Geosci. Model Dev., 10, 2321–2332, 2017
https://doi.org/10.5194/gmd-10-2321-2017

Geoscientific Model Development
Open Access
EGU

A Bayesian posterior predictive framework for weighting ensemble regional climate models

Yanan Fan[1], **Roman Olson**[2], **and Jason P. Evans**[3]

[1]School of Mathematics and Statistics, UNSW, Sydney, Australia
[2]Department of Atmospheric Sciences, Yonsei University, Seoul, South Korea
[3]Climate Change Research Centre and ARC Centre of Excellence for Climate System Science, UNSW, Sydney, Australia

Correspondence to: Yanan Fan (y.fan@unsw.edu.au)

Received: 28 November 2016 – Discussion started: 4 January 2017
Revised: 17 May 2017 – Accepted: 22 May 2017 – Published: 23 June 2017

Abstract. We present a novel Bayesian statistical approach to computing model weights in climate change projection ensembles in order to create probabilistic projections. The weight of each climate model is obtained by weighting the current day observed data under the posterior distribution admitted under competing climate models. We use a linear model to describe the model output and observations. The approach accounts for uncertainty in model bias, trend and internal variability, including error in the observations used. Our framework is general, requires very little problem-specific input, and works well with default priors. We carry out cross-validation checks that confirm that the method produces the correct coverage.

1 Introduction

Regional climate models (RCMs) are powerful tools to produce regional climate projections (Giorgi and Bates, 1989; Christensen et al., 2007; van der Linden and Mitchell, 2009; Evans et al., 2013, 2014; Mearns et al., 2013; Solman et al., 2013; Olson et al., 2016b). These models take climate states produced by global climate models (GCMs) as boundary conditions, and solve equations of motion for the atmosphere on a regional grid to produce regional climate projections. The main advantages of RCMs over GCMs are increased resolution, more parsimony in terms of representing sub-grid-scale processes, and often improved modelling of spatial patterns, particularly in regions with coastlines and considerable topographic features (e.g. van der Linden and Mitchell, 2009; Prömmel et al., 2010; Feser et al., 2011). Current computing power is now allowing for ensembles of regional climate models to be performed, allowing for sampling of model structural uncertainty (Christensen et al., 2007; Giorgi and Bates, 1989; van der Linden and Mitchell, 2009; Mearns et al., 2013; Solman et al., 2013).

Along with these ensemble modelling studies, methods for extracting probabilistic projections have followed (Buser et al., 2010; Fischer et al., 2012; Kerkhoff et al., 2015; Olson et al., 2016a; Wang et al., 2016). While these studies all take a Bayesian approach, the implementations differ. For example, Buser et al. (2010) and Kerkhoff et al. (2015) model both the RCM output and the observations as a function of time. However, this implementation uses too many parameters to be applicable to short (e.g. 20-year) time series common in regional climate modelling. Furthermore, the results are affected by climate model convergence: the output from the outlier models is pulled towards clusters of converging models. The Wang et al. (2016) method is applicable to relatively short time series; however, convergence still influences model predictions.

Olson et al. (2016a) introduced Bayesian model averaging to the RCM model processing. In their framework, model clustering does not affect the results, incorporating their belief that clustering can occur due to common model errors. Furthermore, they provide model weights – a useful diagnostic of model performance. The weights depend on model performance in terms of trend, bias, and internal variability. However, their approach still suffers from shortcomings. Specifically, the observations are modelled as a function of smoothed model output. However, the smoothing requires subjective choices, and the uncertainty in the smooth-

ing choice is not explicitly considered. Second, in the projection stage the Olson et al. (2016a) implementation does not fully account for the uncertainty in model biases and in standard deviation of the model–data residuals.

Several authors have shown that in many regions, future changes are positively correlated with present-day internal variability in the models: see Buser et al. (2009) and Huttunen et al. (2017). This means that knowing internal variability may provide important information and potentially improve future projections. While previous works have included information from internal variability in their statistical model, the information was not used to directly penalise the models for getting the internal variability wrong: see for example Buser et al. (2010) and Kerkhoff et al. (2015). Olson et al. (2016a) was the first attempt to incorporate this information via penalising model priors. However, the priors were chosen ad hoc. A fundamental improvement of this work is weighting the models not just by their performance in terms of the mean, but also in terms of the internal variability in a principled way.

In this article, we propose a new method to obtain model weights using raw model output, so the method better accounts for model output uncertainty. Our framework allows us to compute weights efficiently, simultaneously penalising for model bias, deviations in trend and model internal variability. One of the main advantages of the current approach is that improper and vague priors for the model parameters can be used, which makes implementation of the method much more straightforward. In the Olson et al. (2016a) framework, subjective and informative parameter choices are required. Such choices impact strongly on the resulting weights and inference. In addition, their framework cannot accommodate improper priors since they need to be able to sample directly from the prior.

Below the Bayesian methodology developed is described followed by a Markov chain Monte Carlo (MCMC) method to obtain solutions for the posterior distributions. The technique is then applied to a regional climate model ensemble and compared with results found in previous work (Olson et al., 2016a).

2 Posterior predictive weighting

In this section, we introduce the Bayesian methodology for weighting model output based on current day observations. The framework we describe below is not limited to any particular distributional form, although the analysis presented is based on the univariate normal distribution. We have also implemented the same procedure using the asymmetric Laplace distribution for median regression to obtain robust estimators for our analyses, but we have excluded them from presentation as the procedure produced similar results to that of the normal error assumption (indicating no major violations from normality).

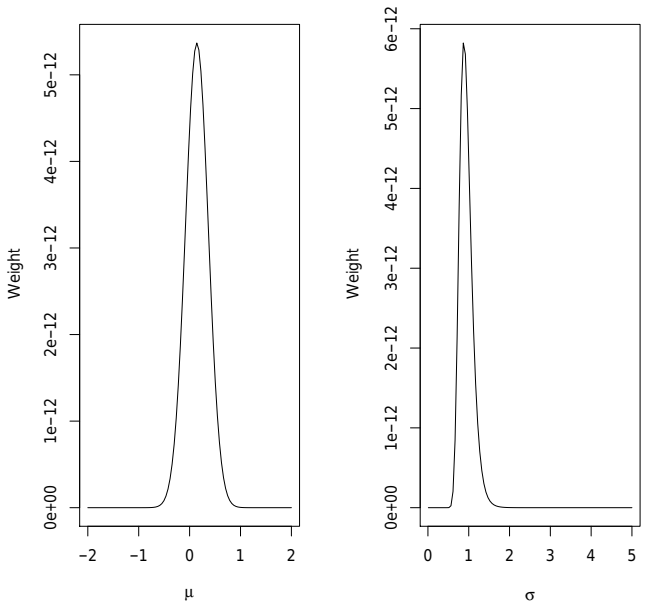

Figure 1. Pictorial representation of the weight distribution on μ and σ.

We suppose that current day observations are denoted as y_t, where $t = 1, \ldots, T$ is a set of indices for time. We assume that the present-day observations over time can be described by

$$y_t = a_p + b_p(t - t_1) + \epsilon_t \tag{1}$$

where $\epsilon_t \sim N(0, \sigma_p)$, $t = t_0, \ldots t_0 + T$, and t_0 is the first year that the observation is available, and $t_1 = t_0 + T/2$. Formulating the equation in terms of t_1 allows us to interpret a_p as the mean value of the observations. This model is reasonable for the type of short time series temperature data that we consider. We assume that the data y_t are independent between observations. Let $x_t^m, t = 1, \ldots, T$ denote data generated by the mth model over the same time period, where $m = 1, \ldots, M$, and we assume that each set of model outputs can be adequately modelled by

$$x_t^m = a_m + b_m(t - t_1) + \epsilon_t \tag{2}$$

with $\epsilon_i \sim N(0, \sigma_m)$. Again, x_ts are assumed independent.

