Calibration, Bridging, and Merging to Improve GCM Seasonal Temperature Forecasts in Australia

ANDREW SCHEPEN

CSIRO Land and Water, Dutton Park, Queensland, Australia

Q. J. WANG

CSIRO Land and Water, Clayton, Victoria, Australia

YVETTE EVERINGHAM

James Cook University, Townsville, Queensland, Australia

(Manuscript received 1 November 2015, in final form 1 March 2016)

ABSTRACT

There are a number of challenges that must be overcome if GCM forecasts are to be widely adopted in climate-sensitive industries such as agriculture and water management. GCM outputs are frequently biased relative to observations and their ensembles are unreliable in conveying uncertainty through appropriate spread. The calibration, bridging, and merging (CBaM) method has been shown to be an effective tool for postprocessing GCM rainfall forecasts to improve ensemble forecast attributes. In this study, CBaM is modified and extended to postprocess seasonal minimum and maximum temperature forecasts from the POAMA GCM in Australia. Calibration is postprocessing GCM forecasts using a statistical model. Bridging is producing additional forecasts using statistical models that have other GCM output variables (e.g., SST) as predictors. It is demonstrated that merging calibration and bridging forecasts through CBaM effectively improves the skill of POAMA seasonal minimum and maximum temperature forecasts for Australia. It is demonstrated that CBaM produces bias-corrected forecasts that are reliable in ensemble spread and reduces forecasts to climatology when there is no evidence of forecasting skill. This work will help enable the adoption of GCM forecasts by climate-sensitive industries for quantitative modeling and decision-making.

1. Introduction

Seasonal climate forecasts from coupled ocean–atmosphere general circulation models (GCMs) are routinely issued by meteorological forecasting centers around the world (Troccoli 2010). Forecasts of climate variables such as temperature and rainfall are highly sought after by climate-sensitive industries such as agriculture, water management, mining, and insurance. Not only can quality seasonal forecasts forewarn of changes in climatic conditions, but they can influence decision-making and, in turn, affect business and environmental outcomes (e.g., Everingham et al. 2008, 2012).

In mid-2013, the Australian Bureau of Meteorology first issued official seasonal climate forecasts based on outputs from the Predictive Ocean Atmosphere Model for Australia (POAMA). POAMA is a coupled ocean–atmosphere GCM set up specifically for seasonal forecasting (Hudson et al. 2013; Marshall et al. 2014; Wang et al. 2011). GCMs produce forecasts for a wide array of oceanic, atmospheric, and land surface variables. Forecasts for local climate conditions can, therefore, be viewed in the context of an overall global climate forecast. For example, a seasonal climate forecast for an important agricultural region can be considered in the wider context of the El Niño–Southern Oscillation (ENSO), subsurface ocean temperatures, and polar circulations. GCMs are able to provide forecasts many months into the future. For example, POAMA forecasts up to nine months ahead from the model initialization date. Appreciably then, GCMs are potentially useful for

Corresponding author address: Andrew Schepen, 41 Boggo Rd., Dutton Park, QLD 4102, GPO Box 2583, Brisbane, QLD 4001, Australia.
E-mail: andrew.schepen@csiro.au

DOI: 10.1175/MWR-D-15-0384.1

forecasting the evolution of large-scale climatic patterns such as ENSO.

There are a number of challenges that must be overcome if GCMs are to be widely adopted by the aforementioned industries. GCM outputs are frequently biased relative to observations and their ensembles are unreliable in conveying uncertainty through appropriate spread. Barnston et al. (2015) provide an up-to-date discussion of this. The consequence is that raw GCM forecasts are unsuitable as direct inputs to the types of quantitative models used in, for example, agriculture and water management. For practical applications in such industries, GCM forecasts can be statistically postprocessed to correct bias and ensemble uncertainty spread (e.g., Feddersen et al. 1999; Gneiting et al. 2005).

Simple pooling of forecasts to create multimodel ensembles—such as the North American Multimodel Ensemble (Kirtman et al. 2014)—can improve forecast accuracy by cancelling uncorrelated model errors. However, multimodel forecasts can still have limited skill in discriminating climate events and cannot be assured to reliably convey forecast uncertainty. In the 1990s and first decade of the 2000s, postprocessing methods predominantly used singular value decomposition analysis and canonical correlation analysis in conjunction with multiple linear regression equations (Bartman et al. 2003; Feddersen et al. 1999; Lim et al. 2011). However, statistical corrections can be unstable due to short data records and therefore forecast uncertainty requires more rigorous treatment. Lim et al. (2011) suggested that Bayesian approaches could be better suited to postprocessing GCM outputs. Over the past decade, Bayesian methods have become increasingly popular in hydro-climate forecasting in order to better capture forecast uncertainty, improve reliability, and combine multiple model forecasts (e.g., Coelho et al. 2004; DeChant and Moradkhani 2014; Dutton et al. 2013; Herr and Krzysztofowicz 2015; Jo et al. 2012; Luo et al. 2007).

Even though Bayesian methods can improve forecast accuracy and reliability through forecast calibration and combination, forecast skill can still be low after postprocessing. Individual GCMs differ in their representation of large-scale climate patterns such as ENSO and the Indian Ocean dipole (Barnston and Tippett 2013; Shi et al. 2012). Additionally, individual models vary in their representation of the teleconnections between the large-scale climate patterns and local rainfall and temperature (e.g., Kim et al. 2012). In Australia, several studies have analyzed the relationships between POAMA SST patterns and rainfall and temperature (Lim et al. 2009; White et al. 2014; Zhao and Hendon 2009). Forecast skill can be unrealized if teleconnections are poorly represented. Thus statistical postprocessing methods that can induce additional skill by overcoming poorly represented teleconnections will be at an advantage.

Recent work has led to the calibration, bridging, and merging (CBaM) method for postprocessing GCM rainfall forecasts (Schepen and Wang 2014; Schepen et al. 2014). CBaM makes the best use of GCM outputs by capturing and combining the information available in multiple output fields. Calibration is postprocessing GCM forecasts using a statistical model. In CBaM, the Bayesian joint probability modeling approach (Wang and Robertson 2011; Wang et al. 2009) is used. Bridging is the production of alternative forecasts, also through a statistical model, that has the GCM's forecasts of climate indices as predictors (Lim et al. 2011; Schepen and Wang 2014; Schepen et al. 2014). Merging is the optimal combination of calibration and bridging forecasts from one or more GCMs. CBaM uses model averaging (Schepen and Wang 2015; Schepen et al. 2012; Wang et al. 2012) for merging. The current preferred merging method is Bayesian model averaging (BMA). The BMA method, detailed by Wang et al. (2012), builds on the earlier BMA methods by Hoeting et al. (1999) and Raftery et al. (2005). CBaM has been demonstrated to produce rainfall forecasts that are reliable in ensemble spread and to maximize skill through the combination of calibration and bridging. CBaM has been tested for rainfall forecasting in Australia (Hawthorne et al. 2013; Schepen and Wang 2013, 2014; Schepen et al. 2014) and China (Peng et al. 2014). CBaM needs to be tested as an effective tool in postprocessing a wider range of climate variables. For example, a crop forecasting model may take temperature as an input alongside rainfall. The Australian Bureau of Meteorology makes publicly available forecasts of minimum and maximum temperature. However, for applications such as the aforementioned crop modeling, it is essential that the necessary climate variables are postprocessed in a consistent way. It is therefore a logical next step to test CBaM on minimum and maximum temperature forecasts. Through testing of CBaM for postprocessing temperature, we refine the methods to produce a more widely applicable and useful version of CBaM.

In this study, we investigate the application of CBaM to postprocess forecasts of POAMA seasonal (three month) minimum temperature (Tmin) and maximum temperature (Tmax) in Australia. We extend the work of Schepen et al. (2014) who investigated CBaM for postprocessing POAMA forecasts of Australian seasonal rainfall. POAMA is still the current operational forecasting GCM in Australia. This work targets forecasts issued one month in advance. We consider the entire Australian continent and forecasts issued monthly. The analysis is therefore consistent with the type of climate outlooks that are issued operationally in Australia. Although the focus is on development of CBaM using POAMA and Australian data, it is anticipated that the method will be transferrable to other GCMs and countries.

The remainder of this paper is organized as follows. Section 2 details the GCM and observed data used in this study. Section 3 details the CBaM and verification methods. Section 4 presents the results in a straightforward manner. Section 5 contains a detailed discussion of the results and outlines avenues for future research. Section 6 summarizes the study and points out the main conclusions.

2. Data

a. GCM forecasts of temperature

The GCM used in this investigation is POAMA. POAMA version M2.4 is a coupled ocean–atmosphere GCM set up specifically for seasonal forecasting in Australia (Hudson et al. 2013; Marshall et al. 2014; Wang et al. 2011). The atmospheric spatial resolution is approximately 250 km. POAMA comprises coupled atmospheric, oceanic, and land surface modules as well as ocean and atmosphere–land initialization schemes. Atmospheric and land initial conditions are created by forcing POAMA's atmospheric module with observed SST and nudging winds, temperatures, and humidity toward an observationally based analysis (Hudson et al. 2011). Ocean initial conditions are produced separately using an ensemble ocean data assimilation system (PEODAS; Yin et al. 2011). Final perturbed ocean and atmosphere initial conditions are created jointly using a coupled breeding technique (Hudson et al. 2013). The number of ensemble members is 33.

Long periods of data (at least 25 years) are needed for establishing CBaM models and for forecast verification. Fortunately, POAMA hindcasts (simulated forecasts of past events) are available and used in this study. Specifically, we use hindcasts initialized from 1 December 1981 to 1 November 2010. These start dates correspond to 1-month ahead forecasts for the period January–March (JFM) 1982 to December–February (DJF) 2010–11 (hereafter all 3-month ranges will be abbreviated by the first letter of each month).

The variables of interest in this study are minimum temperature (Tmin) and maximum temperature (Tmax). Tmin is the minimum temperature in the preceding 24-h period. Similarly, Tmax is the maximum temperature in the preceding 24-h period. The POAMA hindcast dataset makes available monthly Tmin and Tmax output fields whereby monthly values represent a monthly average of daily values. Seasonal forecasts of Tmin and Tmax are obtained by averaging monthly values across three sequential months.

b. GCM forecasts of climate indices

Climate patterns in the Pacific and Indian Oceans affect Australian climate in certain parts of the year; climate indices are therefore useful indicators or predictors of Australian climate (e.g., Risbey et al. 2009; Schepen et al. 2012). POAMA forecasts of climate indices are calculated from ensemble-mean POAMA sea surface temperature (SST) fields. The indices represent future SST conditions in the Pacific and Indian Oceans for the forecast period. The Niño-3, Niño-3.4, Niño-4, and El Niño Modoki (EMI) indices are used to capture ENSO teleconnections from the Pacific region. The Indian Ocean dipole mode index (DMI), its poles, and the Indonesian SST index capture teleconnections from the Indian Ocean region. Seasonal forecasts of climate indices are obtained by averaging monthly values across three sequential months.

c. Observed temperature

Seasonal Tmin and Tmax values are obtained from the Australian Water Availability Project's 0.05° × 0.05° gridded climate dataset (Jones et al. 2009). The data are regridded to match the POAMA grid of approximately 2.5° resolution. To match the period of M2.4 hindcasts, observed data from January 1982 to February 2011 are used.

