forecasting will continue to be challenging. Statistical prediction systems using climate indices as predictors remain relevant, although they can be improved. In the literature, many climate indices have been linked to Australian seasonal rainfall, although the

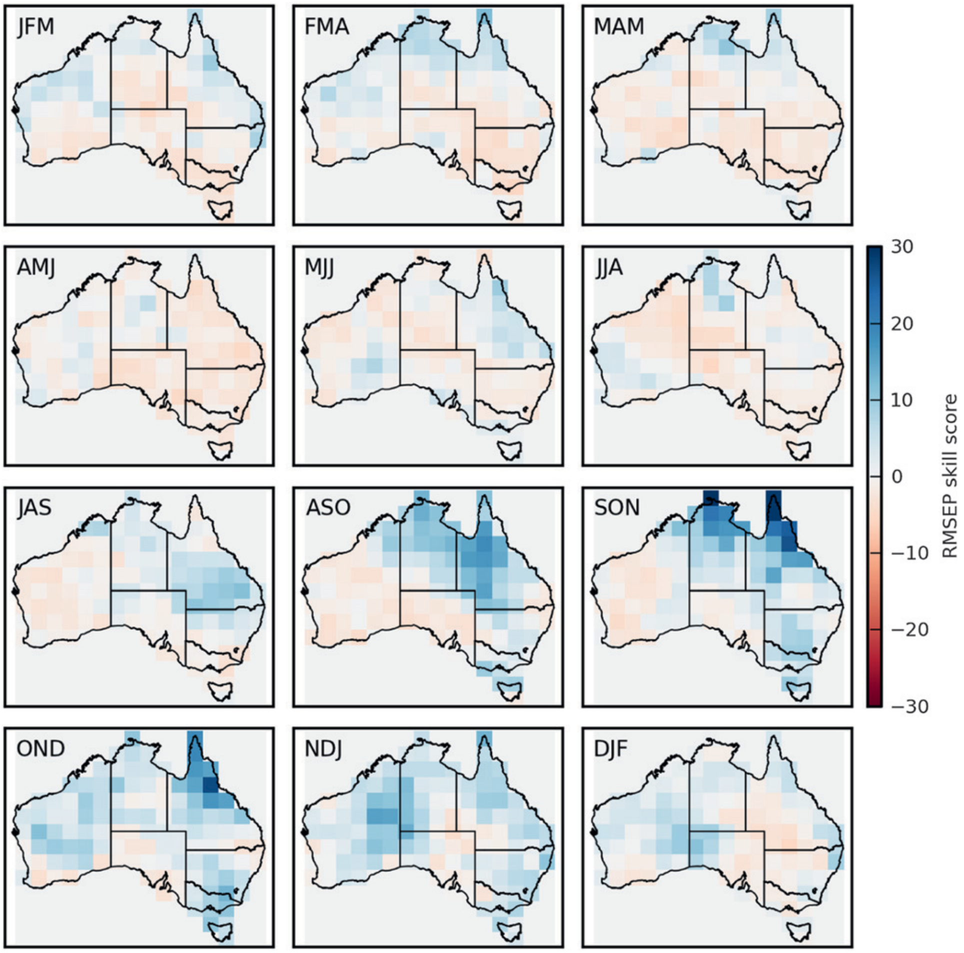

FIG. 8. Skill of fully merged BMA forecasts of seasonal rainfall with 1-month lead time.

strengths of the links vary with season and location. One strategy for improving statistical seasonal rainfall forecasts, given that there is uncertainty about which predictors to use, is to merge forecasts of many models with different predictors.

