Combining the strengths of statistical and dynamical modeling approaches for forecasting Australian seasonal rainfall

Andrew Schepen,[1] Q. J. Wang,[2] and David E. Robertson[2]

Received 27 April 2012; revised 7 September 2012; accepted 7 September 2012; published 20 October 2012.

[1] Forecasting rainfall at the seasonal time scale is highly challenging. Seasonal rainfall forecasts are typically made using statistical or dynamical models. The two types of models have different strengths, and their combination has the potential to increase forecast skill. In this study, statistical-dynamical forecasts of Australian seasonal rainfall are assessed. Statistical rainfall forecasts are made based on observed relationships with lagged climate indices. Dynamical forecasts are made by calibrating raw outputs from multiple general circulation models. The statistical and dynamical forecasts are then merged using a Bayesian model averaging (BMA) method. The skill and reliability of the forecasts is assessed through cross-validation for the period 1980–2010. We confirm that the dynamical and statistical groups of models give skill in different locations and seasons and the merged statistical-dynamical forecasts represent a significant improvement in terms of maximizing spatial and temporal coverage of skillfulness. We find that the merged statistical-dynamical forecasts are reliable in representing forecast uncertainty.

Citation: 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.

1. Introduction

[2] The demand for accurate and reliable forecasts of seasonal climate, particularly rainfall, continues to be strong. Objective seasonal rainfall forecasts are produced using either statistical models or dynamical models such as general circulation models (GCMs). In Australia, the Bureau of Meteorology issues its operational seasonal rainfall forecasts based on a statistical model [*Fawcett et al.*, 2005; *Fawcett*, 2008], which uses climate indices as predictors. The climate indices represent the El Niño Southern Oscillation (ENSO) and the status of the Indian Ocean. In line with international trends, the Bureau is also developing a GCM with the view that it will eventually replace the statistical model as the operational model. The coupled ocean–atmosphere GCM, the Predictive Ocean Atmosphere Model for Australia (POAMA), is tailored to seasonal forecasting.

[3] Statistical models have been used extensively throughout the world to forecast seasonal climate [e.g., *Barnston et al.*, 1999; *Fawcett et al.*, 2005; *Folland et al.*, 1991; *Landsea and Knaff*, 2000; *Rajeevan et al.*, 2007; *Shabbar and Barnston*, 1996]. They are based on observational relationships, and their main advantage is that they are simple to implement and operate. The main concern with statistical models is the fact that they are reliant on stationary relationships between the predictor and predictand variables, which is not guaranteed in a changing climate.

[4] GCMs are increasingly being used to forecast seasonal climate [e.g., *Lim et al.*, 2009; *Palmer et al.*, 2004; *Saha et al.*, 2006; *Yasuda et al.*, 2007]. They are based on the laws of physics and their main advantages are that they have the ability to capture nonlinear interactions of the atmosphere, land and ocean, and are adaptable to shifts in climate. GCMs also provide spatially and temporally coherent forecasts of multiple variables (e.g., temperature, pressure) at high temporal resolution (e.g., daily). However, they suffer from the general problem that the spread of ensemble members tends to be too narrow (overconfident) [e.g., *Graham et al.*, 2005, *Lim et al.*, 2011] and the climatology of the simulations is not always aligned with that of the corresponding observations. It is therefore common to apply statistical models to calibrate raw GCM output for bias correction and variance adjustment so that the forecast climatology more closely matches the observed and the forecasts are statistically reliable [e.g., *Doblas-Reyes et al.*, 2005; *Feddersen et al.*, 1999; *Landman and Goddard*, 2002]. Of course, statistical calibration also assumes stationary predictor-predictand relationships and thus can be viewed as interim solution while GCMs continue to be improved.

[5] Given their unique strengths, both statistical and dynamical models are attractive approaches for seasonal climate forecasting. Recently, *Barnston et al.* [2012] compared the real-time performance of 12 GCMs and 8 statistical models for forecasting ENSO (Niño3.4) from 2002 to 2011. At this large scale, the group of GCMs was able to

[1]Bureau of Meteorology, Brisbane, Queensland, Australia.
[2]CSIRO Land and Water, Highett, Victoria, Australia.

Corresponding author: A. Schepen, Bureau of Meteorology, GPO Box 413, Brisbane, Qld 4001, Australia. (a.schepen@bom.gov.au)

0148-0227/12/2012JD018011

Table 1. Climate Indices Used as Predictors of Australian Seasonal Rainfall

Climate Index	Description
Southern Oscillation Index (SOI)	Pressure difference between Tahiti and Darwin as defined by *Troup* [1965]
NINO3	Average sea surface temperature anomaly over 150W–90W and 5N–5S
NINO3.4 (NINO34)	Average sea surface temperature anomaly over 170W–120W and 5N–5S
NINO4	Average sea surface temperature anomaly over 150E–160E and 5N–5S
ENSO Modoki Index (EMI)	C – 0.5 (E + W) where the components are average sea surface temperature anomalies over: C: 165E–140W and 10N–10S E: 110W–70W and 5N–15S W: 125E–145E and 20N–10S
Indian Ocean West Pole Index (WPI)	Average sea surface temperature anomaly over 50E–70E and 10N–10S
Indian Ocean East Pole Index (EPI)	Average sea surface temperature anomaly over 90E–110E and 0N–10S
Indian Ocean Dipole Mode Index (DMI)	WPI – EPI
Indonesia Index (II)	Average sea surface temperature anomaly over 120E–130E and 0N–10S

outperform the group of statistical models during the onset or transition phases of ENSO events, suggesting dynamical models may be starting to edge ahead in terms of skill. However, a general conclusion was that there remains a place in operational climate prediction for statistical modeling when GCMs fail to provide useful information. There is also potential to improve seasonal forecasts by objectively combining the information from statistical and dynamical models, particularly for sub-grid scale processes such as rainfall.