The parameters a_m, b_m, σ_m can be obtained under the Bayesian paradigm by first specifying a prior distribution $p(a_m, b_m, \sigma_m)$, and the posterior distribution given data x^m is subsequently obtained via the Bayes rule,

$$p(a_m, b_m, \sigma_m | x^m) \propto L(x^m | a_m, b_m, \sigma_m) p(a_m, b_m, \sigma_m), \tag{3}$$

where $L(x^m|\cdot)$ denotes the likelihood of obtaining data x^m from model m. In this work, vague priors are used throughout. The use of a vague prior allows the data to discriminate amongst models, whereas informative priors reflect the scientist's personal knowledge, and can lead to more subjec-

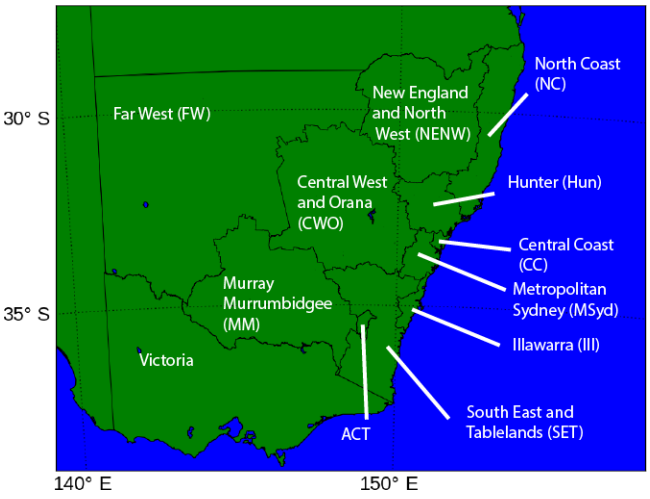

Figure 2. New South Wales planning regions, the ACT and the state of Victoria.

tive analyses. Vague priors are sometimes considered preferable when data contain sufficient information or when subjective knowledge is uncertain. Conjugate analyses for certain classes of models, including Gaussian error models, are often possible, leading to analytical forms for the posterior distributions. In this work, we choose to present the results with non-standard priors, and use MCMC for computation. This approach is much easier when extending to more complex modelling scenarios.

We would like to weight the models based on the similarity of output x_t^m to the observation data. We note that a model that performs well under recent conditions does not guarantee that it will perform well under future climate conditions, but we assume that good performance under recent conditions is an indication of reliable performance in future climates. This translates to preferring models whose parameters a_m, b_m, σ_m are similar to a_p, b_p, σ_p. In practice σ_p has additional terms, due to instrumental and gridding error associated with collecting observational data. This additional error is not reflected in the model output. Jones et al. (2009) performed error analyses for 2001–2007 for Australian climate data, and found that the root mean squared error for monthly temperature data ranges between 0.5 and 1 K. For our analyses of seasonally averaged temperature data in Sect. 2.2, we set the additional error to be $\delta = 0.5$ K. Resulting weights were largely insensitive to values of δ between 0.5 and 1.

Finally, we define the weight for each model m to be of the form

$$w^m = \int L(y|a_m, b_m, \sqrt{\sigma_m^2 + \delta^2})\, p(a_m, b_m, \sigma_m|x^m)\mathrm{d}a_m\mathrm{d}b_m\mathrm{d}\sigma_m \quad (4)$$

where $L(y|a_m, b_m, \sqrt{\sigma_m^2 + \delta^2})$ denotes the likelihood of observational data y, given the parameters of the mth model, a_m, b_m and σ_m. The weight w^m fully accounts for the uncertainties associated with the estimates of a_m, b_m and σ_m, by averaging over the posterior distribution of $p(a_m, b_m, \sigma_m|x^m)$. Clearly, the right-hand side of Eq. (4) will be larger if a_m, b_m and $\sqrt{\sigma_m^2 + \delta^2}$ are similar to a_p, b_p and σ_p, i.e. if the distributions of y and x^m are similar (up to a difference of observational error δ). We term these weights the posterior predictive weights. Note that Eq. (4) is simply the marginal likelihood $p(y|x^m)$, i.e. the probability of observing data y given x_m, averaging over any model parameter uncertainties. The term a_m and its deviation from a_p in the observation model can be considered as penalising bias between model output and observation, the deviation between b_m and b_p can be thought of as a penalty for trend, and the terms σ_m and σ_p account for the differences of model and observation internal variability.

The ensemble models can now be combined into a single posterior model, using the weights

$$p(a_{\mathrm{BMA}}, b_{\mathrm{BMA}}, \sigma_{\mathrm{BMA}}|x^1, \ldots, x^M) = \sum_{m=1}^{M} w^m\, p(a_m, b_m, \sigma_m|x^m). \quad (5)$$

The above expression gives us an ensemble estimate for the posterior distribution of the parameters for a, b and σ from the M model outputs, and we denote these as a_{BMA}, b_{BMA} and σ_{BMA}. Note that the weights should be normalised by $\sum_{m=1}^{M} w^m = 1$.

In order to understand this weight, we suppose for the moment that the data y come from, say, a $N(0, 1)$. Suppose also that x^m comes from $N(\mu, \sigma)$. Then if the posterior distributions of μ and σ are centered around 0 and 1, x^m should be assigned a higher weight. As the values of μ and σ diverge away from 0 and 1, we should see a decrease in the respective weights. Figure 1 plots the likelihood of 50 simulated y values from $N(0, 1)$ distribution, the left panel shows the weights for a fixed value of $\mu = -2, \ldots, 2$ and $\sigma = 1$, and the right panel shows the weights for a fixed value of $\sigma = 0.01, \ldots, 5$ with $\mu = 0$. The figure corresponds to a single term inside the weight Eq. (4), where $a_{m,\mathrm{I}}, b_{m,\mathrm{I}}$ correspond to μ and $\sqrt{\sigma_{m,\mathrm{I}}^2 + \delta^2}$ corresponds to σ. See also Eq. (6) below. The figure shows the changes in the weight, as parameter values move away from the true values of 0 and 1. In the case of single fixed values of μ and σ, the weights simply correspond to the likelihood at these values. In practice, the weights in Eq. (4) average over the set of posterior values of μ and σ.

It is worth noting that even if we specify non-informative priors in Eq. (3) for all models, the implied priors used in our approach are not uninformative. As pointed out by H. R. Künsch, some form of informative priors must be used because the data available simply do not contain information for certain parameters of the model for the future (see Buser et al., 2009 for an alternative formulation which also requires some form of informative prior specifications.) In the current case, our modelling approach assumes that the relationship