In this study, we develop a BMA method for merging forecasts from multiple models. We aim for a BMA method that is capable of producing relatively stable weights in presence of significant sampling variability, leading to robust forecasts for future events. We apply the BMA method to merge forecasts from multiple statistical models using climate indices as predictors. We show that the fully merged forecasts effectively combine the best skills of the models to maximize the spatial coverage of positive skill scores. Overall, the skill is low for the first half of the year but more positive for the second half of the year. Models in the Pacific group contribute the most skill, and models in the Indian and extratropical groups also produce useful and sometimes distinct skills. The fully merged forecasts are found to be reliable; that is, the BMA method produces forecast probability distributions that reliably reflect the forecast uncertainty spread. We also show that the skill of 3-month rainfall total forecasts holds well when the forecast lead time is increased from 0 to 1 month.

The BMA method outperforms the approach of using a model with two fixed predictors chosen a priori (Niño-3.4 and DMI) and the approach of selecting the best model based on predictive performance. The fixed-predictor approach is competitive, but it is difficult to augment further to take advantage of additional predictive information contained in other climate indices. The best model approach produces forecasts that have sharply negative skill scores for many grid cells, despite the safeguard of a selection threshold. The BMA method allows for model uncertainty while taking advantage of the better performing models, and is shown to be robust in our cross-validation assessment.

The BMA method used in this study can be easily applied to merging forecasts of multiple dynamical climate models and to merging forecasts of both dynamical and statistical models. Indeed, the results from this study have clearly demonstrated that the forecast skill from statistical models alone is still low for many seasons and locations. A BMA approach to combining both statistical and dynamical models is likely to yield more skillful forecasts.

Acknowledgments. This work was completed as part of the Water Information Research and Development Alliance (WIRADA), a collaboration between CSIRO and the Bureau of Meteorology to facilitate the transfer of research to operations. The work was also partly funded by the CSIRO OCE Science Leader Scheme. We thank Dr. Eun-Pa Lim for providing the grid coordinates used in the Predictive Ocean Atmosphere Model for Australia (POAMA), Cathy Bowditch for proofreading and editing an early version of the manuscript, and Dr. William Wang for providing valuable comments on the work. We are most grateful to two anonymous reviewers, whose insightful critique and constructive suggestions have led us to improve the paper significantly.

REFERENCES

Casanova, S., and B. Ahrens, 2009: On the weighting of multimodel ensembles in seasonal and short-range weather forecasting. *Mon. Wea. Rev.,* **137,** 3811–3822.

Casey, T., 1995: Optimal linear combination of seasonal forecasts. *Aust. Meteor. Mag.,* **44,** 219–224.

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

Clarke, B., 2003: Comparing Bayesian model averaging and stacking when model approximation error cannot be ignored. *J. Mach. Learn. Res.,* **4,** 683–712.

Coelho, C. A. S., S. Pezzulli, M. Balmaseda, F. J. Doblas-Reyes, and D. B. Stephenson, 2004: Forecast calibration and combination: A simple Bayesian approach for ENSO. *J. Climate,* **17,** 1504–1516.

Domingos, P., 2000: Bayesian averaging of classifiers and the overfitting problem. *Proceedings of the Seventeenth International Conference on Machine Learning,* Morgan Kaufmann Publishers, 223–230.

Drosdowsky, W., and L. E. Chambers, 2001: Near-global sea surface temperature anomalies as predictors of Australian seasonal rainfall. *J. Climate,* **14,** 1677–1687.

Eklund, J., and S. Karlsson, 2007: Forecast combination and model averaging using predictive measures. *Econ. Rev.,* **26,** 329–363.

Fawcett, R., D. Jones, and G. Beard, 2005: A verification of publicly issued seasonal forecasts issued by the Australian Bureau of Meteorology: 1998-2003. *Aust. Meteor. Mag.,* **54,** 1–13.

Geweke, J., and C. Whiteman, 2006: Bayesian forecasting. *Handbook of Economic Forecasting,* G. Elliot, C. W. J. Granger, and A. Timmermann, Eds., *Handbooks in Economics,* Vol. 1, North-Holland, 3–80.

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

Jackson, C. H., S. G. Thompson, and L. D. Sharples, 2009: Accounting for uncertainty in health economic decision models by using model averaging. *J. Roy. Stat. Soc.,* **172A,** 383–404.