3. Methods

The CBaM method for postprocessing seasonal climate forecasts links together Bayesian joint probability (BJP) modeling (Wang and Robertson 2011; Wang et al. 2009) and BMA (Wang et al. 2012). A calibration model is a BJP model for correcting bias and ensemble spread in raw GCM forecasts of a climate variable. A bridging model is a BJP model that produces alternative forecasts of the climate variable by using predictors derived from other GCM output fields. Each of the bridging models is established with a single climate index predictor. CBaM BJP models have a bivariate normal distribution at their core, and allow data transformation in fitting the model. Inverse transformation is used in forecasting mode, which is achieved by conditioning the model on new predictor values. BJP models produce ensemble forecasts, which can be merged using BMA in a separate step. The BMA model assigns weights to BJP calibration and bridging models based on performance for predicting historical events. The BMA method described by

Wang et al. (2012) includes a weight-stabilizing prior and separates BJP model establishment from combination. The final BMA merged forecast is obtained by sampling from the calibration and bridging forecasts in proportion to the BMA weights.

The methods for establishing BJP and BMA models and their use for ensemble forecasting are described in mathematical detail in the subsequent subsections. The methods for verification follow. The experimental process for this study is outlined in Fig. 1.

a. BJP model description

A bivariate BJP model relates a predictor variable x and a predictand variable y. The relationship of the variables is assumed to conform to a continuous bivariate normal distribution after allowing for data transformation. If g and h are the transformed variables for x and y, respectively, then

$$\begin{bmatrix} g \\ h \end{bmatrix} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma}), \qquad (1.1)$$

where and $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ are, respectively, a vector of theoretical means and the theoretical covariance matrix. More precisely,

$$\boldsymbol{\mu} = \begin{bmatrix} \mu_g \\ \mu_h \end{bmatrix} \qquad (1.2)$$

and

$$\boldsymbol{\Sigma} = \begin{bmatrix} \sigma_g^2 & r_{gh}\sigma_g\sigma_h \\ r_{gh}\sigma_g\sigma_h & \sigma_h^2 \end{bmatrix}, \qquad (1.3)$$

where, omitting subscripts, μ is a mean, σ is a standard deviation, and r is a correlation coefficient.

Variables that contain both positive and negative values, as in this study, are transformed using the Yeo–Johnson transform (Yeo and Johnson 2000). For the predictor,

$$g = \begin{cases} [(x+1)^{\lambda_x} - 1]/\lambda_x & \lambda_x \neq 0, x \geq 0 \\ \log(x+1) & \lambda_x = 0, x \geq 0 \\ -[(-x+1)^{2-\lambda_x} - 1]/(2-\lambda_x) & \lambda_x \neq 2, x < 0 \\ -\log(-x+1) & \lambda_x = 2, x < 0 \end{cases}. \qquad (1.4)$$

The predictand y is similarly transformed, with λ_y, to yield h. The full collection of model parameters is therefore $\boldsymbol{\theta} = \{\lambda_x, \lambda_y, \mu_g, \mu_h, \sigma_g, \sigma_h, r_{gh}\}$.

The bivariate normal distribution can be conditioned on a value of g, which is related to the conditional distribution of y through

$$p(y \mid x, \boldsymbol{\theta}) = J_{h\to y}\, p(h \mid g, \boldsymbol{\theta}), \qquad (1.5)$$

where $J_{h\to y}$ is the Jacobian determinant of the transformation;

$$J_{h\to y} = \begin{cases} (y+1)^{\lambda_y - 1} & y \geq 0 \\ (-y-1)^{1-\lambda_y} & y < 0 \end{cases} \qquad (1.6)$$

and

$$h \mid g, \boldsymbol{\theta} \sim N\left[\mu_h + r_{gh}\frac{\sigma_h}{\sigma_g}(g - \mu_g), (1 - r_{gh}^2)\sigma_g^2\right]. \qquad (1.7)$$

For new values of x or, equivalently g, Eq. (1.5) represents the forecasting equation.

In the Bayesian formulation, the posterior distribution of the parameters is

$$p(\boldsymbol{\theta} \mid \mathbf{D}) \propto p(\boldsymbol{\theta})p(\mathbf{D} \mid \boldsymbol{\theta}) = p(\boldsymbol{\theta}) \prod_{t=1}^{T} p(\mathbf{D}^t \mid \boldsymbol{\theta}), \qquad (1.8)$$

where $p(\boldsymbol{\theta})$ is the prior distribution of the parameters, $p(\mathbf{D} \mid \boldsymbol{\theta})$ is the likelihood, and $\mathbf{D}^t$ contains the pair of predictor and predictand values for time period $t(t = 1, 2, \ldots, T)$. The prior distribution of parameters is specified as

$$p(\boldsymbol{\theta}) \propto p(\lambda_x)p(\lambda_y)p(r_{gh})p(\mu_g, \sigma_g)p(\mu_h, \sigma_h). \qquad (1.9)$$

The terms in the prior are given in the next section, which covers parameter inference.

A forecast incorporating parameter uncertainty corresponds to the posterior predictive density. Following from Eqs. (1.5) and (1.8), the posterior predictive density is obtained by integrating over the parameter space:

$$f(y) = \int p(y \mid x, \boldsymbol{\theta})p(\boldsymbol{\theta} \mid \mathbf{D})d\boldsymbol{\theta}. \qquad (1.10)$$

b. BJP inference

1) Prior specification

Prior to parameter inference, the data are rescaled to account for the fact that the effect of transformation for a given λ depends on the data range. In this study we rescale each variable by subtracting the sample mean and dividing by the sample standard deviation. A preliminary scaling of the data allows for more efficient and robust transformation parameter inference. Subsequently, in this study, the priors for the transformation parameters are specified to follow a normal

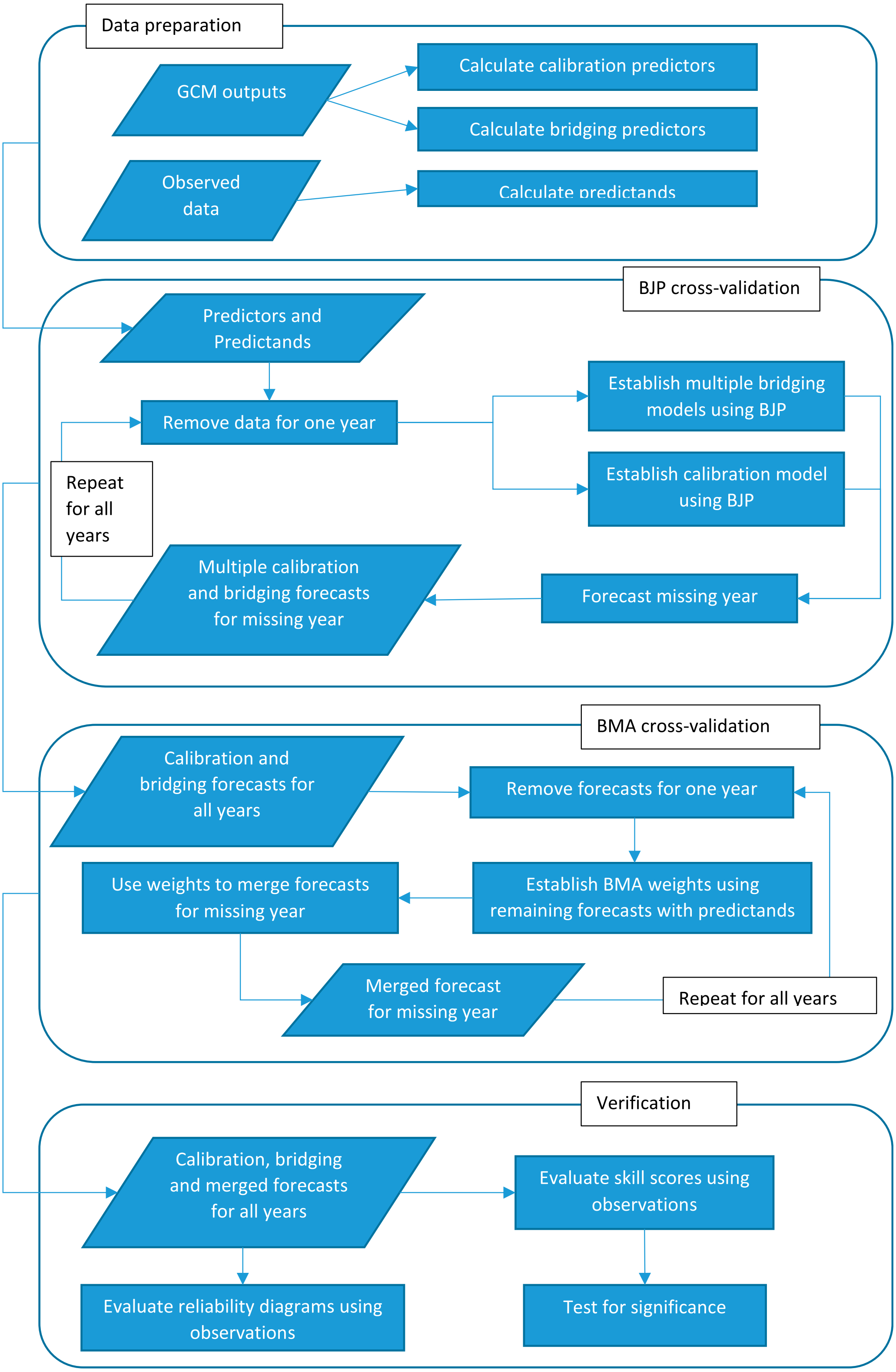
Data preparation
GCM outputs
Calculate calibration predictors
Calculate bridging predictors
Observed data
Calculate predictands
BJP cross-validation
Predictors and Predictands
Remove data for one year
Establish multiple bridging models using BJP
Establish calibration model using BJP
Repeat for all years
Multiple calibration and bridging forecasts for missing year
Forecast missing year
BMA cross-validation
Calibration and bridging forecasts for all years
Remove forecasts for one year
Use weights to merge forecasts for missing year
Establish BMA weights using remaining forecasts with predictands
Merged forecast for missing year
Repeat for all years
Verification
Calibration, bridging and merged forecasts for all years
Evaluate skill scores using observations
Evaluate reliability diagrams using observations
Test for significance

FIG. 1. Process for producing and verifying cross-validation CBaM forecasts for a hindcast period.

distribution with a mean of 1.0 and a standard deviation of 0.4

$$p(\lambda_x)p(\lambda_y) \propto N(1, 0.4^2)N(1, 0.4^2) \tag{1.11}$$

with density calculated accordingly. The prior on the transformation parameters given in Eq. (1.11) favors weak transformation (i.e., λ closer to 1.0), and maintains λ within a reasonable range. The priors for the mean and standard deviation parameters are

$$p(\mu_g, \sigma_g)p(\mu_h, \sigma_h) \propto \frac{1}{\sigma_g}\frac{1}{\sigma_h}. \tag{1.12}$$

The prior for the correlation coefficient is

$$p(r_{gh}) \propto 1. \tag{1.13}$$

2) Inference of fixed transformation parameters

In this study, we opt for fixed data transformations, as it is our experience that this leads to a more stable forecasting model. The transformation parameters λ_x and λ_y are inferred individually by considering the marginal distributions of the predictor and predictand. By way of example, in the case of the predictor, the posterior distribution of the parameters is

$$p(\lambda_x, \mu_g, \sigma_g \,|\, \mathbf{D}_x) \propto p(\lambda_x, \mu_g, \sigma_g)\prod_{t=1}^{T} p(\mathbf{D}_x^t \,|\, \lambda_x, \mu_g, \sigma_g), \tag{1.14}$$

where $\mathbf{D}_x$ contains the predictor data. A point estimate of the parameters is found by maximizing the rhs of Eq. (1.14) using the shuffled complex evolution (SCE) optimizer (Duan et al. 1992). This corresponds to the maximum a posteriori (MAP) estimate of the parameters. The MAP estimate is preferred over the maximum likelihood estimate because the number of the data records is small. For the purposes of transformation only λ_x is retained. The same steps are followed to find λ_y.