[6] Some recent studies have investigated Bayesian approaches for combining statistical forecasts with raw GCM output. *Coelho et al.* [2004] developed a Bayesian methodology for combining statistical forecasts with raw GCM output to provide improved and well-calibrated long lead Niño3.4 predictions. A statistical forecast was used as an informative prior distribution, which was updated with raw GCM output. *Luo et al.* [2007] adopted a similar approach to update a climatology model with raw output from multiple GCMs, but assumed model independence for simplicity. A more flexible Bayesian approach to combine multiple models, which does not assume model independence, is to establish the models individually and weight and merge forecasts based on past predictive performance through Bayesian Model Averaging (BMA) [*Hoeting et al.*, 1999; *Raftery et al.*, 2005]. *Wang et al.* [2012] developed a BMA method for merging multiple statistical seasonal rainfall forecasting models based on climate indices. Here, we apply the method to merge multiple statistical forecasts with calibrated dynamical forecasts from multiple GCMs.

[7] In this paper, we combine statistical and dynamical forecasts of Australian seasonal rainfall. We use a consistent Bayesian modeling approach to establish statistical models based on lagged climate indices and to calibrate raw rainfall forecasts from dynamical models. We demonstrate that the statistical and dynamical models have different strengths and weaknesses. In particular, they produce skillful seasonal rainfall forecasts in different Australian regions and seasons. We further demonstrate that by weighting and merging the forecasts from the different models, we take advantage of their respective strengths, and achieve greater spatial and temporal coverage of skillfulness.

[8] The remainder of this paper is structured as follows. In the next section, we present the statistical and dynamical forecasting models and data. In section 3, we outline the BMA method for merging forecasts and verification methods for assessing the skill and reliability of probabilistic forecasts. In section 4, we present maps and diagrams showing the skill and reliability of the forecasts and show some examples of model weights. Section 5 provides some supplementary discussion. Section 6 completes the paper with a summary and conclusions.

2. Model Formulation and Data

[9] We seek to forecast seasonal (three month) rainfall totals at lead times of 0 and 1 month for 2.5 degree grid cells covering Australia. Multiple statistical models and dynamical models are established for each season, grid cell and lead time independently of the others. We also establish a climatology model, which is a model with no predictor. Forecasts are made on the first day of each of the 12 months. We use a common period of 1980–2010 for all data in this study to coincide with the availability of POAMA hindcasts.

[10] The statistical models are established using lagged climate indices as predictors. Only single predictor models are used. The climate indices included in this study are well known to represent anomalies in the large-scale circulations in the tropical Pacific and Indian Ocean regions. As such, they reflect the phases of the El Niño Southern Oscillation and fluctuations in Indian Ocean SSTs which are the dominant drivers of Australian rainfall [e.g., *Risbey et al.*, 2009]. *Schepen et al.* [2012] showed significant lagged relationships exist for some regions and seasons and therefore the climate indices are also suitable for forecasting. In total there are nine monthly climate indices. For lead time 0 forecasts, each index is lagged by 1 month. For lead time 1 forecasts, each index is lagged by two months. Table 1 summarizes the nine climate indices and provides a brief description of each. Climate indices based on sea surface temperatures were derived from the NCAR Extended Reconstruction of Sea Surface Temperature version 3 [*Smith et al.*, 2008]. The Southern Oscillation Index is sourced from the Australian Bureau of Meteorology.

[11] The dynamical models used here are calibration models that aim to correct biased and under-dispersed raw GCM ensembles. The GCMs included in this study are three variants of the Predictive Ocean Atmosphere Model for Australia (POAMA), denoted as P24A, P24B and P24C [*Wang et al.*, 2011]. Compared to P24C, P24A and

P24B have an alternative parameterization of atmospheric physics associated with shallow convection. P24B additionally has a flux correction scheme to reduce biases in the ocean–atmosphere climatologies that can arise as lead time is increased. Each POAMA variant is a coupled ocean–atmosphere GCM that produces a 10 member forecast ensemble by perturbing ocean conditions. The dynamical models are established using the raw GCM ensemble mean rainfalls as predictors in statistical models. Again only single predictor models are used.

[12] The rainfall data used in this study are derived from the Australian Water Availability Project (AWAP) 0.05 degree × 0.05 degree gridded data set of monthly rainfall [*Jones et al.*, 2009]. Monthly rainfall is upscaled to 2.5 degree × 2.5 degree grid by averaging within POAMA grid cells and then aggregated to obtain seasonal totals.

[13] A flexible multivariate Bayesian joint probability (BJP) modeling approach is used to establish the statistical forecast models and calibrate the dynamical models. Only a brief description of the BJP modeling approach is provided here, but full details are given by *Wang et al.* [2009] and *Wang and Robertson* [2011, 2012]. For ease of understanding, we relax the description of the BJP formulation here to the bivariate case. That is, we consider only single predictor and predictand models.

[14] Given a predictor variable x and a predictand variable y, Yeo–Johnson transforms [*Yeo and Johnson*, 2000] are applied to transform each to normality. The transformed predictor and predictand variables are thus assumed to jointly follow a bivariate normal distribution. A Bayesian inference of the Yeo–Johnson transform parameters and the bivariate normal distribution parameters (means, standard deviations and correlation coefficient) is made by using historical (or hindcast) data from 1980 to 2010. The inference is numerically implemented through Markov chain Monte Carlo (MCMC) sampling based on the Metropolis algorithm. Rainfall is a variable that is bounded by zero from below. In the BJP implementation, this is handled by treating the rainfall predictor and predictand variables as left-censored at 0. More complete details on the handling of this problem are given by *Wang and Robertson* [2011] and *Robertson and Wang* [2012].

[15] Probabilistic forecasts are readily generated by the BJP modeling approach. Consider a model M_k with parameters $\boldsymbol{\theta}$. If $(\mathbf{x}_k^T, \mathbf{y}^T)$ contains the predictor and predictand data used for parameter inference and x_k is a predictor value, then a probabilistic forecast for a single event is given by the posterior predictive density:

$$f_k(y \mid x_k) = p\left(y \mid x_k; \mathbf{x}_k^T, \mathbf{y}^T, M_k\right) = \int p(y \mid x_k; \boldsymbol{\theta}, M_k) \cdot p\left(\boldsymbol{\theta} \mid \mathbf{x}_k^T, \mathbf{y}^T, M_k\right) \cdot d\boldsymbol{\theta}. \quad (1)$$

3. Statistical-Dynamical Forecast Merging

3.1. Bayesian Model Averaging

[16] To merge the forecasts of the statistical models and the statistically calibrated dynamical models, we apply a Bayesian model averaging (BMA) approach. Complete details of the approach are presented by *Wang et al.* [2012]. Here, we present only the key features necessary for understanding.