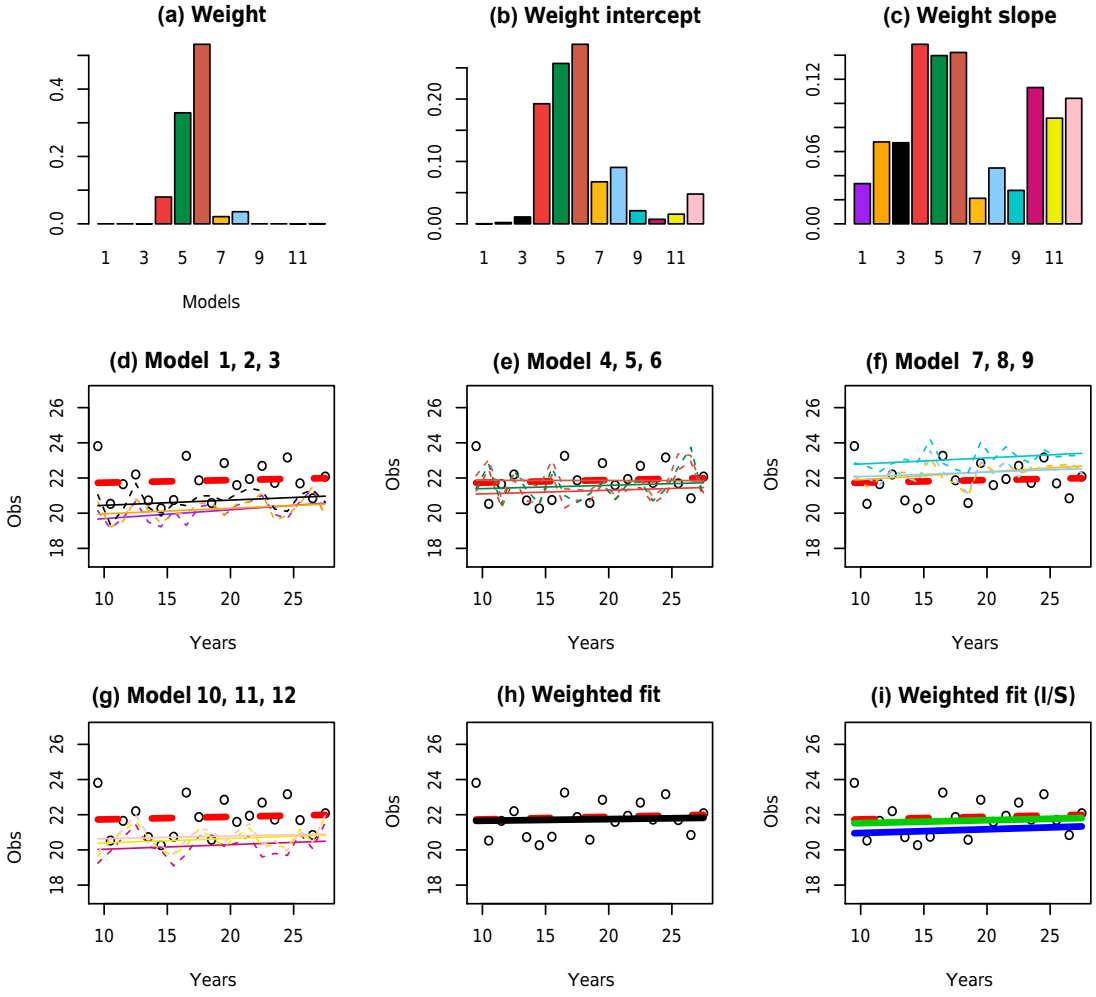

Figure 3. Results for the CC region of south-eastern Australia, in the DJF season. Top row: weights w^m of 12 models based on Eq. (4) (L), Eq. (8), $w^{m,\mathrm{I}}$ (M) and Eq. (9) $w^{m,\mathrm{T}}$ (R). Each triplet represents a GCM (MIROC3.2, ECHAM5, CCCMA3.1, and CSIRO-Mk3.0). Middle row and first plot of last row: fitted observations according to Eq. (1) (red dashed line) and fitted model output according to Eq. (2) for 12 models. Last row: weighted fit based on w^m in solid black line (M), weighted fit based on $w^{m,\mathrm{I}}$ in solid green line and weighted fit based on $w^{m,\mathrm{T}}$ in solid blue line (R).

between future climate and future model output behaves in a similar way to the relationship between present-day climate and present-day model output. We consider that there is a perfect model that has the same parameters (intercept, slope and standard deviation) in both the present and the future. We then compute the probability that any model m is this perfect model, based on present-day data. These assumptions can be seen as an informative prior on the parameters governing future observations, although these parameters are not explicitly modelled.

2.1 Computation

The procedure for the calculation of weights is designed to be applicable regardless of the distributional forms chosen to model the data. In most cases, the posterior distributions $p(a_m, b_m, \sigma_m|x^m)$ in Eq. (3) will be analytically intractable; however, samples from this distribution can easily be obtained via MCMC. Many software packages performing MCMC are available. For the analysis in this paper, we used the MCMCpack library of the R statistical package (R Core Team, 2014). MCMC is an iterative algorithm, and it is necessary to check for convergence and throw away an initial burn-in period of the chain. For our simulations, we used 5000 chain iterations, throwing away the initial 500 iterations as burn in, retaining $N = 4500$ MCMC samples to work with. Default priors from MCMCpack were used throughout this paper. For the model and data used in this paper, only a routine application of MCMC was required. However, more complex model and data typically require advanced knowledge of MCMC; see Gilks et al. (1996) for more on MCMC.

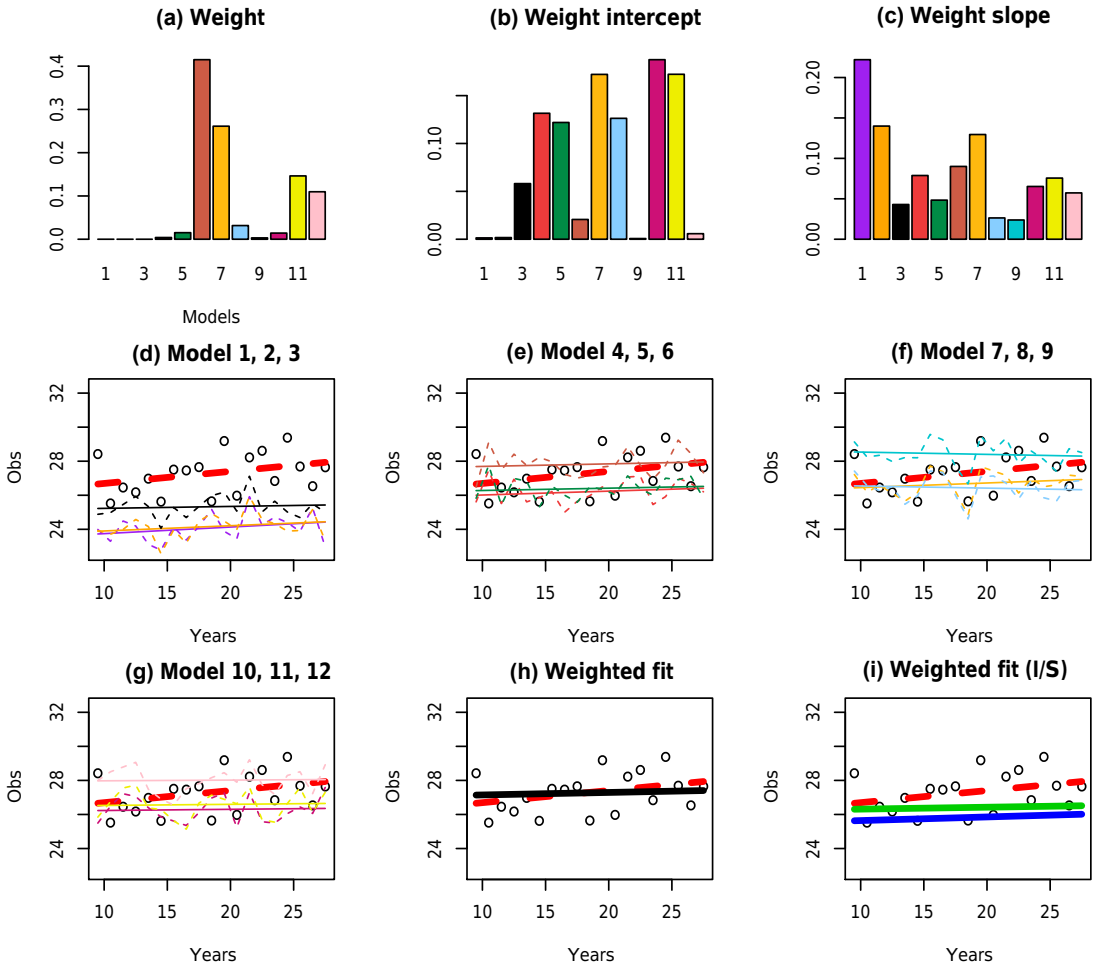

Figure 4. Results for the FW region of south-eastern Australia, in the DJF season. Top row: weights w^m of 12 models based on Eq. (4) (L), Eq. (8), $w^{m,\mathrm{I}}$ (M) and Eq. (9) $w^{m,\mathrm{T}}$ (R). Each triplet represents a GCM (MIROC3.2, ECHAM5, CCCMA3.1, and CSIRO-Mk3.0). Middle row and first plot of last row: fitted observations according to Eq. (1) (red dashed line) and fitted model output according to Eq. (2) for 12 models. Last row: weighted fit based on w^m in solid black line (M), weighted fit based on $w^{m,\mathrm{I}}$ in solid green line and weighted fit based on $w^{m,\mathrm{T}}$ in solid blue line (R).

In addition to obtaining simulations from the posteriors of the M ensemble models, the weight calculation in Eq. (4) also involves an intractable integral, which we can approximate using standard Monte Carlo

$$w^m \approx \sum_{a_{m,\mathrm{I}},b_{m,\mathrm{I}},\sigma_{m,\mathrm{I}}} L(y|a_{m,\mathrm{I}},b_{m,\mathrm{I}},\sqrt{\sigma_{m,\mathrm{I}}^2+\delta^2}) \tag{6}$$

where $L(y|a_{m,\mathrm{I}},b_{m,\mathrm{I}},\sqrt{\sigma_{m,\mathrm{I}}^2+\delta^2})$ denotes the likelihood of y under the ith sample of $a_{m,\mathrm{I}},b_{m,\mathrm{I}}$ and $\sigma_{m,\mathrm{I}}$ from the posterior distribution $p(a_m,b_m,\sigma_m|x^m)$. Thus, the 4500 MCMC samples obtained for each model are then used to compute the Monte Carlo sum in Eq. (6). Again, the weights should be normalised by the constraint $\sum_{m=1}^{M} w^m = 1$.