3) Inference of bivariate normal parameters

With the transformation parameters λ_x and λ_y held fixed, a Markov chain Monte Carlo inference is then made for the remaining bivariate normal distribution parameters $\boldsymbol{\theta}_{\text{bn}} = \{\mu_g, \mu_h, \sigma_g, \sigma_h, r_{gh}\}$ to capture uncertainty in the model parameters. Since the transformation parameters are held fixed, the posterior distribution of Eq. (1.8) is modified to

$$\begin{aligned} p(\boldsymbol{\theta}_{\text{bn}} \,|\, \mathbf{D}; \lambda_x, \lambda_y) &\propto p(\boldsymbol{\theta}_{\text{bn}})p(\mathbf{D} \,|\, \boldsymbol{\theta}_{\text{bn}}; \lambda_x, \lambda_y) \\ &= p(\boldsymbol{\theta}_{\text{bn}})\prod_{t=1}^{T} p(\mathbf{D}^t \,|\, \boldsymbol{\theta}_{\text{bn}}; \lambda_x, \lambda_y). \end{aligned} \tag{1.15}$$

A Metropolis sampler is used to obtain numerous sets of parameters, representing a numerical sample of the posterior distribution.

c. *BJP forecast ensemble generation*

In this study, 1000 parameter sets are sampled, representing the posterior distribution of $\boldsymbol{\theta}_{\text{bn}}$. The parameter sets are used to generate a forecast ensemble representing the posterior predictive density given by Eq. (1.10). Given the large number of parameter sets, we generate one forecast ensemble member per parameter set. With reference to Eq. (1.5), one ensemble member is generated from $p(h \,|\, g, \boldsymbol{\theta})$ and then back transformed using the inverse of Eq. (1.4). Hence, each BJP model yields a forecast ensemble of size 1000.

d. *BMA model description*

The specific version of BMA used in this study was developed by Wang et al. (2012). A BMA-merged forecast is a weighted average of multiple individual model density forecasts:

$$f_{\text{BMA}}(y) = \sum_{k=1}^{K} w_k f_k(y), \tag{1.16}$$

where w_k is the weight for model $k(k = 1, 2, \ldots, K)$ and f_k is the density forecast for model k.

If the model weights are collected in $\boldsymbol{\pi} = [w_1, w_2, \ldots, w_K]$ and the BMA forecasts for events $t = 1, 2, \ldots, T$ are collected in $\mathbf{F} = [f_{\text{BMA}}^1(y), f_{\text{BMA}}^2(y), \ldots, f_{\text{BMA}}^T(y)]$, then the posterior distribution of the weights is

$$p(\boldsymbol{\pi} \,|\, \mathbf{F}, \mathbf{D}_y) \propto p(\boldsymbol{\pi})p(\mathbf{D}_y \,|\, \mathbf{F}), \tag{1.17}$$

where $p(\boldsymbol{\pi})$ is the prior distribution of the weights, $p(\mathbf{D}_y \,|\, \mathbf{F})$ is the likelihood, and $\mathbf{D}_y$ contains the predictand data.

e. *BMA inference*

1) Prior specification

To encourage more even weights among the models, the prior is specified as a Dirichlet distribution with concentration parameter $\alpha = 1 + \alpha_0/K$ and $\alpha_0 = 0.5$. This is the same prior as specified by Wang et al. (2012):

$$p(\boldsymbol{\pi}) \propto \prod_{k=1}^{K} (w_k)^{\alpha - 1}. \tag{1.18}$$

2) Model weights inference

A MAP estimate of the weights is obtained by maximizing the posterior distribution of the weights. To estimate the weights according to the model predictive abilities rather than fitting abilities, cross-validation predictive densities $f_k^{(t)}(y)$ (Shinozaki et al. 2010; Wang et al. 2012) are used in the likelihood calculation. It follows that the posterior distribution of the weights is

$$p(\boldsymbol{\pi}\,|\,\mathbf{F},\mathbf{D}_y) \propto \prod_{k=1}^{K}(w_k)^{\alpha-1}\prod_{t=1}^{T}\sum_{k=1}^{K} w_k f_k^{(t)}(d_y^t), \qquad (1.19)$$

where d_y^t is the predictand value for event t.

A modified iterative expectation–maximization (E-M) algorithm that accounts for the effect of the prior (Cheng et al. 2006; Wang et al. 2012; Zivkovic and van der Heijden 2004) is used to find the MAP estimates of the weights.

3) BMA forecast ensemble merging

BJP modeling yields a large number of ensemble members for each model forecast. The BMA-merged forecast ensemble is obtained by randomly drawing ensemble members from each model's forecast ensemble according to the model weights. For example, if the number of ensemble members for each forecast is 1000 and $w_k = 0.5$ then 500 ensemble members will sampled from the ensemble representing $f_k(y)$. After drawing ensemble members for all models, the result is a forecast ensemble of size 1000 representing $f_{\mathrm{BMA}}(y)$.

f. Verification

In this study, we assess the performance of forecasts for the period JFM 1982 to DJF 2010–11. Forecasts are produced for each grid cell and season independently. The skill of calibration, bridging, and merging forecasts are compared to evaluate the benefit of bridging over calibration alone. We also contrast the skill and reliability of CBaM postprocessed forecasts with mean-corrected raw forecasts to demonstrate the need for more sophisticated postprocessing methods.

The mean-corrected and CBaM forecast results are cross validated using leave-one-year-out cross validation. For each historical forecast event to be tested, the data points for the year to be forecast are omitted from the BJP and BMA inferences. This procedure is repeated for each event in the historical record. Each forecast event to be verified is therefore made with a separate cross-validation model. Although leave-one-year-out cross validation does not assure testing of perfectly independent events, we consider it to be suitable for the conclusions drawn in this paper; particularly as we are analyzing the relative performance of different postprocessing measures in a consistent way.

Forecasts from CBaM are probabilistic. We therefore assess the essential attributes of probabilistic forecasts including accuracy, sharpness, resolution, and reliability. The continuous ranked probability score (CRPS; Matheson and Winkler 1976) is used to assess full forecast probability distributions, therefore involving forecast sharpness as well as forecast accuracy. For convenience we define CRPS as the average CRPS across forecast time periods $t (t = 1, 2, \ldots, T)$,

$$\mathrm{CRPS} = \frac{1}{T}\sum_{t=1}^{T}\int [F(y^t) - H(y^t - d_y^t)]^2\, dy^t\,, \qquad (1.20)$$

where y^t is the forecast variable for time period t, d_y^t is the corresponding predictand value, $F(\cdot)$ is the forecast CDF, and $H(\cdot)$ is the Heaviside step function, which equals 0 if $y^t < d_y^t$ and equals 1 otherwise. To assess forecast value, the CRPS score is converted to a skill score, to measure the relative improvement of the CRPS over $\mathrm{CRPS}_{\mathrm{ref}}$, the CRPS for leave-one-year-out climatology reference forecasts:

$$\mathrm{CRPS}_{\mathrm{SkillScore}} = \frac{\mathrm{CRPS}_{\mathrm{ref}} - \mathrm{CRPS}}{\mathrm{CRPS}_{\mathrm{ref}}} \times 100. \qquad (1.21)$$

The CRPS skill score is positively oriented (whereas CRPS is negatively oriented). A maximum score of 100 is indicative of perfect forecasts. A score of 0 indicates no overall improvement compared to the reference forecast. A negative score indicates poor quality forecasts.

A bootstrap procedure is used to detect if the skill scores of merged forecasts are significantly higher or lower compared to calibration forecasts. Uncertainty in the skill scores is primarily due to a limited number of forecast events. A bootstrap procedure is used to build up a plausible distribution of calibration CRPS skill scores, which is represented by the CDF F_{css}. If the CRPS skill score of the merged forecasts exceeds $F_{\mathrm{css}}^{-1}(0.95)$, it is determined that the skill of the merged forecasts is higher than the skill of the calibration forecasts at the 5% significance level. In other words, bridging significantly improves skill over calibration. Similarly, it is tested whether bridging significantly worsens skill compared to calibration alone.

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

Sharpness is checked by plotting the relative proportions of forecasts in bins of the forecast probability. We note that the test against the observed climatological median is more stringent than a test against a modeled median. Any systematic residual bias will become obvious when testing against the observed median.

In construction of the attributes diagrams the forecast probabilities are binned into bins of width 0.1. The plotting points for the *x* axis are the average forecast probabilities within each bin. The plotting points on the *y* axis are the observed relative frequencies. The size of the dots represents the proportion of forecasts in the bin and, therefore, forecast sharpness. If forecast probabilities are reliable (i.e., consistent with the observed relative frequencies of occurrence) the centers of the dots will align with the 1:1 line.

4. Results

a. Skill—minimum temperature

Prior to verifying the CBaM Tmin forecasts, we establish the skill of mean-corrected raw Tmin forecasts, to provide a further basis for comparison. Spatial and seasonal CRPS skill scores for simple mean-corrected Tmin forecasts are presented in Fig. 2. There is one panel for each season from JFM to DJF. The skill scores represent the skill in the original 33-member forecast ensembles after mean bias has been removed. It is clear that simple mean bias correction is ineffective as a method for calibrating POAMA Tmin forecasts. CRPS skill scores for many regions and seasons are largely negative. We will reflect further on these results in the discussion (section 5).

We now turn to results from CBaM. Spatial and seasonal CRPS skill scores for Tmin-calibration forecasts are presented in Fig. 3. Since calibration applies corrections to bias and ensemble spread, these skill scores represent the recovery of inherent skill in the raw Tmin forecasts for each grid cell. CRPS skill scores for Tmin-calibration forecasts are in the realm of −5 to 30. The range of skill scores is congruent with expectations for seasonal climate forecasts, which are generally of low skill (e.g., Langford and Hendon 2013). There is positive skill across parts of eastern Australia from SON to FMA and across northern Australia from JJA to ASO. Slightly negative skill is evident in many regions, for example, in southeast Australia in JJA. In the absence of real forecasting skill, CBaM returns the forecasts to climatology, resulting in marginally negative skill in cross validation.