[17] For a single forecast event, a merged probabilistic forecast from K models is given by the BMA predictive density:

$$f_{BMA}(y \mid x_1, \ldots, x_K) = \sum_{k=1}^{K} w_k f_k(y \mid x_k) \quad (2)$$

where for model k, x_k and y are respectively the predictor and predictand variables and w_k is the model weight.

[18] We make a Bayesian inference of the weights based on the performance of leave-one-out cross validation predictive densities. Cross-validation predictive densities are used as a safeguard against overfitting. Our Bayesian inference of the weights follows a finite mixture model approach [e.g., *Raftery et al.*, 2005], rather than the classical model posterior probability approach [e.g., *Hoeting et al.*, 1999]. In the mixture model approach, the BMA weights are inferred from evaluation of predictive performance of the merged forecasts.

[19] The short data period (31 years) available leads to uncertainty about the optimum model weights due to significant sampling variability. Compounding this problem, it has been noted previously that BMA can give small weights to all but the best model and hence increases the risk of overfitting, leading to poor forecasts [e.g., *Casanova and Ahrens*, 2009]. We therefore apply a Dirichlet prior that helps to stabilize the weights:

$$p(w_k, k = 1, \ldots K) \propto \prod_{k=1}^{K} (w_k)^{\alpha - 1}. \quad (3)$$

[20] The concentration parameter α, is set to $\alpha = 1.0 + \alpha_0/K$, where $\alpha_0 = 1.0$ is used in this study. This gives a slight preference toward more evenly distributed weights. In this setup, the posterior distribution of the weights is then proportional to:

$$A = \prod_{k=1}^{K} (w_k)^{\alpha - 1} \prod_{t=1}^{T} \sum_{k=1}^{K} w_k f_k^{(t)}\left(y^t \mid x_k^t\right) \quad (4)$$

where $f_k^{(t)}(y^t|x_k^t)$ is the cross-validation predictive density. We find a point estimate of the weights by maximizing A using a highly efficient expectation- maximization (EM) algorithm [*Cheng et al.*, 2006; *Zivkovic and van der Heijden*, 2004]. The event to be forecast is left out. Initially, all weights are set to equal. The EM algorithm is then iterated until convergence is achieved. In this study, convergence is achieved when the change in ln(A) is smaller than 0.0001.

3.2. Forecast Assessment

[21] Leave-one-out cross-validation is used to assess forecasts during the period 1980–2010. The short data period does not permit validation of forecasts in an independent period. It is therefore important that we note that the forecasts are not completely independent of the training period. For example, ENSO related predictors may not be independent from year-to-year. However, this type of cross-validation is fairly standard and, for seasonal rainfall forecasting, is likely to provide reasonable skill estimates.

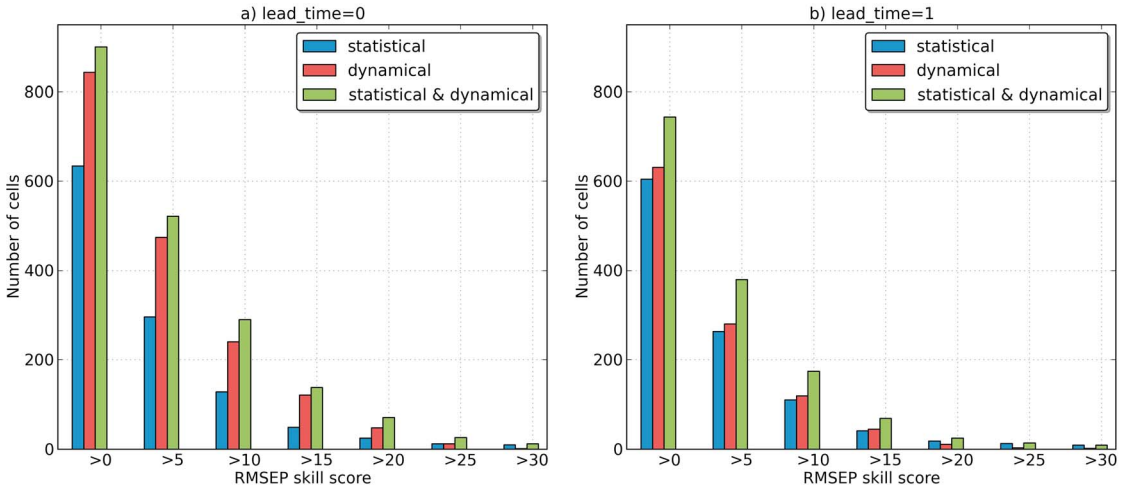

Figure 1. Total number of grid cells where RMSEP skill score exceeds a range of thresholds. All grid cells for all seasons have been pooled together.

[22] To assess forecast accuracy, we use the root mean squared error in probability (RMSEP) [*Wang and Robertson*, 2011] skill score to assess forecast medians, using the climatological median as the reference forecast. This score is formulated so that incorrect near median forecasts are penalized more severely than extreme forecasts that have a larger error but are in the right direction. Mathematically, RMSEP is given by:

$$\mathrm{RMSEP} = \sqrt{\frac{1}{T}\sum_{t=1}^{T}\left[F_{c\,\mathrm{lim}}\left(y^{t}_{fcst}\right) - F_{c\,\mathrm{lim}}\left(y^{t}_{obs}\right)\right]^{2}} \qquad (5)$$

where F_{clim} is the cumulative distribution function of the climatological data and y^{t}_{fcst} and y^{t}_{obs} are the forecast median and observed values, respectively. The RMSEP skill score is then formulated as a generalized skill score:

$$\mathrm{SS}_{\mathrm{RMSEP}} = 100 \times \frac{RMSEP_{c\,\mathrm{lim}} - RMSEP_{fcst}}{RMSEP_{c\,\mathrm{lim}}}. \qquad (6)$$