Finally, the predictive distribution for the future climate $y_t^{\mathrm{f}}, t = 1,\ldots,T'$, given future model output denoted as $x^{\mathrm{f},1},\ldots,x^{\mathrm{f},m}$, is defined as

$$p(y_1^{\mathrm{f}},\ldots,y_{T'}^{\mathrm{f}}|x^{\mathrm{f},1},\ldots,x^{\mathrm{f},M}) = \int p(y_1^{\mathrm{f}},\ldots,y_{T'}^{\mathrm{f}}|a_{\mathrm{BMA}}^{\mathrm{f}},b_{\mathrm{BMA}}^{\mathrm{f}},\sigma_{\mathrm{BMA}}^{\mathrm{f}}) p(a_{\mathrm{BMA}}^{\mathrm{f}},b_{\mathrm{BMA}}^{\mathrm{f}},\sigma_{\mathrm{BMA}}^{\mathrm{f}}|x^{\mathrm{f},1},\ldots,x^{\mathrm{f},M}) \mathrm{d}a_{\mathrm{BMA}}^{\mathrm{f}}\mathrm{d}b_{\mathrm{BMA}}^{\mathrm{f}}d\sigma_{\mathrm{BMA}}^{\mathrm{f}}. \tag{7}$$

2.2 Application

Here we consider the same data as Olson et al. (2016a) – temperature output from NARCliM (New South Wales/ACT Regional Climate Modeling Project, Evans et al., 2014). This project is the most comprehensive regional modelling project for south-eastern Australia, and the first to systematically explore climate model structural uncertainties. The NARCliM ensemble downscales four GCMs (MIROC3.2, ECHAM5,

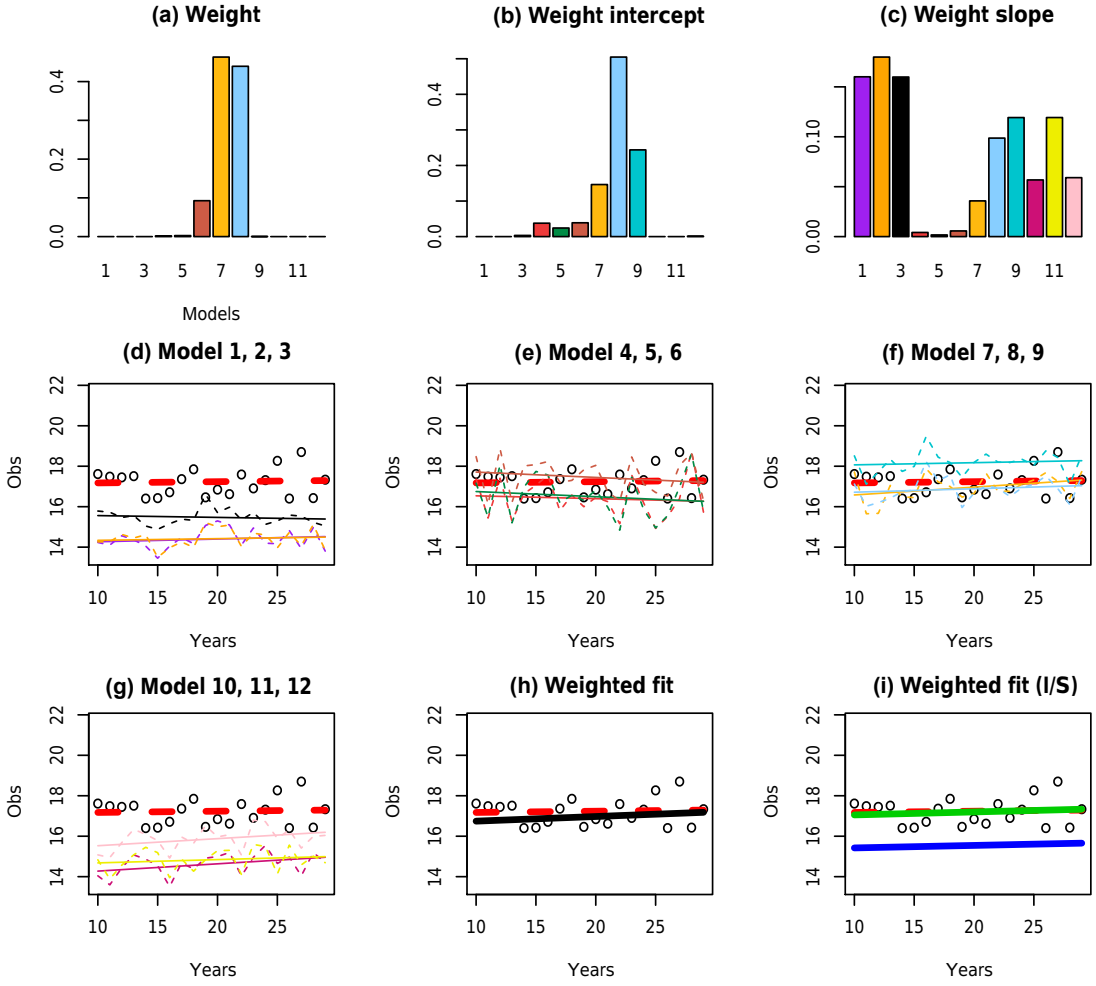

Figure 5. Results for the CWO region of south-eastern Australia, in the MAM season. Top row: weights w^m of 12 models based on Eq. (4) (L), Eq. (8), $w^{m,\mathrm{I}}$ (M) and Eq. (9) $w^{m,\mathrm{T}}$ (R). Each triplet represents a GCM (MIROC3.2, ECHAM5, CCCMA3.1, and CSIRO-Mk3.0). Middle row and first plot of last row: fitted observations according to Eq. (1) (red dashed line) and fitted model output according to Eq. (2) for 12 models. Last row: weighted fit based on w^m in solid black line (M), weighted fit based on $w^{m,\mathrm{I}}$ in solid green line and weighted fit based on $w^{m,\mathrm{T}}$ in solid blue line (R).

CCCMA3.1, and CSIRO-Mk3.0) with three versions of the WRF modelling framework (which we call R1, R2, and R3, Skamarock et al., 2008) that differ in parameterisations of radiation, cumulus physics, surface physics, and planetary boundary layer physics. NARCliM output has been evaluated in terms of its ability to reproduce the observed mean climate (Ji et al., 2016; Olson et al., 2016b; Grose et al., 2015), climate extremes (Cortés-Hernández et al., 2015; Perkins-Kirkpatrick et al., 2016; Walsh et al., 2016; Kiem et al., 2016; Sharples et al., 2016), and important regional climate phenomena (Di Luca et al., 2016; Pepler et al., 2016). These studies demonstrate that while the downscaling has provided added value (Di Luca et al., 2016), a range of model errors are present within the ensemble. For the analysis, we focus on seasonal–mean temperature differences as modelled by the inner NARCliM domain RCMs between years 1990–2009 (present) and 2060–2079 (far future). We discard partial seasons from the analysis.

Here we average the temperatures over south-eastern Australian regions that include New South Wales (NSW) planning regions, ACT, and Victoria; see Fig. 2. Corresponding temperature observations are derived from the AWAP project (Jones et al., 2009). The models are generally cooler than the observations; however, in many cases the observations span the mean model climate.