Spatial and seasonal CRPS skill scores for Tmin-bridging forecasts are presented in Fig. 4. These skill scores represent the skill that can be obtained by ignoring the direct temperature output and considering solely the large-scale climate patterns. CRPS skill scores for POAMA Tmin forecasts are again in the realm of −5 to 30. However, positive skill is evident over a larger proportion of grid cells compared to calibration. There is positive skill across large swathes of Australia from JAS to DJF. Positive skill is limited for JFM–JJA except for AMJ, when weakly positive skill is widespread across Western Australia.

The results for Tmin-calibration skill scores and Tmin-bridging skill scores (Figs. 3 and 4) suggest that bridging can be very effective for improving the skill of POAMA Tmin forecasts. Merging calibration and bridging forecasts is therefore expected to yield higher overall skill. Spatial and seasonal CRPS skill scores for Tmin-merged forecasts are presented in Fig. 5. Green triangles indicate significant improvement in skill due to bridging whereas red triangles indicate significant worsening of skill due to bridging. The regions of positive skill are consistent with the regions of positive skill for Tmin calibration and Tmin bridging. There are regions and seasons where calibration and bridging are similarly skillful (e.g., in southeastern Australia in SON). In this case merging does not significantly improve skill. There are regions and seasons where calibration yields no skill, yet bridging yields considerable skill. An example is the eastern coast in ASO. Overall, the addition of bridging significantly improves skill over calibration in approximately 17% of grid cells. The addition of bridging worsens skill in approximately 1% of grid cells. The CBaM Tmin-merged forecasts are considerably more skillful than the mean-corrected raw Tmin forecasts (cf. Figs. 5 and 2).

b. Skill—maximum temperature

Spatial and seasonal CRPS skill scores for simple mean-corrected Tmax forecasts are presented in Fig. 6, reaffirming the result for Tmin that simple mean bias correction is ineffective for calibrating POAMA temperature forecasts.

Spatial and seasonal CRPS skill scores for Tmax-calibration forecasts are presented in Fig. 7. The Tmax-calibration skill scores are evidently higher than the Tmin-calibration skill scores in a high proportion of grid cells. There is positive skill across eastern Australia in all seasons. There is positive skill across northern Australia, most pervasively from FMA to JAS, but also evident in other seasons. Positive skill for Western Australia is most pervasive FMA–MJJ but is also evident in other seasons, notably OND–DJF. There is no skill for few regions and seasons, for example, in central Australia in DJF.

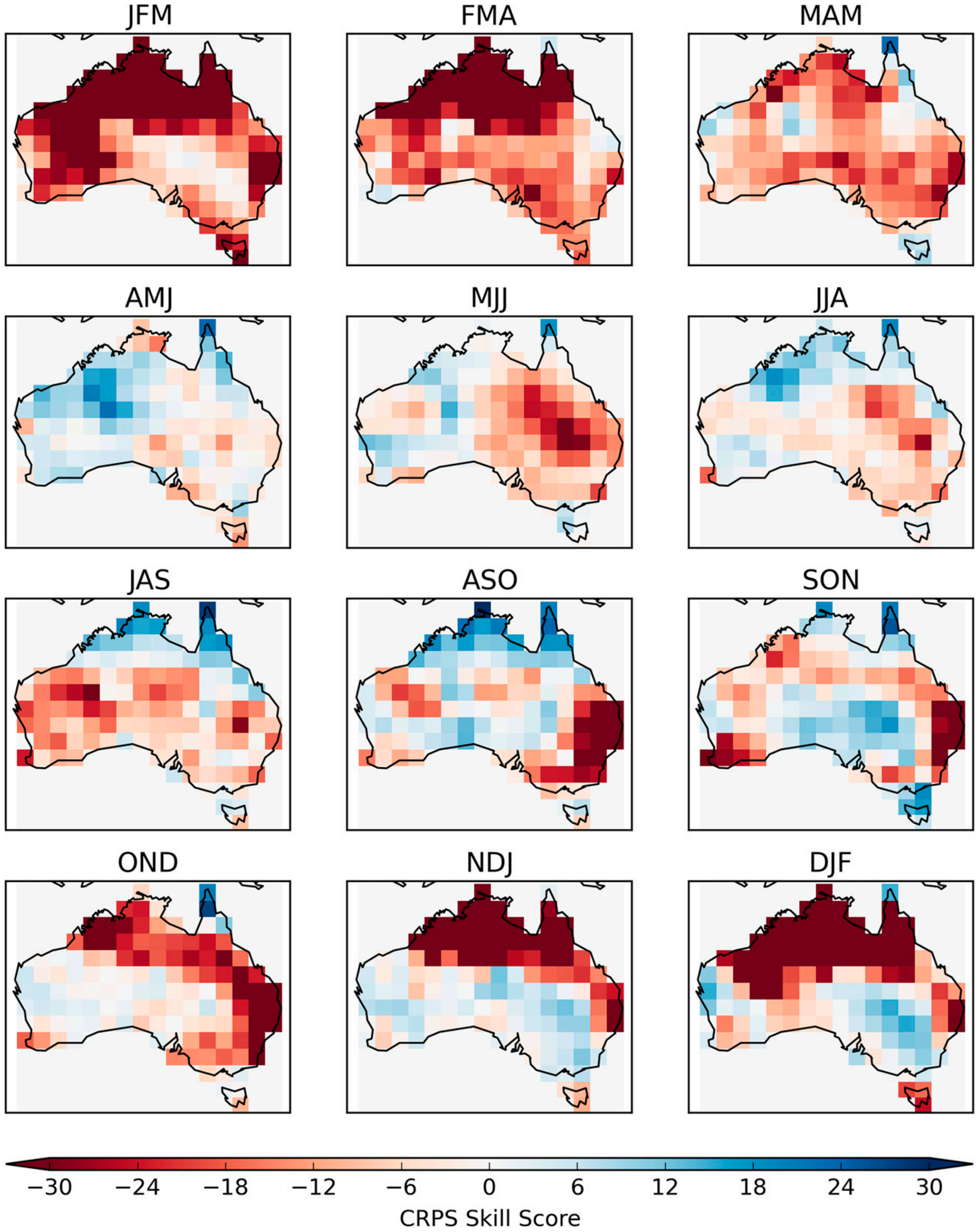

FIG. 2. Cross-validation CRPS skill scores for raw mean-corrected Tmin forecasts. The CRPS skill scores are calculated from hindcasts for the period 1982–2010.

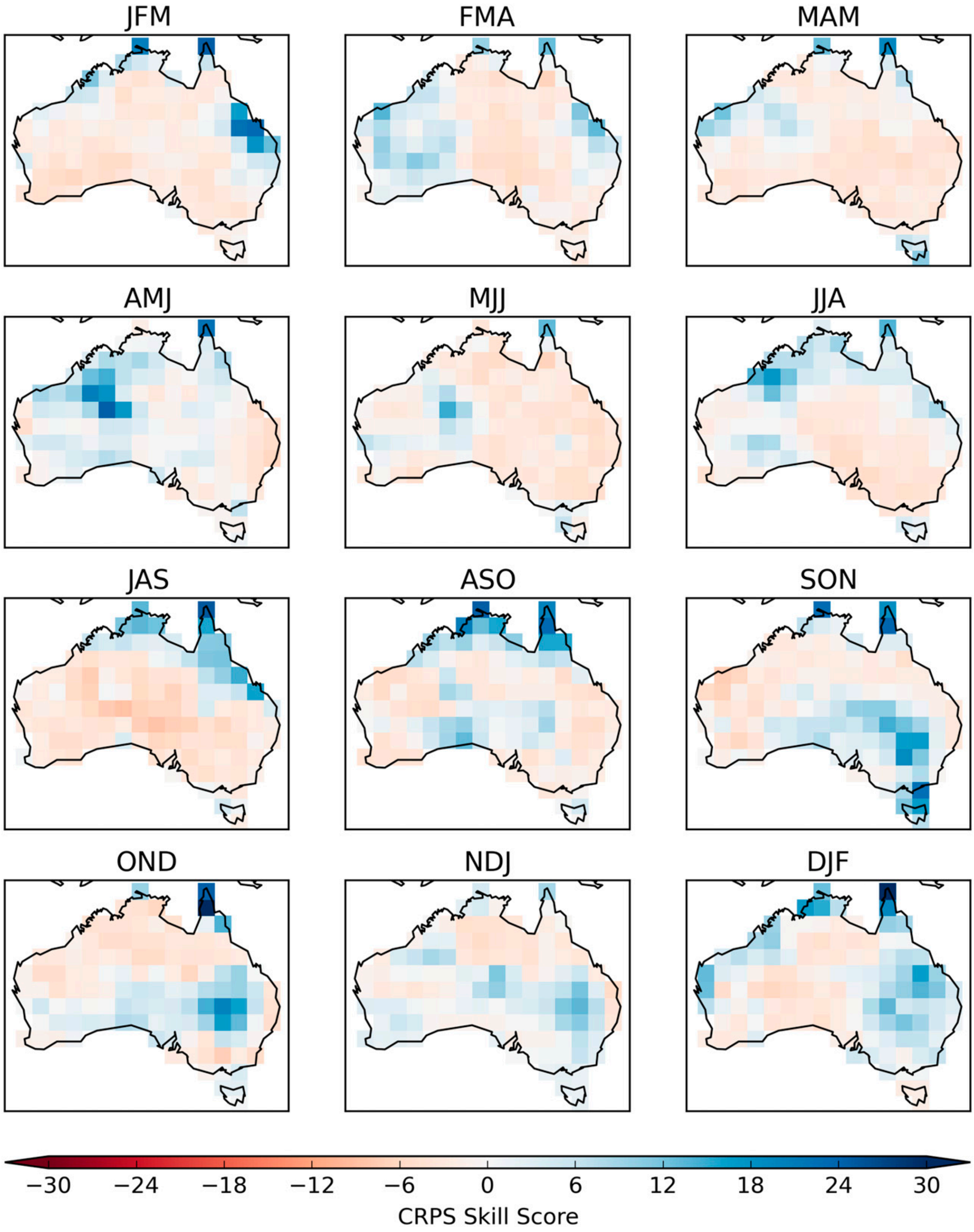

FIG. 3. Cross-validation CRPS skill scores for Tmin-calibration forecasts. The CRPS skill scores are calculated from hindcasts for the period 1982–2010.