[23] The reference forecasts used are the corresponding cross validation climatological medians. 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. A negative skill score means the climatological median is a more skillful forecast. We also assess forecast accuracy using the continuous ranked probability score (CRPS) which assesses full forecast distributions. Conclusions drawn from the results

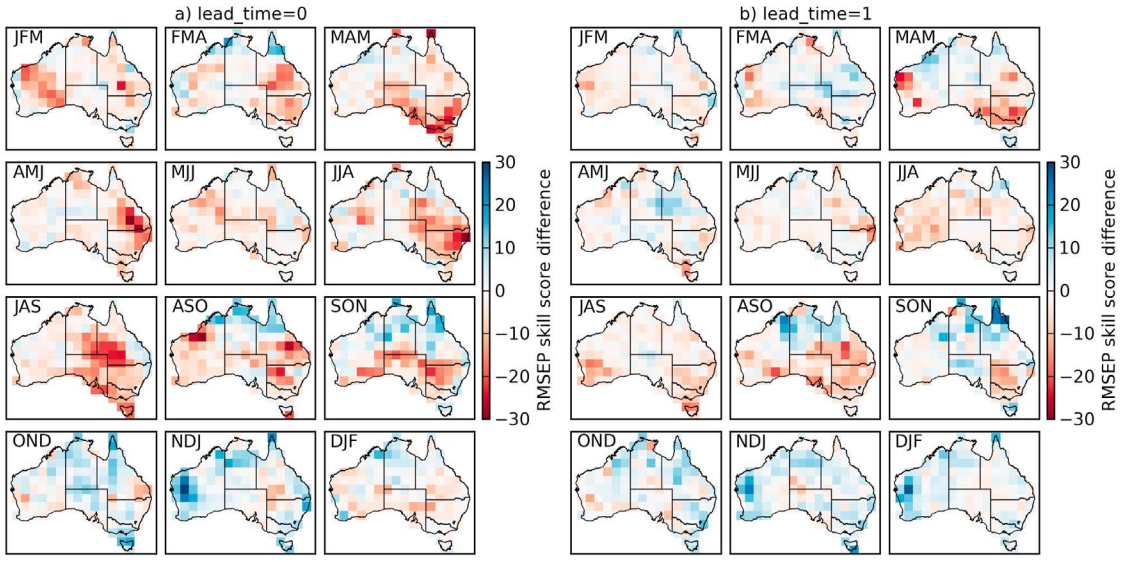

Figure 2. Differences in cross-validated RMSEP skill scores between statistical and dynamical BMA models for forecasting Australian seasonal rainfall (1980–2010). Negative areas (red) indicate where dynamical models are more skillful and positive areas (blue) indicated where statistical models are more skillful.

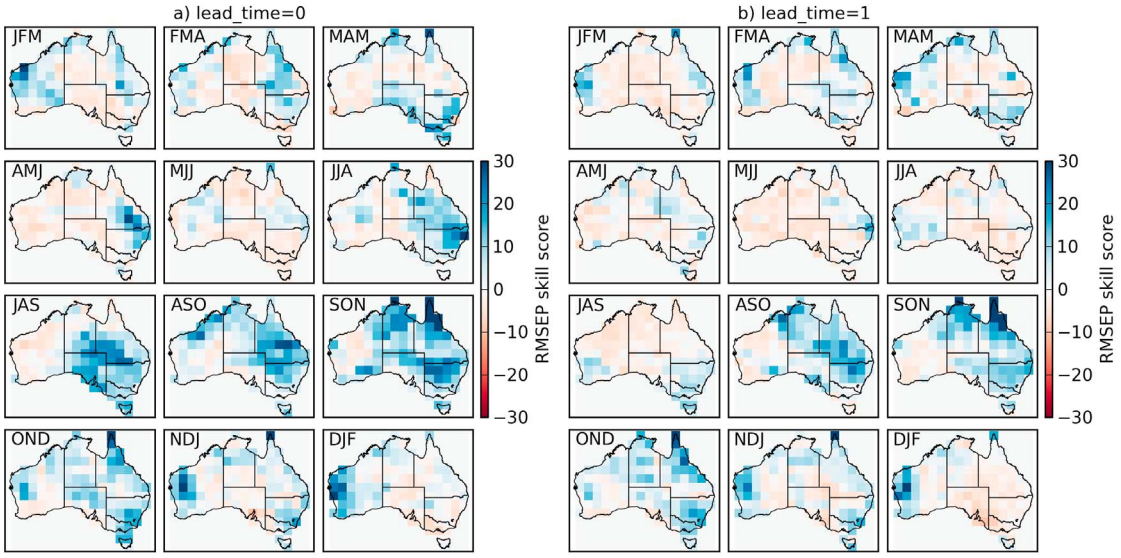

Figure 3. Cross-validated RMSEP skill scores of Australian seasonal rainfall forecasts (1980–2010) using statistical-dynamical BMA models.

CRPS-based skill scores are consistent with RMSEP, and therefore are not presented in this paper.

[24] While sharp and accurate forecasts are desirable, it is important that the forecasts have realistic uncertainty bounds or spread in the ensembles. To check forecast reliability and forecast sharpness, we plot an attributes diagram [*Hsu and Murphy*, 1986]. In the diagram, reliability is assessed by plotting the forecast probabilities of events against their observed frequencies. Forecast sharpness, i.e., the tendency for forecasts to move away from climatological averages, is also represented. In this study, we assess the forecasts of probability of exceeding the climatological median. As for skill assessment discussed in the last paragraph, the cross validation climatological medians consistent with the forecast models are used here also.