In addition to computing weights of the form in Eq. (4), we also compute two variants of the weight: one based on penalising only the intercept a_m and internal variability σ_m, and an alternative weight based on penalising only the slope term b_m and internal variability σ_m. This is achieved by modifying

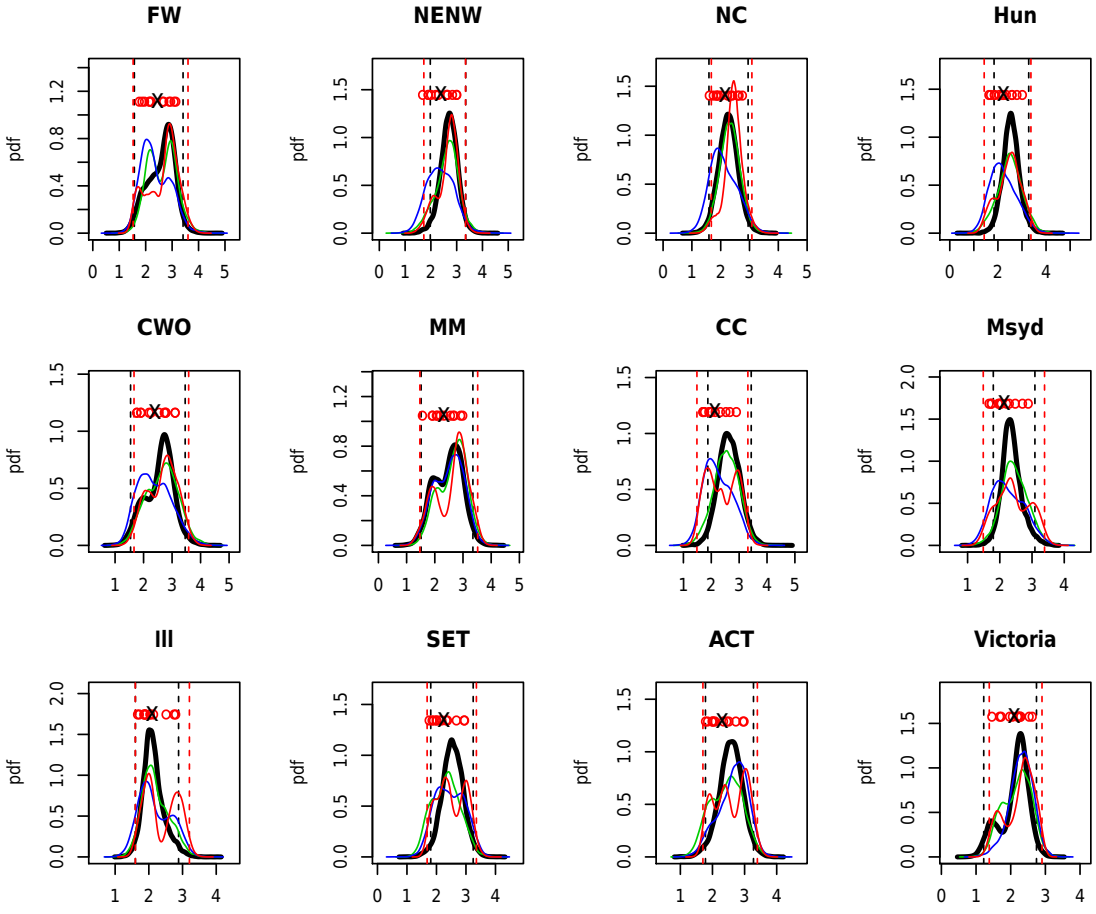

Figure 6. Posterior predictive projections of DJF temperature change in 2060–2079 compared to 1990–2009 for regions in south-eastern Australia. Black lines correspond to w^m weights, green lines to $w^{m,\mathrm{I}}$ weights and blue lines to $w^{m,\mathrm{T}}$ weights. Red lines are results from Olson et al. (2016a). Black vertical lines represent 95 % credible intervals, and red vertical lines represent the 95 % credible intervals obtained by Olson et al. (2016a). Circles represent the difference between the changes in temperature using the individual models. Black crosses indicate the simple ensemble mean of the changes in temperature.

Eq. (4) to

$$w^{m,\mathrm{I}} = \int L(y|a_m, b_p, \sqrt{\sigma_m^2 + \delta^2})\, p(a_m, \sigma_m|x^m)\mathrm{d}a_m\mathrm{d}\sigma_m \qquad (8)$$

or

$$w^{m,\mathrm{T}} = \int L(y|a_p, b_m, \sqrt{\sigma_m^2 + \delta^2})\, p(b_m, \sigma_m|x^m)\mathrm{d}b_m\mathrm{d}\sigma_m \qquad (9)$$

where $w^{m,\mathrm{I}}$ penalises models with large biases and wrong internal variability, and $w^{m,\mathrm{T}}$ penalises models with the wrong trend and internal variability. Note that our proposed weight w^m penalises bias, trend and internal variability simultaneously. The weights $w^{m,\mathrm{I}}$ and $w^{m,\mathrm{T}}$ can be computed by fitting the observation data to the model in Eq. (1) to obtain estimates for a_p and b_p, and using only the posterior samples of a_m, b_m and σ_m to complete the calculation.

Figure 3 shows the weight calculation of each model based on Eq. (4), for the CC region in season DJF. We used the observed data and the corresponding model output for the years 1990–2009. One can see how the three different types of weights behave relative to the bias and slope of the model output. For example, in Fig. 3, models 1, 2, 3 (left figure, middle row) and 10, 11, 12 (left figure, bottom row) have large bias compared to the other models; consequently, w^m and $w^{m,\mathrm{I}}$ give these models almost no weight. On the other hand these models simulated the trend well, and are preferred by $w^{m,\mathrm{T}}$.

The weighted fits are shown in the last two plots in the bottom row of Fig. 3. The black line is computed using w^m, according to

$$\hat{y}_t = \sum_{m=1}^{M} w^m (a_m + b_m \cdot t) \qquad (10)$$

where a_m and b_m are taken as the posterior means of the MCMC samples, and $t = -9.5, \ldots, 9.5$. A similar calculation is done based on $w^{m,\mathrm{I}}$, with $w^{m,\mathrm{T}}$ shown in green and blue respectively. The plots here suggest that the weights w^m

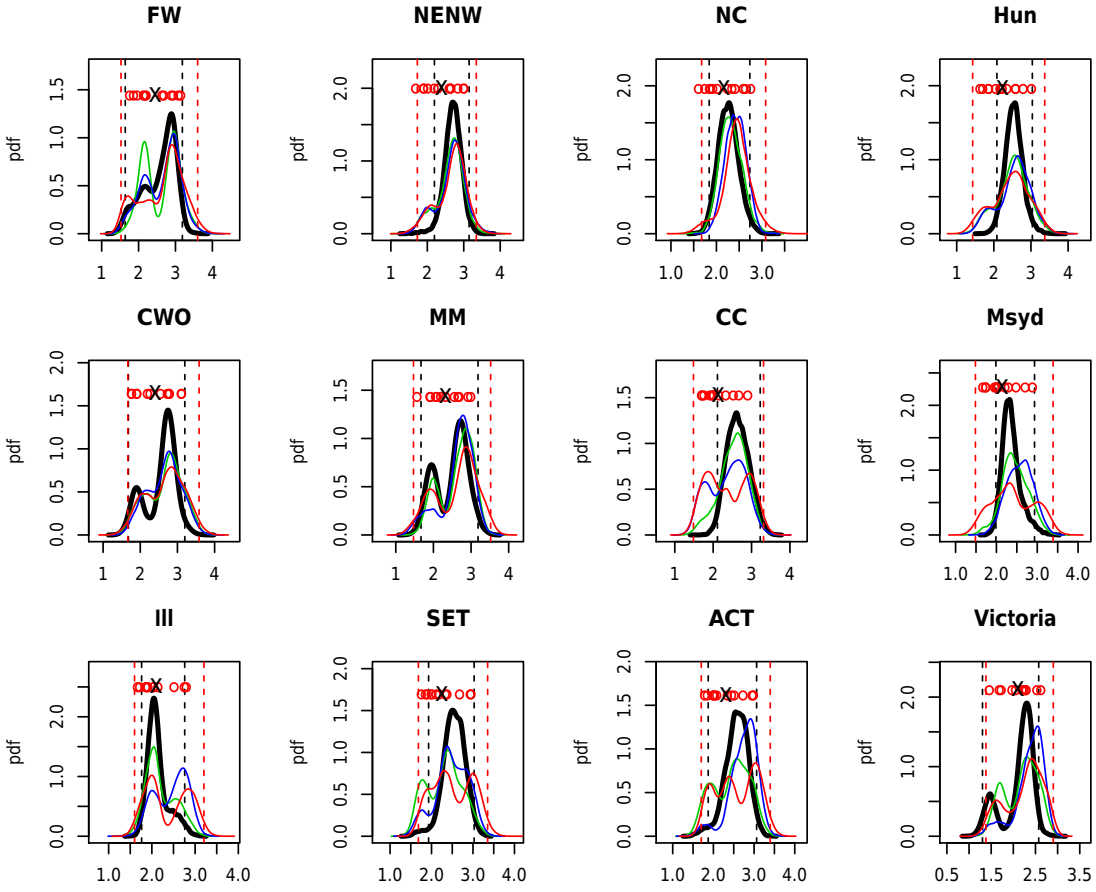

Figure 7. Bootstrapped weighted projections of DJF temperature change in 2060–2079 compared to 1990–2009 for regions in south-eastern Australia. Black lines correspond to w^m weights, green lines to $w^{m,\mathrm{I}}$ weights and blue lines to $w^{m,\mathrm{T}}$ weights. Red lines are results from Olson et al. (2016a). Black vertical lines represent 95 % credible intervals, and red vertical lines represent the 95 % credible intervals obtained by Olson et al. (2016a). Circles represent the difference between the changes in temperature using the individual models. Black crosses indicate the simple ensemble mean of the changes in temperature.