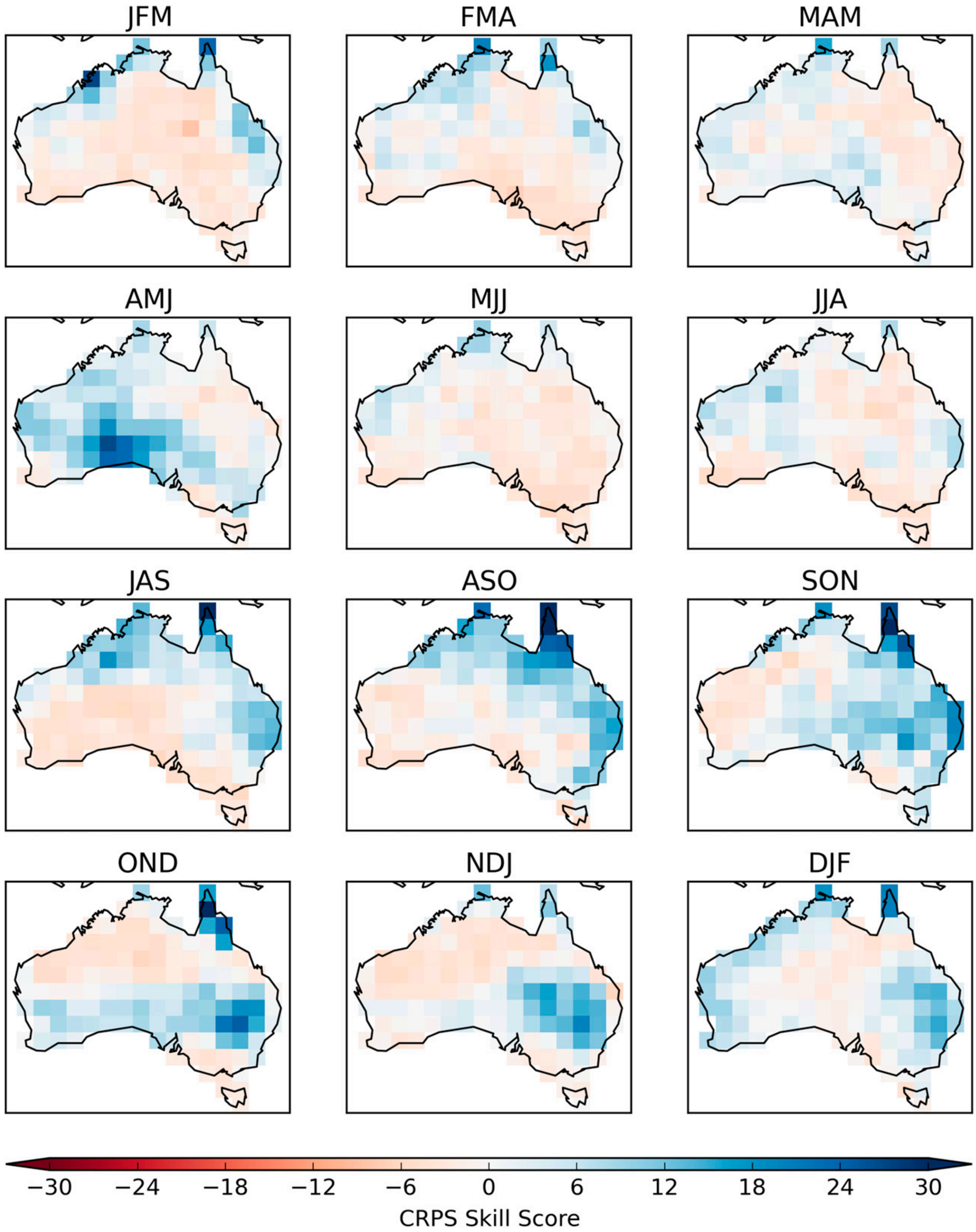

FIG. 4. Cross-validation CRPS skill scores for Tmin-bridging forecasts. The CRPS skill scores are calculated from hindcasts for the period 1982–2010.

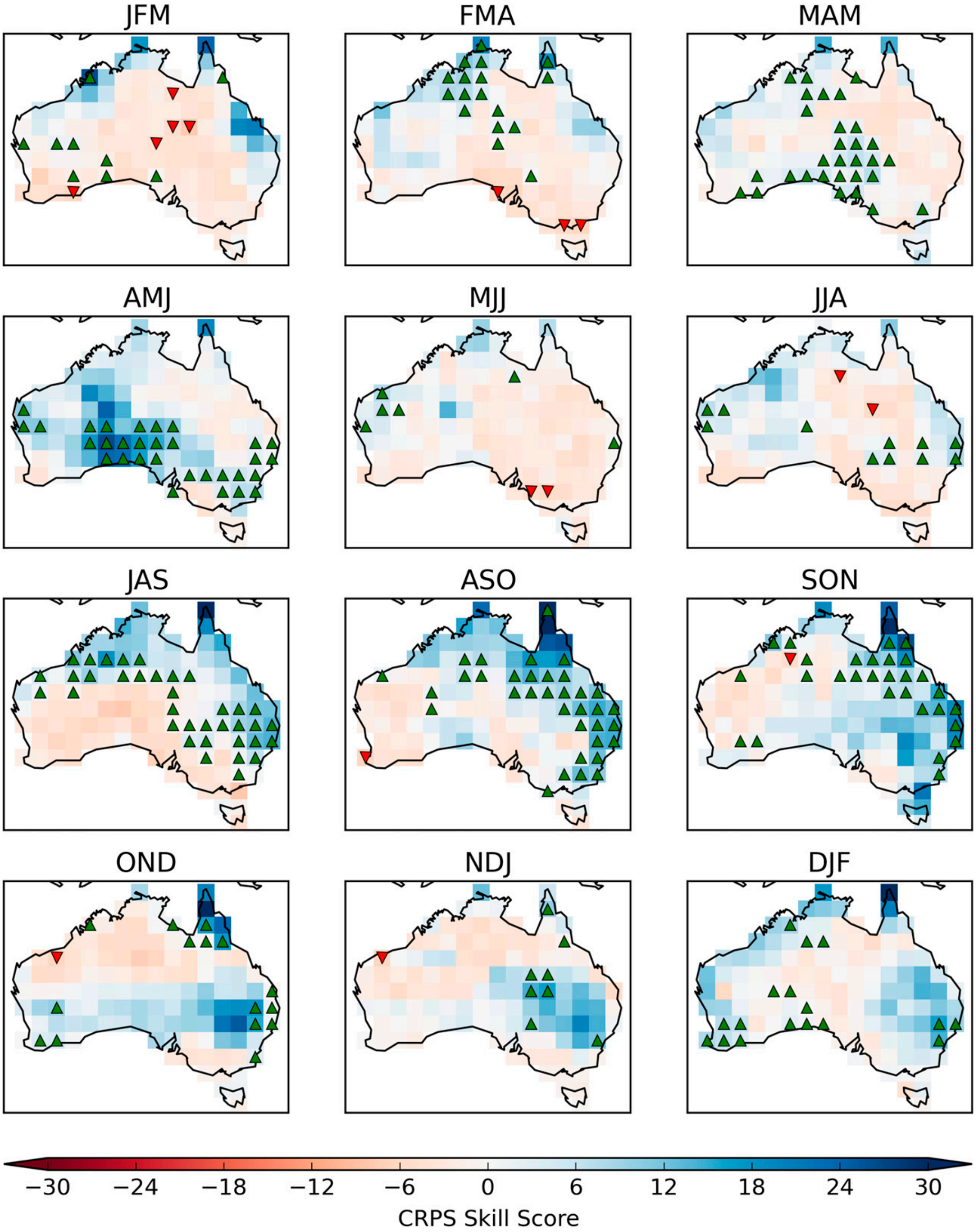

FIG. 5. Cross-validation CRPS skill scores for Tmin-merged forecasts. The CRPS skill scores are calculated from hindcasts for the period 1982–2010. Significance markers at the 5% level are as follows: bridging significantly improves skill (green triangles) and bridging significantly worsens skill (red triangles).

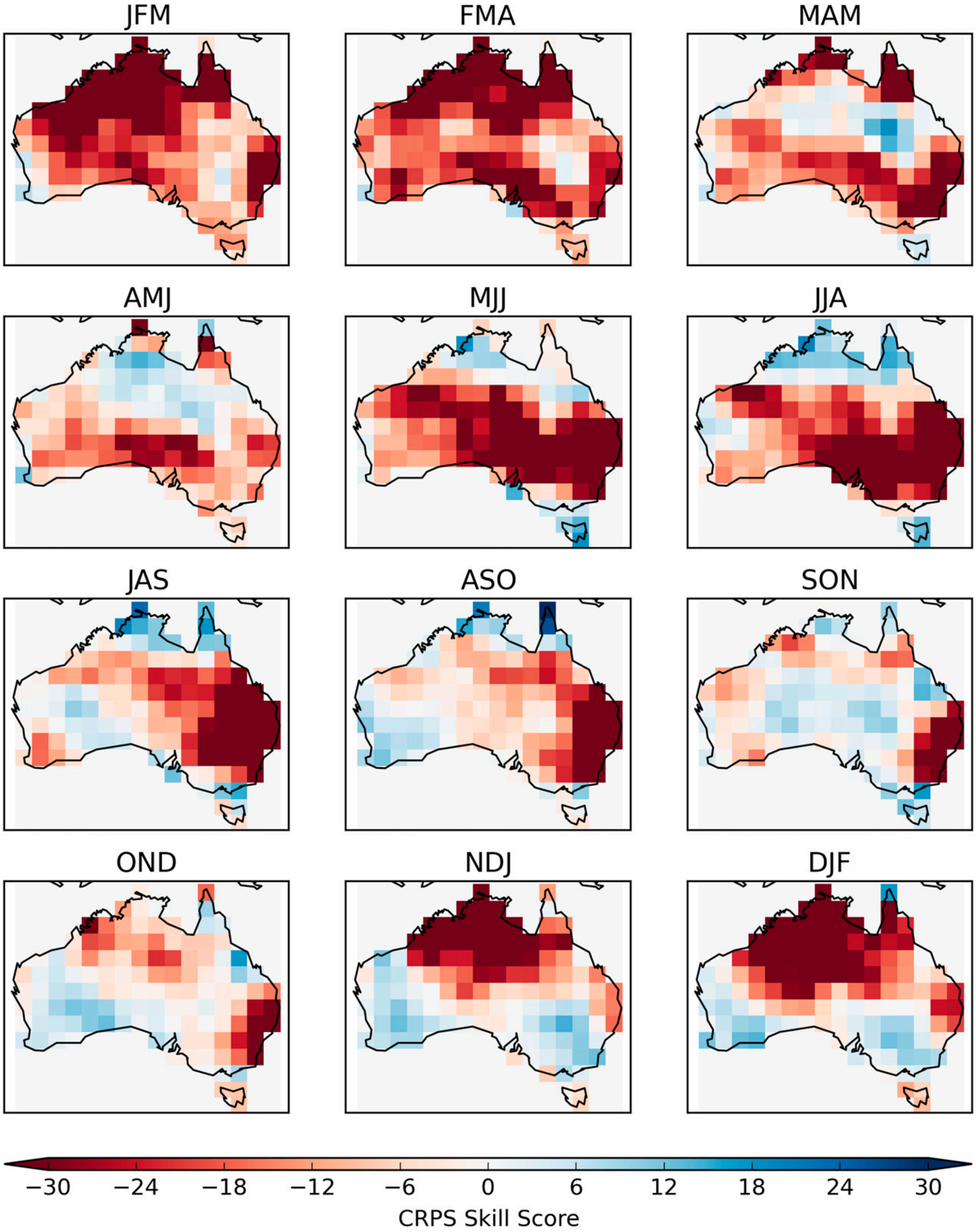

FIG. 6. Cross-validation CRPS skill scores mean-corrected raw Tmax forecasts. The CRPS skill scores are calculated from hindcasts for the period 1982–2010.