4. Results and Discussion

4.1. Relative Skills of Statistical, Dynamical and Statistical-Dynamical BMA Models

[25] It is useful to first compare the skills of the statistical models and the dynamical models. We apply BMA to three groupings of the models. We merge the statistical models as one group and the dynamical models as another group. The third group contains all statistical and dynamical models. In each case, the climatology model is also included as a candidate model. Figure 1 compares the overall forecast performance of the statistical, dynamical and statistical-dynamical BMA models in terms of the number of grid cells where the RMSEP skill scores exceed a range of thresholds. These numbers have been calculated by considering all events (i.e., all seasons and grid cells) pooled together and therefore they provide a good summary of overall forecast performance. We first note that the skill scores are typically quite modest (<30), highlighting the difficulties of seasonal rainfall forecasting. At both lead times, the statistical-dynamical BMA models have the highest number of grid cells exceeding each skill score threshold, clearly demonstrating that overall improvement is achieved by merging statistical and dynamical models. At lead time 0 (Figure 1a), the dynamical BMA models are observed to have more grid cells with positive skill than the statistical BMA models. However, at lead time 1 (Figure 1b), the statistical BMA models and the dynamical BMA are observed to have similar numbers of grid cells with positive skill. The

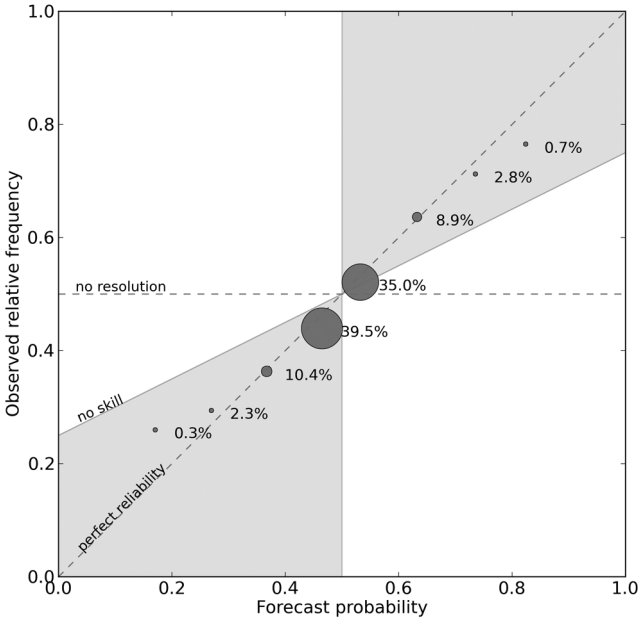

Figure 4. Attributes diagram for forecasts of probability of exceeding median rainfall. The diagram assesses reliability, resolution, sharpness and skill of the binary forecasts. All grid cells, seasons and lead times have been pooled together. Forecast probability is binned with width 0.1. Forecast sharpness is represented by the relative size of the dots, also written as the proportion of forecasts in the bin.

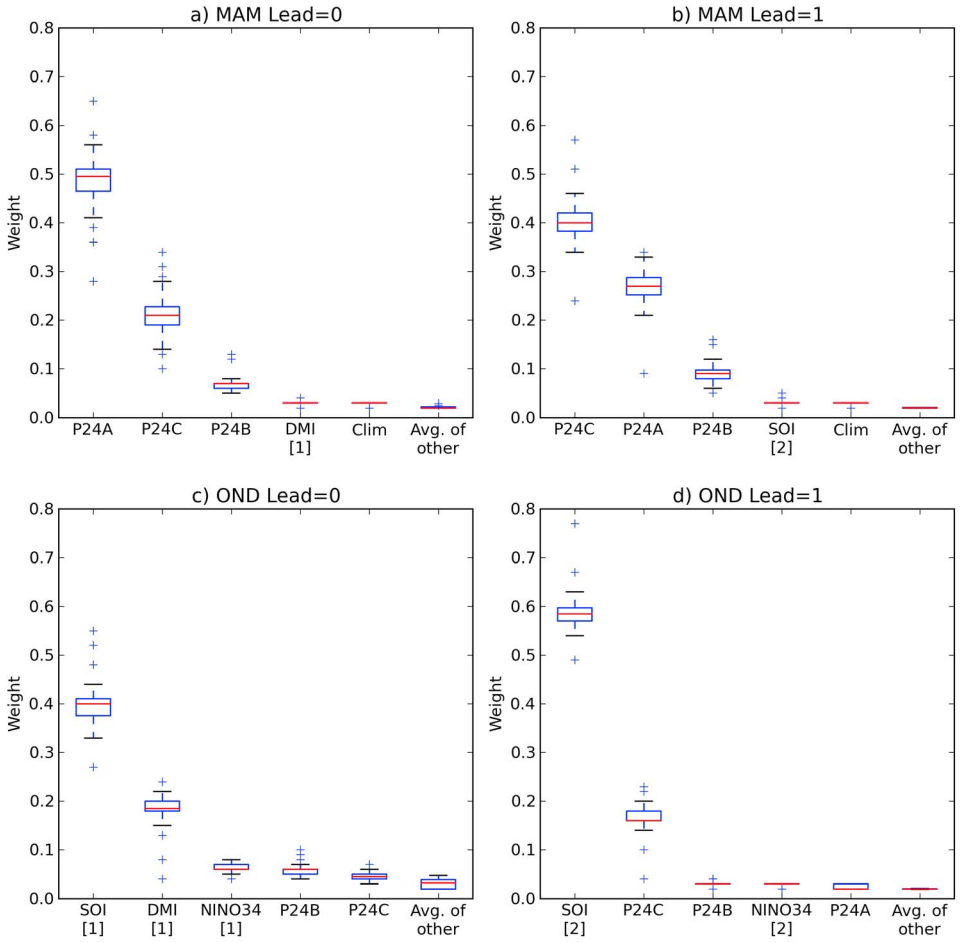

Figure 5. Box plots showing the distributions of model weights in cross-validation for a grid cell in southeastern Australia. The seasons MAM and OND and lead times of 0 and 1 month are shown. Number in brackets denotes lag time of statistical model. The box represents the interquartile range (IQR) with a line at the median. The whiskers reach the most extreme point within the 1.5*IQR range. The crosses outside the 1.5*IQR range are considered outliers.

benefit of merging is more apparent at lead time 1 than at lead time 0.

[26] We now analyze the spatial and temporal distributions of skill in more detail. The differences in skill between the statistical and dynamical BMA models at lead times of 0 and 1 month are plotted in Figures 2a and 2b, respectively. Positive (blue) values show extra skill achieved by the statistical BMA models and negative (red) values show extra skill achieved by the dynamical BMA models. If the difference is close to zero, the models are similarly skillful. The group with the most skill depends on the season and location and, to a lesser extent, lead time.