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

Kalnay, E., and Coauthors, 1996: The NCEP/NCAR 40-Year Reanalysis Project. *Bull. Amer. Meteor. Soc.,* **77,** 437–471.

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.

Madigan, D., and A. E. Raftery, 1994: Model selection and accounting for model uncertainty in graphical models using OCCAM's window. *J. Amer. Stat. Assoc.,* **89,** 1535–1546.

Meyers, G., P. McIntosh, L. Pigot, and M. Pook, 2007: The years of El Niño, La Niña, and interactions with the tropical Indian Ocean. *J. Climate,* **20,** 2872–2880.

Minka, T., 2000: Bayesian model averaging is not model combination. MIT Media Lab Note, 2 pp.

Monteith, K., J. Carroll, K. Seppi, and T. Martinez, 2011: Turning Bayesian model averaging into Bayesian model combination. *Proceedings of the IEEE International Joint Conference on Neural Networks,* IEEE, 2657–2663.

Raftery, A. E., D. Madigan, and J. A. Hoeting, 1997: Bayesian model averaging for linear regression models. *J. Amer. Stat. Assoc.,* **92,** 179–191.

——, T. Gneiting, F. Balabdaoui, and M. Polakowski, 2005: Using Bayesian model averaging to calibrate forecast ensembles. *Mon. Wea. Rev.,* **133,** 1155–1174.

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.

Robertson, D. E., and Q. J. Wang, 2012: A Bayesian approach to predictor selection for seasonal streamflow forecasting. *J. Hydrometeor.,* **13,** 155–171.

Rust, R. T., and D. C. Schmittlein, 1985: A Bayesian cross-validated likelihood method for comparing alternative specifications of quantitative models. *Mark. Sci.,* **4,** 20–40.

Schepen, A., Q. J. Wang, and D. E. Robertson, 2012: Evidence for using lagged climate indices to forecast Australian seasonal rainfall. *J. Climate,* **25,** 1230–1246.

Shinozaki, T., S. Furui, and T. Kawahara, 2010: Gaussian mixture optimization based on efficient cross-validation. *IEEE J. Sel. Top. Signal Process.,* **4,** 540–547.

Smith, T. M., R. W. Reynolds, T. C. Peterson, and J. Lawrimore, 2008: Improvements to NOAA's historical merged land–ocean surface temperature analysis (1880–2006). *J. Climate,* **21,** 2283–2296.

Smyth, P., 1996: Clustering using Monte Carlo cross-validation. *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining,* E. Simoudis, J. Han, and U. Fayyad, Eds., AAAI Press, 126–133.

——, 2000: Model selection for probabilistic clustering using cross-validated likelihood. *J. Stat. Comput.,* **10,** 63–72.

Stephenson, D. B., C. A. S. Coelho, F. J. Doblas-Reyes, and M. Balmaseda, 2005: Forecast assimilation: A unified framework for the combination of multi-model weather and climate predictions. *Tellus,* **57A,** 253–264.

Stone, M., 1977: An asymptotic equivalence of choice of model by cross-validation and Akaike's criterion. *J. Roy. Stat. Soc.,* **39B,** 44–47.

Taschetto, A. S., C. C. Ummenhofer, A. Sen Gupta, and M. H. England, 2009: Effect of anomalous warming in the central Pacific on the Australian monsoon. *Geophys. Res. Lett.,* **36,** L12704, doi:10.1029/2009GL038416.

Troup, A. J., 1965: Southern Oscillation. *Quart. J. Roy. Meteor. Soc.,* **91,** 490–506.

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, G., and H. H. Hendon, 2007: Sensitivity of Australian rainfall to inter–El Niño variations. *J. Climate,* **20,** 4211–4226.

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.

Wilks, D. S., 1995: *Statistical Methods in the Atmospheric Sciences: An Introduction.* Academic Press, 467 pp.

Wright, J. H., 2009: Forecasting US inflation by Bayesian model averaging. *J. Forecasting,* **28,** 131–144.

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

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.

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