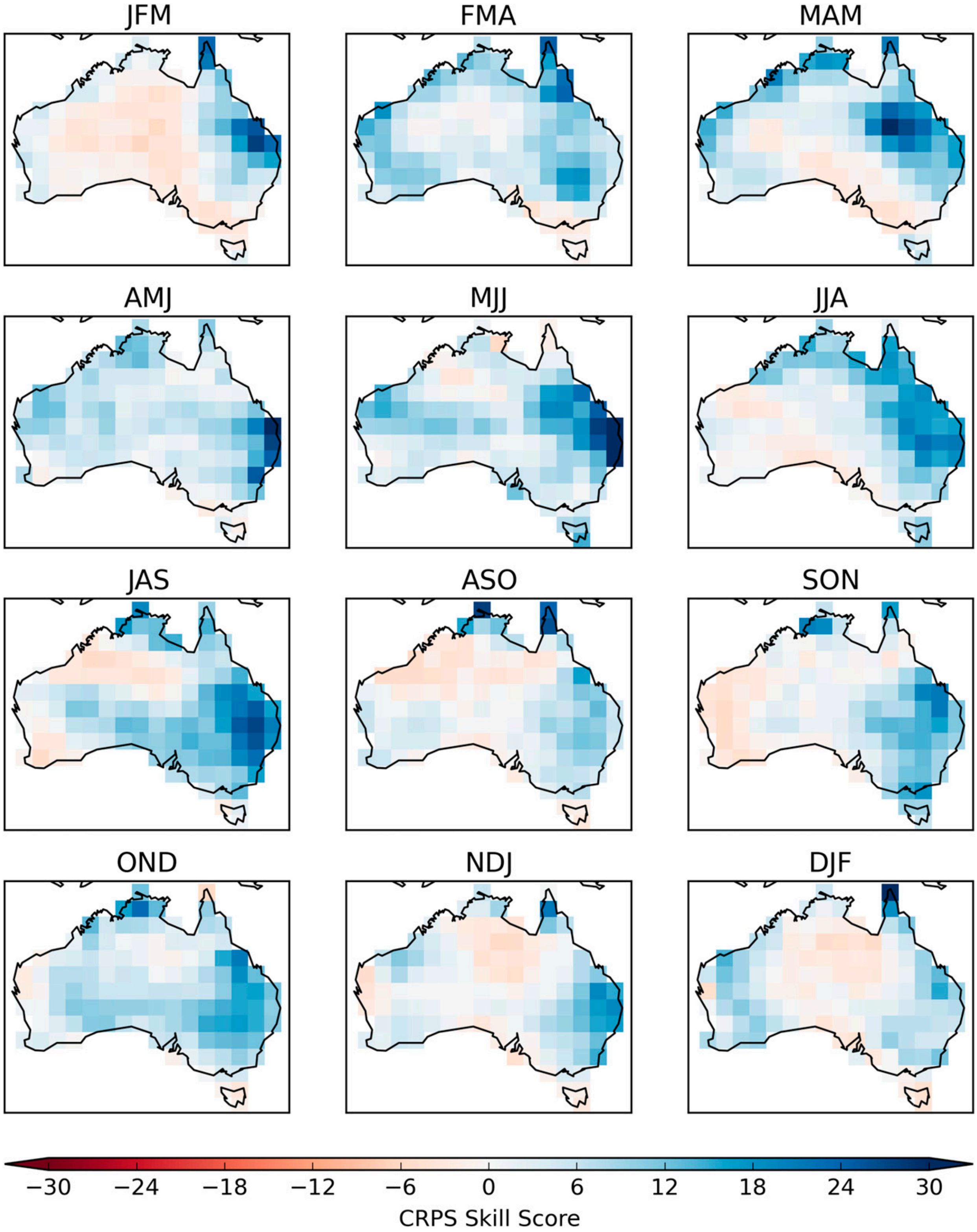

FIG. 7. Cross-validation CRPS skill scores Tmax-calibration forecasts. The CRPS skill scores are calculated from hindcasts for the period 1982–2010.

Spatial and seasonal CRPS skill scores for Tmax-bridging forecasts are presented in Fig. 8. While skill is obtainable by solely considering large-scale climate patterns, Tmax-calibration skill scores exceed Tmax-bridging skill scores for most grid cells. However, Tmax-bridging skill scores exceed Tmax-calibration skill scores chiefly in OND and NDJ. Spatial and seasonal CRPS skill scores for Tmax-merged forecasts are presented in Fig. 9. The green and red triangles indicate significant improvement or worsening in Tmax-merged skill scores due to bridging. The regions of positive skill are consistent with the regions of positive skill for Tmax calibration and Tmax bridging. The addition of bridging is confirmed to significantly improve skill in OND and NDJ. Overall bridging significantly improves skill over calibration in approximately 6% of grid cells. Bridging worsens skill in less than 1% of grid cells. The CRPS skill score results are summarized in Fig. 10, which shows the proportion of grid cells where the merged forecast skill scores are significantly improved or worsened by the inclusion of bridging.

c. *Reliability—minimum and maximum temperature*

The overall reliability and sharpness attributes of the merged forecasts are visually assessed using an attributes diagram. The attributes diagram for Tmin merged is Fig. 11 and the attributes diagram for mean-corrected raw Tmin is Fig. 12. The attributes diagram for Tmax merged is Fig. 13 and the attributes diagram for mean-corrected raw Tmax is Fig. 14. In construction of each diagram, all forecasts for all grid cells and seasons are pooled.

For both Tmin merged and Tmax merged (Figs. 11 and 13), the dots align well with the 1:1 line, indicating reliable forecasts. The dots for more emphatic probabilities deviate somewhat from the 1:1 line; however, due to the generally low skill of the forecasts, there are fewer forecasts and therefore the values are subject to greater uncertainty. Focusing on forecast sharpness, we observe forecasted probabilities are clustered toward the climatological probability and there are few events with emphatic probabilities. The higher skill of Tmax forecasts is reflected in the sharpness attribute with more forecasts deviating from climatological probability than for Tmin. The results for mean-corrected raw Tmin and Tmax (Figs. 12 and 14) are in stark contrast to the CBaM-postprocessed forecasts. The raw mean-corrected forecasts are sharper, as evidenced by higher frequency of more emphatic forecast probabilities; however, the forecast probabilities are not reconciled with observed relative frequencies. The evident reliability of CBaM Tmin and Tmax forecasts confirm that CBaM is far more effective at postprocessing temperature forecasts than simple mean correction.

5. Discussion

a. *Discussion of results*

The results show that CBaM is an effective method for maximizing the skill of POAMA seasonal Tmin and Tmax forecasts for Australia. Skillful Tmin forecasts are primarily achieved through bridging. Skillful Tmax forecasts are primarily achieved through calibration. These results build upon the previous findings of Schepen et al. (2014) who found that CBaM is effective for maximizing the skill of POAMA seasonal rainfall forecasts for Australia.

The failure of the simple mean bias-correction approach to yield acceptable skill scores can be understood in the context of the recent discussion by Barnston et al. (2015). In their study, the authors considered the ensemble spread of Niño-3.4 forecasts from several models composing the North American Multimodel Ensemble (NMME). Although the context is different, we can take insight from their erudition. Barnston et al. (2015) essentially found that GCM forecast ensemble spreads do not vary appropriately from forecast to forecast and therefore contain little useful information about the real uncertainty surrounding a forecast. While the ensemble mean is useful, and mean bias correction helps, other corrections, such as amplitude corrections are necessary for effective postprocessing. Their findings support the need for more sophisticated methods like CBaM.

Bayesian model averaging was applied to merge the calibration and bridging forecasts. An alternative method is quantile model averaging (QMA; Schepen and Wang 2015). QMA averages forecast values across equal quantile fractions (cumulative probabilities). QMA has the effect of preserving the original unimodal shape of the calibration and bridging forecasts and averages their ensemble spread; taking into account the weights. The CBaM method, which first produces multiple ensemble forecasts and then merges with BMA, as detailed in section 3, can theoretically lead to forecasts with a wider than optimal ensemble spread. We tested an alternative approach of applying QMA merging with the same weights as derived through BMA (results not shown). For the Tmax forecasts, which primarily source skill through calibration, QMA merging improved continentally averaged skill scores marginally. For Tmin forecasts, which primarily source skill through bridging, QMA merging did not alter the skill scores noticeably. This suggests that the ensemble spread of bridging forecasts can be marginally wider than optimal due to

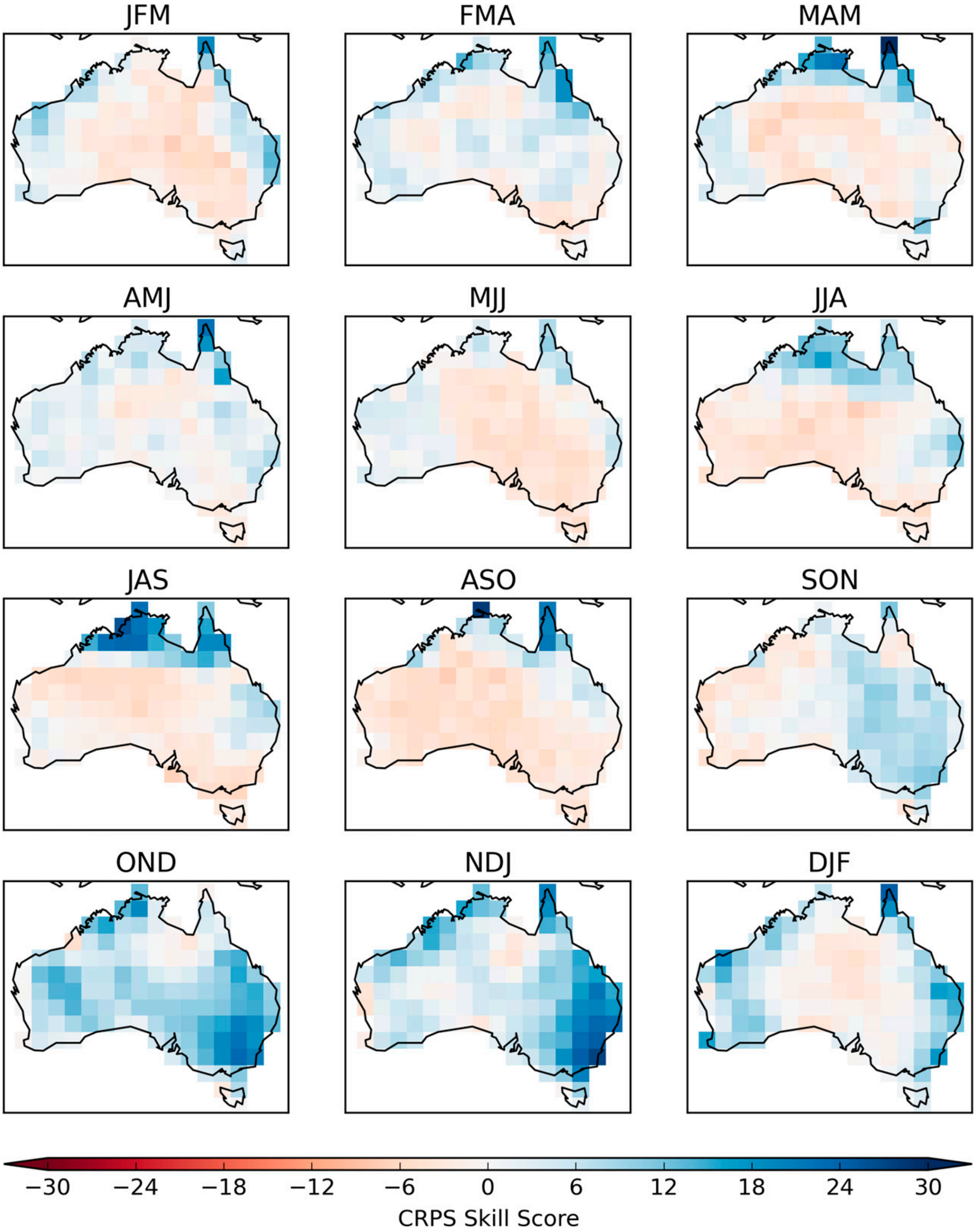

FIG. 8. Cross-validation CRPS skill scores Tmax-bridging forecasts. The CRPS skill scores are calculated from hindcasts for the period 1982–2010.