[27] Consider the lead time 0 forecasts (Figure 2a). During JFM - JAS, the dynamical models are skillful in more regions than the statistical models, particularly in eastern and southeastern Australia. It is expected that the dynamical models are more skillful than the statistical models in the first half of the year because of weak relationships between climate indices and Australian seasonal rainfall. Furthermore, coupled ocean–atmosphere dynamical models are likely to perform better in this period because they can capture and propagate changes in the oceanic-atmospheric circulation on much shorter (e.g., weekly) time scales [*Barnston et al.*, 2012]. In the latter part of the year (ASO-DJF), when the relationships between climate indices and Australian seasonal rainfall are known to be stronger, the statistical models achieve more positive differences in skill. In ASO and SON, the dynamical models are more skilled in parts of eastern and southern Australia, whereas the statistical models are more skilled in northern Australia. During OND - DJF, there is widespread spatial coverage of higher statistical model skill. It has been shown that there are strong empirical relationships between lagged climate indices and seasonal rainfall in these seasons [e.g., *Schepen et al.*, 2012]. However, it is not clear to us why the dynamical models perform so relatively poorly during this period.

[28] Consider now the lead time 1 forecasts (Figure 2b), arguably a more important result as this is the lead time typically used for operational forecasting. The pattern of skill difference is similar to the pattern at lead time 0 but

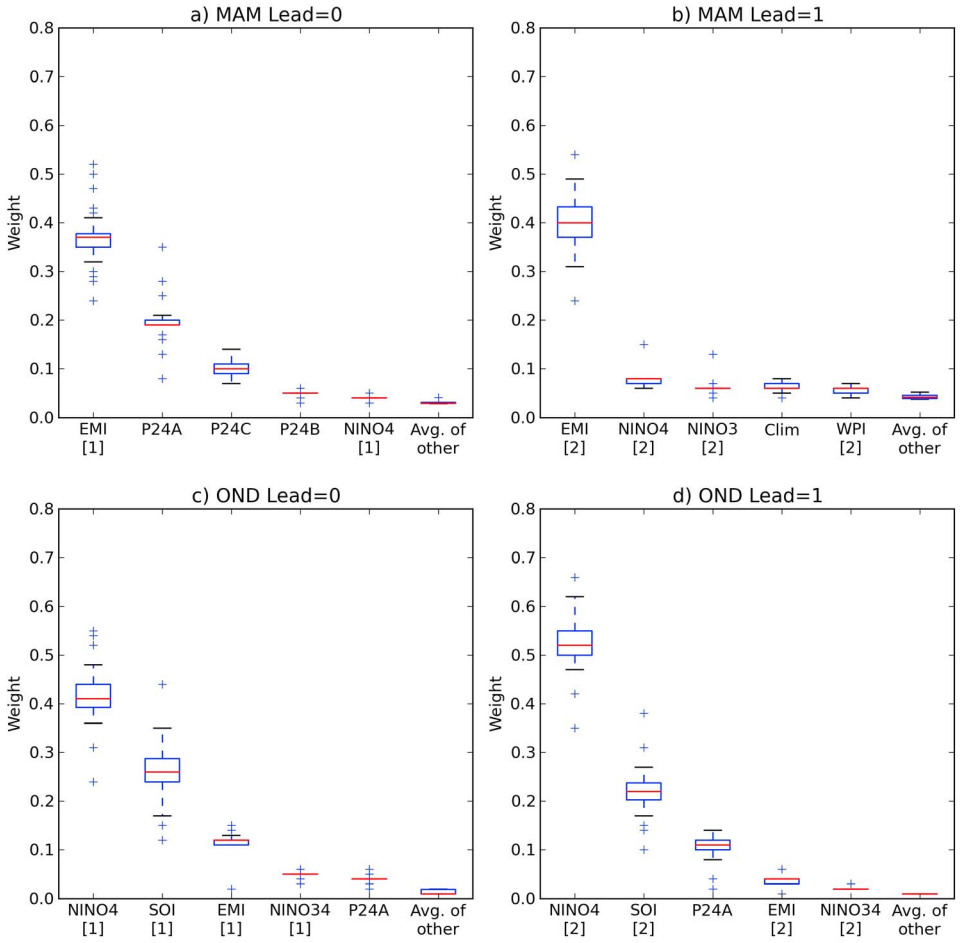

Figure 6. Box plots showing the distributions of model weights in cross-validation for a grid cell in northeastern Australia. The seasons MAM and OND and lead times of 0 and 1 month are shown. Number in brackets denotes lag time of statistical model. The box represents the interquartile range (IQR) with a line at the median. The whiskers reach the most extreme point within the 1.5*IQR range. The crosses outside the 1.5*IQR range are considered outliers.

there is an observed sharp drop in skill of the dynamical models in most seasons. During SON-DJF, the statistical models are clearly more skillful in many regions.

4.2. Skill and Reliability of the Merged Statistical-Dynamical Forecasts

[29] We now analyze the spatial and temporal coverage of skill achieved by merging the forecasts of statistical and dynamical models. The BMA approach will take advantage of the relative strengths of statistical and dynamical models depending on region, season and lead time. Figures 3a and 3b show the coverage of the RMSEP skill scores for the statistical-dynamical BMA models at lead times of 0 and 1, respectively. We note here that the overall spatiotemporal coverage of skill could be improved further by including additional models in the BMA, e.g., calibrated rainfall forecasts from other international coupled GCMs.

[30] The overall sharpness and reliability of the forecasts produced by the statistical-dynamical BMA models is assessed for forecasts of the probability of exceeding the climatological median. We expect the BJP modeling approach to produce reliable forecasts so this step is mainly a diagnostic check. An attributes diagram (Figure 4) has been produced after pooling all events together (i.e., all seasons, grid cells and lead times). Nearly all the points are aligned well to the 1:1 line. Overall the diagrams suggest that the forecast probabilities of events not exceeding the thresholds are consistent with the observed frequencies. In other words, the forecast distributions are reliable in representing forecast uncertainty spread.