Table 1. Mean squared error and 95 % coverage probabilities for the three sets of weights.

	DJF		MAM		JJA		SON	
	MSE	Cov	MSE	Cov	MSE	Cov	MSE	Cov
w^m	48.43	0.938	14.44	0.958	14.15	0.910	41.86	0.917
$w^{m,\mathrm{I}}$	52.06	0.965	21.34	0.979	17.89	0.944	43.55	0.951
$w^{m,\mathrm{T}}$	56.93	0.993	30.74	0.979	20.45	0.972	39.79	1.000

are similar to $w^{m,\mathrm{I}}$, and better than $w^{m,\mathrm{T}}$ in this case. We note that there are dependencies between the RCMs driven by the same GCM. Our weight calculation does not model this dependence. So if different GCMs drive a different number of RCMs, the weights will over-represent some models but not others. While for most cases, the weights given by $w^{m,\mathrm{I}}$ provide similar weighted fits to w^m, Fig. 4 (showing the FW region for season DJF) demonstrates the instances where the weighted fit produced by $w^{m,\mathrm{I}}$ is clearly worse than w^m. The green line in the final plot shows that $w^{m,\mathrm{I}}$ produces a fit which is very close to the observation at the intercept but fails to capture the trend. This is unsurprising since this weight penalises deviations of a_m to a_p. Similarly, the blue line $w^{m,\mathrm{T}}$ appears to better capture the trend, but is clearly underestimating the bias, since it fails to penalise for bias. The weight w^m is a compromise between the two. From the weight plots in the first row, the models that have non-negligible weights under $w^{m,\mathrm{I}}$ are 6, 7, 11 and 12, corresponding to models whose intercepts are closest to the intercept of the observation model. The weights $w^{m,\mathrm{T}}$ are more

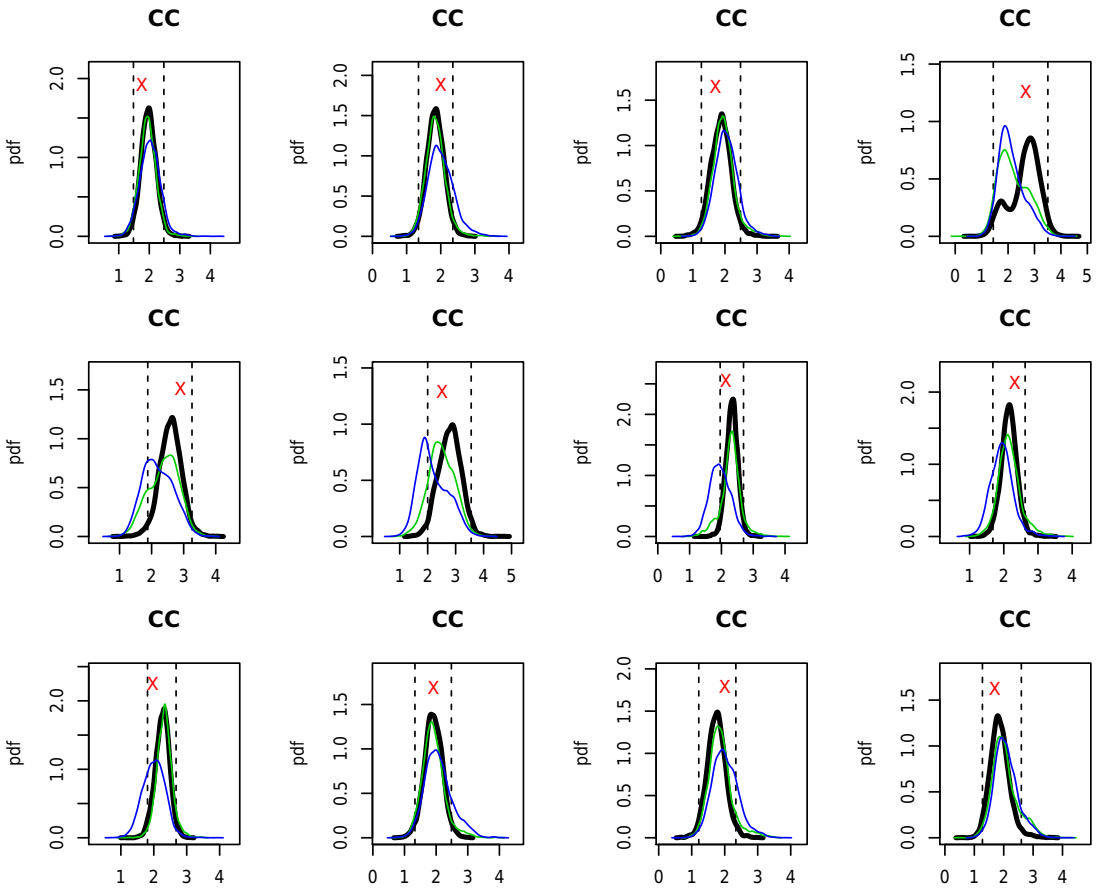

Figure 8. Cross validation of weighted projections of DJF temperature change in 2060–2079 compared to 1990–2009 for region CC in south-eastern Australia. Black lines correspond to w^m weights; green lines correspond to $w^{m,\mathrm{I}}$ weights and $w^{m,\mathrm{T}}$ weights. Each plot represents the weighted posterior predictive distribution of temperature change using the current ith model output as observation and the remaining 11 models are weighted. Vertical lines represent 95 % credible intervals. Crosses indicate the actual changes between the future model output and the current model output of the ith model.

spread out, giving high weights to models 1 and 2, which have large biases but capture the trend well. The weights w^m allocate most weight to models 6 and 7. Both models closely follow the shape of the observed data. In fact, in terms of trend, the weights $w^{m,\mathrm{T}}$ can capture more of the increase in trend better than w^m, this was the case in some of the regions in the SON season. A more formal evaluation of the three different weights will be carried out later in this section.

For seasons JJA and MAM, weights w^m and $w^{m,\mathrm{I}}$ were quite similar in all regions. These weights gave very close fits to the observation model, while $w^{m,\mathrm{T}}$ captured the trend well but gave biased fits to the observation. Generally for these two seasons, fewer models had non-neglible weights compared with DJF and SON. In DJF and SON, the weights were distributed more evenly across the models. This suggests that some of the individual models in JJA and MAM were performing strongly. Interestingly for MAM, the two models that dominated most regions are models 8 and 9; see for example the results for region CWO in Fig. 5. We can see the goodness of fit of these two models individually (see second row, right plot), and clearly they were markedly better than the other competing models.

The corresponding posterior predictive distribution of projections of change in temperature for season DJF over the different regions in south-eastern Australia are plotted in Fig. 6. The pdfs show the mean temperature change in the period 2060–2079 compared to 1990–2009. In order to obtain the posterior predictive projection pdf, we begin by first fitting MCMC for each future model output for the period 2060–2079, to obtain the posterior distribution of $p(a_m^{\mathrm{f}}, b_m^{\mathrm{f}}, \sigma_m^{\mathrm{f}}|x^m)$. Here we obtained 5000 posterior samples of $a_m^{\mathrm{f}}.b_m^{\mathrm{f}}$ and σ_m^{f}. We then obtain 10 000 random samples for each pdf. Each sample is obtained as follows.