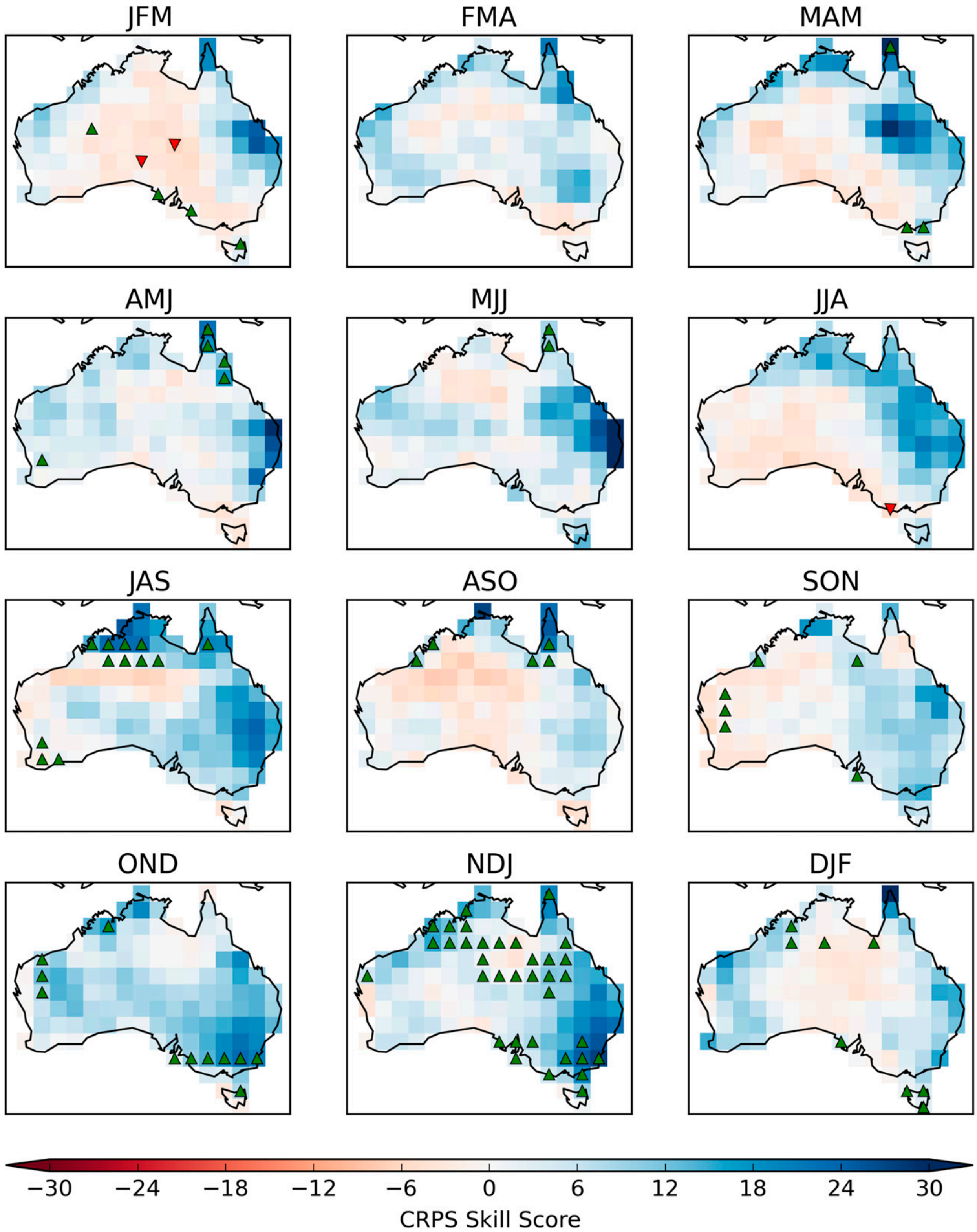

FIG. 9. Cross-validation CRPS skill scores for Tmax-merged forecasts. The CRPS skill scores are calculated from hindcasts for the period 1982–2010: bridging significantly improves skill (green triangles) and bridging significantly worsens skill (red triangles).

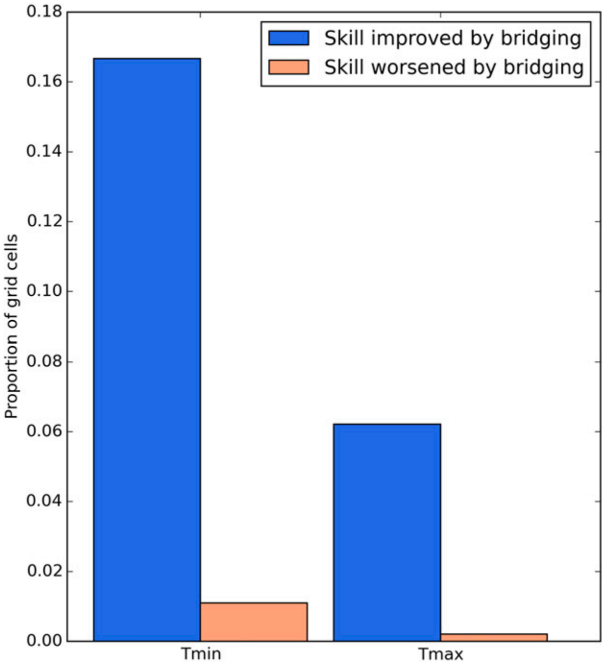

FIG. 10. Proportion of grid cells where improvement or worsening of merged forecast skill scores is attributed to bridging. Improvement or deterioration is tested using a bootstrap method.

the merging of many models. Future research should therefore evaluate QMA more thoroughly as an alternative to BMA in various applications and also investigate alternative methods to further optimize forecast merging.

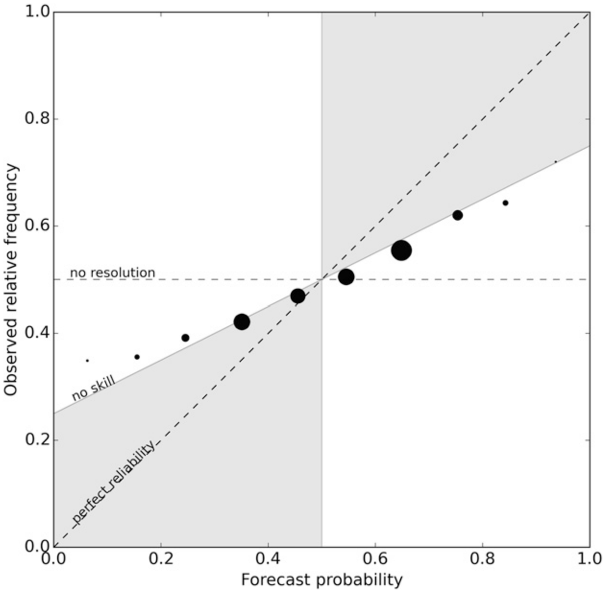

FIG. 12. Attributes diagram for mean-corrected raw Tmin forecasts. All grid cells are considered for the period 1982–2010.

The methodology applied in this study differs from previous studies that applied CBaM. First, parameter uncertainty is not considered for the transformation parameters. Second, the data are subject to scaling. And third, a prior is placed over the transformation parameters. These three measures contributed to a smooth application of CBaM. For example, during the course of the study, it was discovered that forecast values could

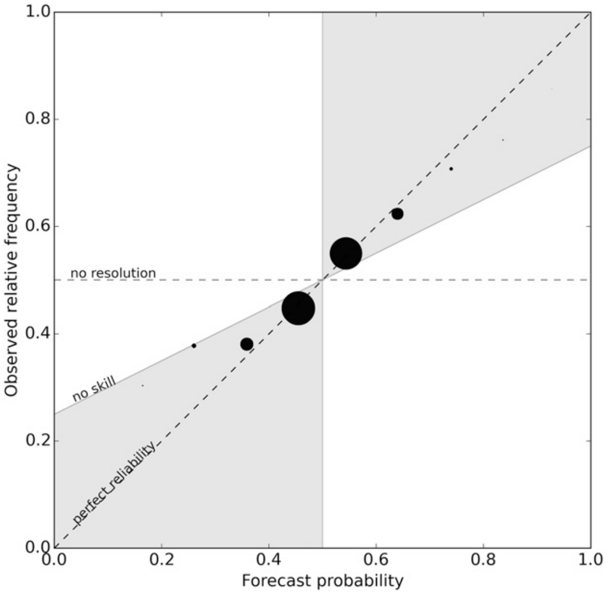

FIG. 11. Attributes diagram for Tmin-merged forecasts. All grid cells are considered for the period 1982–2010.

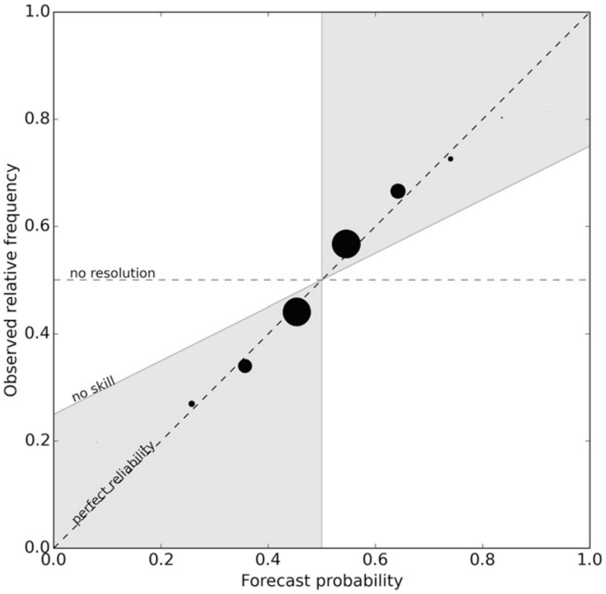

FIG. 13. Attributes diagram for Tmax-merged forecasts. All grid cells are considered for the period 1982–2010.