4.3. Distributions of Model Weights

[31] We now analyze the distribution of weights among individual models for two grid cells in MAM and OND. We note that the weights are not necessarily indicative of skill, although they have more meaning in seasons and locations where there is skill. The distribution of weights among the models varies for each year in the cross-validation and therefore we present the distributions as a box-plot. The weights for the top five models plus the average weights of the remaining models are shown. The variation in weights in cross-validation is due to sampling variability of the limited

data used. In particular, only a small number of ENSO and Indian Ocean Dipole events are present in the data. Although the weights are stable for most events, as seen by the narrow interquartile range of the box-plots, there are a number of events that when excluded cause the weights to vary significantly, as seen by outliers on the box plots. Still, the cross-validation results suggest that our BMA approach is reasonably robust to these variations.

[32] The first grid cell is in southeastern Australia. In MAM (Figure 5a), the weight is distributed mainly among the dynamical models P24A, P24B and P24C for both lead times 0 and 1. In OND, the weight is distributed among the statistical models (Figures 5c and 5d). At lead time 0, the statistical models SOI and DMI are the models with the highest weights. This is consistent with the knowledge that both ENSO and the Indian Ocean influence southeastern Australian rainfall. At lead time 1, the SOI is still the model with the highest weight but the DMI does not receive any significant weight, reflecting its weak relationship with seasonal rainfall at longer lead times.

[33] Referring again to Figures 5a and 5b, it could be argued that because the three variants of POAMA have similar construction, the weights of the three models should be more similar. We previously trialed a more complex hierarchical version of the BMA that puts additional pressure on the POAMA models to be more equal, with little to no impact on the cross-validation results. In order to maintain objectivity, we prefer to let all weights be determined by the EM algorithm from a single pool of candidate models. Furthermore, other international GCMs could be added to the pool of candidate models in the future with no change to the procedure.

[34] The second grid cell is in northeastern Australia. In OND (Figures 6c and 6d), the weight is distributed mainly among the statistical models NINO4 and SOI for both lead times 0 and 1. This reflects the ENSO driven nature of rainfall in this location and season, and in particular a stronger relationship with western Pacific SSTs. In contrast, MAM is a season with low skill. When a model has low skill, its forecasts should fall back to climatology (i.e., the model has no resolution). At lead time 0 (Figure 6a), the weight is distributed among the statistical EMI model and the dynamical models. At lead time 1 (Figure 6b), the EMI model is again the model with the highest weight, however the remaining models don't have significant weight.

5. Supplementary Discussion

[35] At lead time 1, the BMA of statistical models has a similar number of grid cells exceeding each threshold as the dynamical models (Figure 1b). This is despite a two month lag in the climate index predictor variable. The coupled SST forecasts of POAMA are likely to beat persistence and produce good forecasts of climate indices at short lead times (0 and 1 month). Therefore, it is likely that using forecasts of climate indices in statistical models (bridging) will improve the results further. This is an appealing approach to improve the skill of GCM rainfall forecasts. We will report such results in a separate paper. Additionally, future research will investigate the use of all ensemble members and assess the benefit of including additional dynamical models from international modeling centers in the BMA.

[36] We established models and inferred the weights of each model at the grid scale. This has the potential to lead to spatially incoherent forecasts. One improvement that could be made to our method is to incorporate spatial relationships into the calculation of models and also the weights, for example, using similar approaches to *Robertson et al.* [2004].

6. Summary and Conclusions

[37] Statistical models and dynamical models (i.e., GCMs) are both currently widely used to forecast seasonal climate. Since the fundamentals of each type of system are different, there is an opportunity to draw on the strengths of both types of models in order to maximize the spatial and seasonal coverage of forecast skillfulness. In this paper, we applied a Bayesian model averaging method to merge forecasts of Australian seasonal rainfall from multiple statistical and dynamical models.

[38] Statistical forecasts were made using lagged climate indices as predictors and dynamical forecasts were obtained by statistically calibrating rainfall from three variants of the POAMA coupled ocean–atmosphere GCM. Forecast skill was assessed using leave-one-out cross-validation, which only provides an estimate of skill for forecasting independent events.

[39] As separate groups, the statistical and dynamical forecasts are skillful in certain seasons and locations. The statistical-dynamical BMA models represent an improvement in terms of maximizing spatial and temporal coverage of skillfulness compared to statistical or dynamical models alone. The statistical-dynamical BMA model forecasts are reliable in representing forecast uncertainty spread. In the context of limited data, the cross-validation BMA weights are found to be physically reasonable and acceptably stable.

[40] The BMA approach used here merges the forecasts of multiple statistical and dynamical models in a consistent and objective way. Consequently, forecasts from other international coupled GCMs could be readily included as candidate models. Additionally, the approach can be applied to other variables such as temperature or streamflow.

[41] **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. We thank our colleagues from the Centre for Australian Weather and Climate Research for providing hindcast data of the Predictive Ocean Atmosphere Model for Australia. We also thank three anonymous reviewers for their constructive comments, which resulted in a much improved paper.

References

Barnston, A. G., et al. (1999), Predictive skill of statistical and dynamical climate models in SST forecasts during the 1997–98 El Niño episode and the 1998 La Niña onset, *Bull. Am. Meteorol. Soc.*, *80*(2), 217–243, doi:10.1175/1520-0477(1999)080<0217:PSOSAD>2.0.CO;2.

Barnston, A. G., et al. (2012), Skill of real-time seasonal ENSO model predictions during 2002–2011: Is our capability increasing?, *Bull. Am. Meteorol. Soc.*, *93*, 631–651.

Casanova, S., and B. Ahrens (2009), On the weighting of multimodel ensembles in seasonal and short-range weather forecasting, *Mon. Weather Rev.*, *137*(11), 3811–3822, doi:10.1175/2009MWR2893.1.

Cheng, J., et al. (2006), Flexible background mixture models for foreground segmentation, *Image Vis. Comput.*, *24*(5), 473–482, doi:10.1016/j.imavis.2006.01.018.

Coelho, C. A. S., et al. (2004), Forecast calibration and combination: A simple Bayesian approach for ENSO, *J. Clim.*, *17*(7), 1504–1516, doi:10.1175/1520-0442(2004)017<1504:FCACAS>2.0.CO;2.