1. With probability w^m, randomly select a sample from the posteriors of $a_m^{\mathrm{f}} b_m^{\mathrm{f}}$ and σ_m^{f}, say $a_{m,\mathrm{I}}^{\mathrm{f}} b_{m,\mathrm{I}}^{\mathrm{f}}$ and $\sigma_{m,\mathrm{I}}^{\mathrm{f}}$.

2. Simulate a predictive temperature series y_t^{f} according to

$$y_t^{\mathrm{f}} \sim N(a_{m,\mathrm{I}}^{\mathrm{f}} + b_{m,\mathrm{I}}^{\mathrm{f}}(t - t_1), \sigma_{m,\mathrm{I}}^{\mathrm{f}})$$

for $t = 2060, \ldots, 2079$ and $t_1 = 2069.5$. This process produces the posterior predictive samples y_t^{f} according to Eq. (7).

3. Compute current model estimate $\hat{y}_t^m = a_m + b_m \cdot (t - t_1)$, for $t = 1990, \ldots, 2009$ and $t_1 = 1999.5$ where a_m and b_m are posterior means based on model m and current model output x^m.

4. Compute the mean of the differences between future prediction y_t^{f} and $\hat{y}_t^m$.

This process produces the posterior predictive distributions for the mean difference between the posterior predictive samples y_t^{f} and the current estimate of climate.

We present the results for season DJF in Fig. 6. The black lines in Fig. 6 correspond to the pdf given by w^m, the green lines correspond to $w^{m,\mathrm{I}}$ and the blue lines correspond to $w^{m,\mathrm{T}}$. The red circles indicate the difference between the means of $\hat{y}_t$ and $\hat{y}_t^{\mathrm{f}}$ from each of the 12 models; the cross indicates the mean of these differences. Black vertical lines indicate the 95 % credibility interval for predictions made with w^m (black line). We can see that the pdfs based on w^m and $w^{m,\mathrm{I}}$ are similar to each other, while the ones given by $w^{m,\mathrm{T}}$ deviate substantially from the other two. We also superimposed the pdf obtained in Olson et al. (2016a) in red for comparison. The corresponding 95 % credible interval is shown in red vertical lines. It can be seen that our method generally provides a more precise prediction interval. In fact, to properly compare the two predictive distributions, we compute the posterior predictive distribution using the method described by Olson et al. (2016a). Unlike our posterior predictive pdf, the pdf in Olson et al. (2016a) was obtained by bootstrapping the errors, and does not account for the uncertainty in the parameter estimates of a_m, b_m and σ_m. To properly compare the effect of the different weights between our method and that of Olson et al. (2016a), we also show in Fig. 7 the bootstrapped pdf. Here the red line indicates the pdf using Olson et al. (2016a) weights with the 95 % credible interval shown in red vertical lines, and here we can see that Olson et al. (2016a) generally produce significantly larger credible intervals than our approach.

The incident of bimodality or multimodality is reduced in our approach compared to Olson et al. (2016a), suggesting a smoother mixing of models induced by our approach. Our approach generally produced sharper, more definite peaks in the posterior pdf. This could be due to the fact that our penalisation is done simultaneously, whereas Olson et al. (2016a) consider the penalty for bias and internal variability separately.

In order to assess the ensemble pdf, we performed a series of cross-validation checks. For each region at a given season, we have 12 current model outputs and 12 future model outputs. We select 1 of the models, m_i, and treat the current model output for m_i as the truth, and weigh the remaining 11 models. We then cycle through all 12 models, setting $m_i = 1, \ldots, 12$. Figure 8 shows the weighted projections for region CC in season DJF, each plot corresponding to using 1 of the 12 models as truth.

Table 1 shows the empirical coverage probabilities based on 144 sets of cross-validation datasets for each region, DJF, MAM, JJA and SON. The coverage probabilities are computed by counting the number of times the true mean change in temperature falls inside the 95 % credibility intervals, taken as the 0.025th and 0.975th quantile values of the posterior predictive samples. Each weighting method produces a different set of credibility intervals. We see from the table that both w^m and $w^{m,\mathrm{I}}$ perform quite close to the nominal level at 95 %, but the pdfs given by the weight $w^{m,\mathrm{T}}$ are a little too large. Finally, we also computed the mean squared error for each season: this is calculated as the average squared difference between the posterior predictive sample and the true value. The sums over all regions and all cross-validation sets are reported in Table 1. Overall, the weights w^m performed consistently well in this respect. w^m outperforms $w^{m,\mathrm{I}}$ in all seasons. The poorer performance of $w^{m,\mathrm{T}}$ is largely due to the large biases in the $w^{m,\mathrm{T}}$ models. One possibility of making $w^{m,\mathrm{T}}$ models more useful is to perform some kind of post hoc bias correction to the weighted estimates.

3 Conclusions

In this article we have introduced a new framework for computing Bayesian model weights. Our framework is novel, and requires minimal expert knowledge of model parameters. The fact that we do not require subjective expert prior knowledge makes the method more robust, since prior elicitation can sometimes be difficult, and different priors can lead to different conclusions.

We provided two alternative weight specifications under the same framework to aid interpretation of our weighting. One of the weights favours models with intercept terms that are close to the observation intercept. This weight does not penalise for trend deviations very well. An alternative weight which does not penalise for the intercept term can capture trend in the model very well. Both alternatives have deficiencies, and our proposed weight is a combination of the two. However, there are other potential avenues to explore with these alternative weights. For instance, for the weights based on trend and internal variability, it can be seen that the weighted model can capture trend extremely well but fails to account for bias, but applying some kind of post hoc bias correction may be a fruitful direction to pursue.

We validated our approach using cross validation, and showed that our posterior predictive distributions obtained correct empirical coverages, which is a desired property to possess, and provides us with some confidence in our approach. Our posterior predictive distributions also provided narrower confidence intervals than previous approaches. Fi-

nally, our model weighting framework is not restricted to data from univariate normal distributions, or linear models. This approach could be extended to handle dependent Gaussian data via a multivariate normal distribution, as well as non-linear and non-normal models.

Code and data availability. Code and data for the analyses carried out in this article are available in the Supplement.

The Supplement related to this article is available online at https://doi.org/10.5194/gmd-10-2321-2017-supplement.

Competing interests. The authors declare that they have no conflict of interest.

Acknowledgements. This work was supported by the National Research Foundation of Korea Grant funded by the Korean Government (MEST) (NRF-2009-0093069).

Edited by: James Annan
Reviewed by: Hans R. Künsch and one anonymous referee

References

Bhat, K. S., Haran, M., Terando, A., and Keller, K.: Climate Projections Using Bayesian Model Averaging and Space-Time Dependence, J. Agric. Biol. Envir. S., 16, 606?628, https://doi.org/10.1007/s13253-011-0069-3, 2011.

Buser, C. M., Künsch, H. R., Lüthi, D., Wild, M., and Schär, M. C.: Bayesian multi-model projections of climate: bias assumptions and interannual variability, Clim. Dynam., 33, 849–868, 2010.

Buser, C. M., Künsch, H. R., and Schär, C.: Bayesian multi-model projections of climate: generalization and application to ENSEMBLES results, Climate Res., 44, 227–241, 2010.

Christensen, J. H., Carter, T. R., Rummukainen, M., and Amanatidis, G.: Evaluating the performance and utility of regional climate models: the PRUDENCE project, Climatic Change, 81, 1–6, https://doi.org/10.1007/s10584-006-9211-6, 2007.

Cortés-Hernández, V. E., Zheng, F., Evans, J. P., Lambert, M., Sharma, A., and Westra, S.: Evaluating regional climate models for simulating sub-daily rainfall extremes, Clim. Dynam., 47, 1613–1628, https://doi.org/10.1007/s00382-015-2923-4, 2015.

Di Luca, A., Evans, J. P., Pepler, A., Alexander, L. V., and Argüeso, D.: Australian East Coast Lows in a Regional Climate Model ensemble, Journal of Southern Hemisphere Earth Systems Science, 66, 108–124, 2016.

Di Luca, A., Argüeso, D., Evans, J. P., de Elia, R., and Laprise, R.: Quantifying the overall added value of dynamical downscaling and the contribution from different spatial scales, J. Geophys. Res.-Atmos., 121, 1575–1590, https://doi.org/10.1002/2015JD024009, 2016.