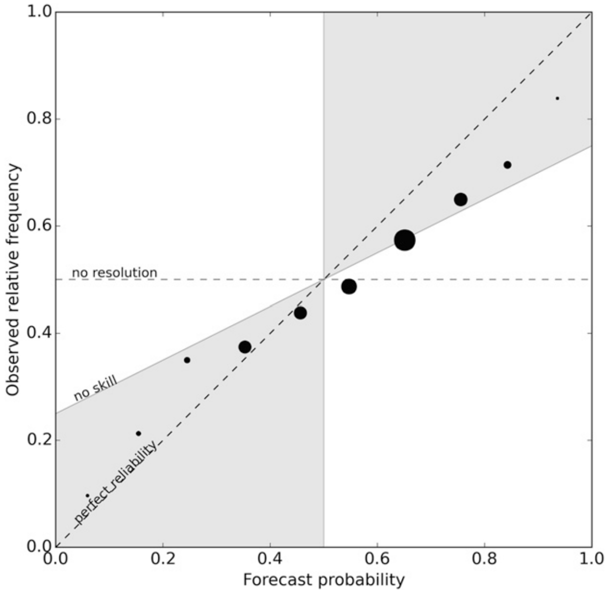

FIG. 14. Attributes diagram for mean-corrected raw Tmax forecasts. All grid cells are considered for the period 1982–2010.

occasionally be sampled beyond reasonable physical limits. The changes to the methodology resolved the problem. As the forecasts proved to be reliable in ensemble spread, it is recommended that future applications of CBaM consider fixed transformation parameters. The continued fortification of CBaM to work consistently well in a wide range of applications is an active area of research.

b. *Future work*

There are a multitude of GCMs in existence, many that are elemental in multimodel ensembles such the NMME and the European Seasonal-to-Interannual Prediction project (EURO-SIP) projects. We expect that CBaM could be uniquely effective in postprocessing temperature, rainfall, and other climate forecasts from different GCMs and in different global regions. Testing of CBaM using other GCMs and for different global regions is therefore a useful avenue for future research. CBaM can also be used to identify priority areas for improvement in GCMs by analyzing the relative contribution of calibration and bridging models to forecast skill. For example, if bridging with a certain predictor is found to improve skill, effort could be invested into improving the teleconnection in the GCM simulations.

A key motivation for applying CBaM to Tmin and Tmax, as given in the introduction, is to contribute to the goal of readily available, consistently postprocessed precipitation and temperature forecasts for use applications such as crop modeling and yield forecasting. Further research is required to realize this goal. First of all, the CBaM Tmin and Tmax forecasts produced in this study represent seasonal (3 month) averages, with 1-month lead time. Crop modeling and forecasting requires monthly data values as a minimum and, more often, daily values, at the crop scale, for many months ahead. CBaM has been proven to be effective at producing monthly time series forecasts of precipitation at the catchment scale for up to 12 months ahead (Schepen and Wang 2014). Properly sequenced ensemble time series forecasts are created by connecting forecast ensemble members using the Schaake shuffle (Clark et al. 2004). It can therefore be assumed that CBaM is readily applicable to produce monthly forecasts of Tmin and Tmax at the crop scale and for many months ahead. The production of daily postprocessed forecasts from CBaM is another challenge altogether. Direct postprocessing of seasonal climate forecasts at the daily time scale is likely to be fraught for a number of reasons, including the lack of clear climate signals and extreme computational requirements. It is therefore our suggestion that CBaM is used to postprocess forecasts at the monthly or seasonal time scale only. Daily forecast values can be obtained through temporal disaggregation. For rainfall, disaggregation may require a sophisticated method that accounts for no-rain days and high variability. For example, disaggregation based on a stochastic weather generator may be a good choice for rainfall (e.g., Hansen and Ines 2005). For temperature, a simpler disaggregation scheme may be applied. Disaggregation of CBaM postprocessed forecasts is earmarked as an area of future research.

The CBaM Tmin and Tmax forecasts are consistent in that the predictor and predictand data sources are the same and the methodology applied is identical. However, Tmin and Tmax are postprocessed separately. The CBaM method can theoretically handle multiple predictors and predictands. It is therefore possible to calibrate Tmin and Tmax jointly. The poor skill of Tmin-calibration forecasts (Fig. 3) may be improved through a joint calibration with Tmax. Furthermore, joint calibration may yield additional skills that are not captured by bridging. On the other hand, the ensemble spread will differ in joint calibration and so a full assessment of forecast skill and reliability would be required. It is therefore an avenue for future research to investigate the value of including joint calibration models in CBaM.

6. Conclusions

Sophisticated postprocessing methods that target the known deficiencies in GCM seasonal climate forecasts need to become prominent to enable climate-sensitive industries to transition to GCM-based forecasts for modeling and decision-making. Very basic methods such as mean bias correction will simply not enable the transition, as the forecasts remain of poor quality.

The calibration, bridging and merging (CBaM) has previously been shown to be effective for postprocessing GCM seasonal rainfall forecasts in Australia and China. In this study, we test CBaM for postprocessing POAMA seasonal temperature forecasts for Australia. Several modifications to the CBaM methodology, including data scaling and prior specification, are found to ease the application to temperature.

CBaM is effective for postprocessing forecasts of POAMA seasonal minimum (Tmin) and maximum (Tmax) temperatures in Australia, considering forecasts issued one month in advance. Tmax skill is primarily maximized by calibration whereas Tmin skill is primarily maximized through bridging. It follows that merging calibration and bridging forecasts yields the best overall skill for postprocessed Tmin and Tmax forecasts.

In addition to maximizing skill, CBaM produces postprocessed forecasts that are bias corrected and reliable in ensemble spread. Further research is anticipated to assist in the adoption of GCM and CBaM forecasts (e.g., through temporal forecast disaggregation).

Acknowledgments. We thank the Australian Bureau of Meteorology and Centre for Australian Weather and Climate Research for making freely available the POAMA and AWAP data sources used in this study. We thank David Robertson for inspired conversations on data transformations and assistance with the Bayesian Joint Probability modeling software.

REFERENCES

Barnston, A. G., and M. K. Tippett, 2013: Predictions of Niño-3. 4 SST in CFSv1 and CFSv2: A diagnostic comparison. *Climate Dyn.*, **41**, 1615–1633, doi:10.1007/s00382-013-1845-2.

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

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

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

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

Coelho, C., S. Pezzulli, M. Balmaseda, F. Doblas-Reyes, and D. Stephenson, 2004: Forecast calibration and combination: A simple Bayesian approach for ENSO. *J. Climate*, **17**, 1504–1516, doi:10.1175/1520-0442(2004)017<1504:FCACAS>2.0.CO;2.

DeChant, C. M., and H. Moradkhani, 2014: Toward a reliable prediction of seasonal forecast uncertainty: Addressing model and initial condition uncertainty with ensemble data assimilation and sequential Bayesian combination. *J. Hydrol.*, **519D**, 2967–2977, doi:10.1016/j.jhydrol.2014.05.045.

Duan, Q., S. Sorooshian, and V. Gupta, 1992: Effective and efficient global optimization for conceptual rainfall-runoff models. *Water Resour. Res.*, **28**, 1015–1031, doi:10.1029/91WR02985.

Dutton, J. A., R. P. James, and J. D. Ross, 2013: Calibration and combination of dynamical seasonal forecasts to enhance the value of predicted probabilities for managing risk. *Climate Dyn.*, **40**, 3089–3105, doi:10.1007/s00382-013-1764-2.

Everingham, Y. L., A. J. Clarke, and S. Van Gorder, 2008: Long lead rainfall forecasts for the Australian sugar industry. *Int. J. Climatol.*, **28**, 111–117, doi:10.1002/joc.1513.

——, N. E. Stoeckl, J. Cusack, and J. A. Osborne, 2012: Quantifying the benefits of a long-lead ENSO prediction model to enhance harvest management—A case study for the Herbert sugarcane growing region, Australia. *Int. J. Climatol.*, **32**, 1069–1076, doi:10.1002/joc.2333.

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

Gneiting, T., A. E. Raftery, A. H. Westveld, and T. Goldman, 2005: Calibrated probabilistic forecasting using ensemble model output statistics and minimum CRPS estimation. *Mon. Wea. Rev.*, **133**, 1098–1118, doi:10.1175/MWR2904.1.

Hansen, J. W., and A. V. Ines, 2005: Stochastic disaggregation of monthly rainfall data for crop simulation studies. *Agric. For. Meteor.*, **131**, 233–246, doi:10.1016/j.agrformet.2005.06.006.

Hawthorne, S., Q. Wang, A. Schepen, and D. Robertson, 2013: Effective use of general circulation model outputs for forecasting monthly rainfalls to long lead times. *Water Resour. Res.*, **49**, 5427–5436, doi:10.1002/wrcr.20453.

Herr, H. D., and R. Krzysztofowicz, 2015: Ensemble Bayesian forecasting system Part I: Theory and algorithms. *J. Hydrol.*, **524**, 789–802, doi:10.1016/j.jhydrol.2014.11.072.

Hoeting, J. A., D. Madigan, A. E. Raftery, and C. T. Volinsky, 1999: Bayesian model averaging: A tutorial (with comments by M. Clyde, David Draper and E. I. George, and a rejoinder by the authors). *Stat. Sci.*, **14**, 382–401, doi:10.1214/ss/1009212519.

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

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

——, A. G. Marshall, Y. Yin, O. Alves, and H. H. Hendon, 2013: Improving intraseasonal prediction with a new ensemble generation strategy. *Mon. Wea. Rev.*, **141**, 4429–4449, doi:10.1175/MWR-D-13-00059.1.

Jo, S., Y. Lim, J. Lee, H.-S. Kang, and H.-S. Oh, 2012: Bayesian regression model for seasonal forecast of precipitation over Korea. *Asia-Pac. J. Atmos. Sci.*, **48**, 205–212, doi:10.1007/s13143-012-0021-7.

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

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

Kirtman, B. P., and Coauthors, 2014: The North American Multimodel Ensemble: Phase-1 Seasonal-to-Interannual Prediction; Phase-2 toward Developing Intraseasonal Prediction. *Bull. Amer. Meteor. Soc.*, **95**, 585–601, doi:10.1175/BAMS-D-12-00050.1.

Langford, S., and H. H. Hendon, 2013: Improving reliability of coupled model forecasts of Australian seasonal rainfall. *Mon. Wea. Rev.*, **141**, 728–741, doi:10.1175/MWR-D-11-00333.1.

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

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

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

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

Peng, Z., Q. Wang, J. C. Bennett, P. Pokhrel, and Z. Wang, 2014: Seasonal precipitation forecasts over China using monthly large-scale oceanic-atmospheric indices. *J. Hydrol.*, **519**, 792–802, doi:10.1016/j.jhydrol.2014.08.012.

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

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

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

——, and ——, 2014: Ensemble forecasts of monthly catchment rainfall out to long lead times by post-processing coupled general circulation model output. *J. Hydrol.*, **519**, 2920–2931, doi:10.1016/j.jhydrol.2014.03.017.

——, and ——, 2015: Model averaging methods to merge operational statistical and dynamic seasonal streamflow forecasts in Australia. *Water Resour. Res.*, **51**, 1797–1812, doi:10.1002/2014WR016163.

——, ——, 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.

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

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

Shinozaki, T., S. Furui, and T. Kawahara, 2010: Gaussian mixture optimization based on efficient cross-validation. *IEEE J. Sel. Topics Signal Process.*, **4**, 540–547, doi:10.1109/JSTSP.2010.2048235.

Troccoli, A., 2010: Seasonal climate forecasting. *Meteor. Appl.*, **17**, 251–268.

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., and D. 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. 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, doi:10.1175/JCLI-D-11-00386.1.

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

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

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

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

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