Doblas-Reyes, F. J., et al. (2005), The rationale behind the success of multi-model ensembles in seasonal forecasting–II. Calibration and combination, *Tellus, Ser. A*, *57*(3), 234–252, doi:10.1111/j.1600-0870.2005.00104.x.

Fawcett, R. J. B. (2008), Verification of the Bureau of Meteorology's seasonal forecasts: 2003–2005, *Aust. Meteorol. Mag.*, *57*(3), 273–278.

Fawcett, R., et al. (2005), A verification of publicly issued seasonal forecasts issued by the Australian Bureau of Meteorology: 1998–2003, *Aust. Meteorol. Mag.*, *54*, 1–13.

Feddersen, H., et al. (1999), Reduction of model systematic error by statistical correction for dynamical seasonal predictions, *J. Clim.*, *12*(7), 1974–1989, doi:10.1175/1520-0442(1999)012<1974:ROMSEB>2.0.CO;2.

Folland, C., et al. (1991), Prediction of seasonal rainfall in the Sahel region using empirical and dynamical methods, *J. Forecast.*, *10*(1–2), 21–56, doi:10.1002/for.3980100104.

Graham, R., et al. (2005), A performance comparison of coupled and uncoupled versions of the Met Office seasonal prediction general circulation model, *Tellus, Ser. A*, *57*(3), 320–339, doi:10.1111/j.1600-0870.2005.00116.x.

Hoeting, J. A., et al. (1999), Bayesian model averaging: A tutorial, *Stat. Sci.*, *14*(4), 382–401.

Hsu, W.-R., and A. H. Murphy (1986), The attributes diagram A geometrical framework for assessing the quality of probability forecasts, *Int. J. Forecast.*, *2*(3), 285–293, doi:10.1016/0169-2070(86)90048-8.

Jones, D. A., et al. (2009), High-quality spatial climate data-sets for Australia, *Aust. Meteorol. Oceanogr. J.*, *58*(4), 233–248.

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

Landsea, C. W., and J. A. Knaff (2000), How much skill was there in forecasting the very strong 1997–98 El Niño?, *Bull. Am. Meteorol. Soc.*, *81*(9), 2107–2119, doi:10.1175/1520-0477(2000)081<2107:HMSWTI>2.3.CO;2.

Lim, E.-P., et al. (2009), Dynamical forecast of inter-El Niño variations of tropical SST and Australian spring rainfall, *Mon. Weather Rev.*, *137*(11), 3796–3810, doi:10.1175/2009MWR2904.1.

Lim, E.-P., et al. (2011), Dynamical, statistical-dynamical, and multimodel ensemble forecasts of Australian spring season rainfall, *Mon. Weather Rev.*, *139*(3), 958–975, doi:10.1175/2010MWR3399.1.

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.

Palmer, T., et al. (2004), Development of a European multi-model ensemble system for seasonal to inter-annual prediction (DEMETER), *Bull. Am. Meteorol. Soc.*, *85*(6), 853–872, doi:10.1175/BAMS-85-6-853.

Raftery, A. E., et al. (2005), Using Bayesian model averaging to calibrate forecast ensembles, *Mon. Weather Rev.*, *133*(5), 1155–1174, doi:10.1175/MWR2906.1.

Rajeevan, M., et al. (2007), New statistical models for long-range forecasting of southwest monsoon rainfall over India, *Clim. Dyn.*, *28*(7), 813–828, doi:10.1007/s00382-006-0197-6.

Risbey, J. S., et al. (2009), On the remote drivers of rainfall variability in Australia, *Mon. Weather Rev.*, *137*(10), 3233–3253, doi:10.1175/2009MWR2861.1.

Robertson, A. W., et al. (2004), Improved combination of multiple atmospheric GCM ensembles for seasonal prediction, *Mon. Weather Rev.*, *132*(12), 2732–2744, doi:10.1175/MWR2818.1.

Robertson, D. E., and Q. J. Wang (2012), A Bayesian approach to predictor selection for seasonal streamflow forecasting, *J. Hydrometeorol.*, *13*(1), 155–171, doi:10.1175/JHM-D-10-05009.1.

Saha, S., et al. (2006), The NCEP Climate Forecast System, *J. Clim.*, *19*(15), 3483–3517, doi:10.1175/JCLI3812.1.

Schepen, A., et al. (2012), Evidence for using lagged climate indices to forecast Australian seasonal rainfall, *J. Clim.*, *25*(4), 1230–1246, doi:10.1175/JCLI-D-11-00156.1.

Shabbar, A., and A. G. Barnston (1996), Skill of seasonal climate forecasts in Canada using canonical correlation analysis, *Mon. Weather Rev.*, *124*(10), 2370–2385, doi:10.1175/1520-0493(1996)124<2370:SOSCFI>2.0.CO;2.

Smith, T. M., et al. (2008), Improvements to NOAA's historical merged land-ocean surface temperature analysis (1880–2006), *J. Clim.*, *21*(10), 2283–2296, doi:10.1175/2007JCLI2100.1.

Troup, A. J. (1965), The Southern Oscillation, *Q. J. R. Meteorol. Soc.*, *91*(390), 490–506, doi:10.1002/qj.49709139009.

Wang, G., et al. (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.

Wang, Q. J., D. E. Robertson, 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.

Wang, Q. J., et al. (2012), Merging seasonal rainfall forecasts from multiple statistical models through Bayesian model averaging, *J. Clim.*, *25*, 5524–5537, doi:10.1175/JCLI-D-11-00386.1.

Yasuda, T., et al. (2007), Asian monsoon predictability in JMA/MRI seasonal forecast system, *CLIVAR Exch.*, *43*, 18–24.

Yeo, I. K., and R. A. Johnson (2000), A new family of power transformations to improve normality or symmetry, *Biometrika*, *87*(4), 954–959, doi:10.1093/biomet/87.4.954.

Zivkovic, Z., and F. van der Heijden (2004), Recursive unsupervised learning of finite mixture models, *IEEE Trans. Pattern Anal. Mach. Intell.*, *26*(5), 651–656.