Duan, Q., Ajami, N. K., Gao, X., and Sorooshian, S.: Multi-model ensemble hydrologic prediction using Bayesian model averaging, Adv. Water Resour., 30, 1371–1386, https://doi.org/10.1016/j.advwatres.2006.11.014, 2007.

Evans, J. P., Fita, L., Argüeso, D., and Liu, Y.: Initial NARCliM Evaluation, in MODSIM2013, 20th International Congress on Modelling and Simulation. Modelling and Simulation Society of Australia and New Zealand, December 2013, Adelaide, Australia, 2013.

Evans, J. P., Ji, F., Lee, C., Smith, P., Argüeso, D., and Fita, L.: Design of a regional climate modelling projection ensemble experiment – NARCliM, Geosci. Model Dev., 7, 621–629, https://doi.org/10.5194/gmd-7-621-2014, 2014.

Feser, F., Rrockel, B., von Storch, H., Winterfeldt, J., and Zahn, M.: Regional climate models add value to global model data: a review and selected examples, B. Am. Meteorol. Soc., 92, 1181–1192, 2011.

Fischer, A. M., Weigel, A. P., Buser, C. M., Knutti, R., Künsch, H. R., Liniger, M. A., Schär, C., and Appenzeller, C.: Climate change projections for Switzerland based on a Bayesian multi-model approach, Int. J. Climatol., 32, 2348–2371, https://doi.org/10.1002/joc.3396, 2012.

Gilks, W. R., Richardson, S., and Spiegelhalter, D. J.: Markov Chain Monte Carlo in Practice, Chapman and Hall, 512 pp., 1996.

Giorgi, F. and Bates, G. T.: The Climatological Skill of a Regional Model over Complex Terrain, Mon. Weather Rev., 117, 2325–2347, https://doi.org/10.1175/1520-0493(1989)117<2325:TCSOAR>2.0.CO;2, 1989.

Giorgi, F., Jones, C., and Asrar, G. R.: Addressing climate information needs at the regional level: the CORDEX framework, WMO Bull., 58, 175–183, 2009.

Goes, M., Urban, N. M., Tonkonojenkov, R., Haran, M., Schmittner, A., and Keller, K.: What is the skill of ocean tracers in reducing uncertainties about ocean diapycnal mixing and projections of the Atlantic Meridional Overturning Circulation?, J. Geophys. Res.-Oceans, 115, C12006, https://doi.org/10.1029/2010JC006407, 2010.

Grose, M. R., Bhend, J., Argüeso, D., Ekström, M., Dowdy, A., Hoffman, P., Evans, J. P., and Timbal, B.: Comparison of various climate change projections of eastern Australian rainfall, Aust. Meteorol. Oceanogr. J., 65, 72–89, 2015.

Hoeting, J. A., Madigan, D., Raftery, A. E., and Volinsky, C. T.: 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, 382–417, https://doi.org/10.1214/ss/1009212519, 1999.

Huttunen, J. M. J., Räiänen, J., Nissinen, A., Lipponen, A., and Kolehmainen, V.: Cross-validation analysis of bias models in Bayesian multi-model projections of climate, Clim. Dynam., 48, 1555–1570, https://doi.org/10.1007/s00382-016-3160-1, 2017.

Ji, F., Evans, J. P., Teng, J., Scorgie, Y., Argüeso, D., Di Luca, A., and Olson, R.: Evaluation of long-term precipitation and temperature WRF simulations for southeast Australia, Clim. Res., 67, 99–115, https://doi.org/10.3354/cr01366, 2016.

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

Kerkhoff, C., Künsch, H. R., and Schär, C.: A Bayesian hierarchical model for heterogeneous RCM-GCM multimodel ensembles,

J. Climate, 28, 6249–6266, https://doi.org/10.1175/JCLI-D-14-00606.1, 2015.

Kiem, A., Johnson, F., Westra, S., van Dijk, A., Evans, J. P., O'Donnell, A., Rouillard, A., Barr, C., Tyler, J., Thyer, M., Jakob, D., Woldemeskel, F., Sivakumar, B., and Mehrotra, R.: Natural hazards in Australia: droughts, Climatic Change, 139, https://doi.org/10.1007/s10584-016-1798-7, 2016.

Kirtman, B., Power, S. B., et al.: Near-term Climate Change: Projections and Predictability, in: Climate Change 2013: The Physical Science Basis, Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change, edited by: Stocker, v, Qin, D., Plattner, G.-K., Tignor, M., Allen, S. K., Borshung, J., Nauels, A., Xia, Y., Bex, V., and Midgley, P. M., Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA, 2013.

Mearns, L. O., Sain, S., Leung, L. R., Bukovsky, M. S., McGinnis, S., Biner, S., Caya, D., Arritt, R. W., Gutowski, W., Takle, E., Snyder, M., Jones, R. G., Nunes, A. M. B., Tucker, S., Herzmann, D., and McDaniel, L.: L. Sloanet al.: Climate change projections of the North American Regional Climate Change Assessment Program (NARCCAP), Climatic Change, 120, 965–975, https://doi.org/10.1007/s10584-013-0831-3, 2013.

Mendoza, P. A., Rajagopalan, B., Clark, M. P., Ikeda, K., and Rasmussen, R. M.: Statistical postprocessing of high-resolution regional climate model output, Mon. Weather Rev., 143, 1533–1553, https://doi.org/10.1175/MWR-D-14-00159.1, 2015.

Montgomery, J. M. and Nyhan, B.: Bayesian Model Averaging: Theoretical Developments and Practical Applications, Polit. Anal., 18, 245–270, https://doi.org/10.1093/pan/mpq001, 2010.

Olson, R., Fan, Y., and Evans, J. P.: A simple method for Bayesian model averaging of regional climate model projections: Application to southeast Australian temperatures', Geophys. Res. Lett., 43, 7661–7669, https://doi.org/10.1002/2016GL069704, 2016a.

Olson, R., Evans, J. P., Di Luca, A., and Argüeso, D.: The NARCliM project: model agreement and significance of climate projections, Climate Res., 69, 209–227, 2016b.

Pepler, A. S., Di Luca, A., Ji, F., Alexander, L. V., Evans, J. P., and Sherwood, S. C.: Projected changes in east Australian midlatitude cyclones during the 21st century, Geophys. Res. Lett., 43, 334–340, https://doi.org/10.1002/2015GL067267, 2016.

Perkins-Kirkpatrick, S., White, C., Alexander, L., Argueso, D., Boschat, G., Cowan, T., Evans, J., Ekstrom, M., Oliver, E., Phatak, A., and Purich, A.: Natural hazards in Australia: heatwaves, Climatic Change, 139, 101–114, https://doi.org/10.1007/s10584-016-1650-0, 2016.

Prömmel, K., Geyer, B., Jones, J. M., and Widmann, M.: Evaluation of the skill and added value of a reanalysis-driven regional simulation for Alpine temperature, Int. J. Climatol., 30, 760–773, 2010.

R Core Team: R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria, ISBN 3-900051-07-0, available at: http://www.R-project.org/, 2014.

Raftery, A. E., Gneiting, T., Balabdaoui, F., and Polakowski, M.: Using Bayesian Model Averaging to Calibrate Forecast Ensembles, Mon. Weather Rev., 133, 1155–1174, https://doi.org/10.1175/MWR2906.1, 2005.

Sen, P. K.: Estimates of the Regression Coefficient Based on Kendall's Tau, J. Am. Stat. Assoc., 63, 1379–1389, https://doi.org/10.1080/01621459.1968.10480934, 1968.

Sharples, J. J., Cary, G., Fox-Hughes, P., Mooney, S., Evans, J. P., Fletcher, M., Fromm, M., Baker, P., Grierson, P., and McRae, R.: Natural hazards in Australia: extreme bushfire, Climatic Change, 139, 85–99, https://doi.org/10.1007/s10584-016-1811-1, 2016.

Skamarock, W. C., Klemp, J. B., Dudhia, J., Gill, D. O., Barker, D. M., Duda, M. G., Huang, X.-Y., Wang, W., and Powers, J. G.: A Description of the Advanced Research WRF Version 3 NCAR Technical Note NCAR/TN-475+STR, NCAR, Boulder, CO, USA, 2008.

Solman, S. A., Sanchez, E., Samuelsson, P., da Rocha, R. P., Li, L., Marengo, J., Pessacg, N. L., Remedio, A. R. C., Chou, S. C., Berbery, H., Le Treut, H., de Castro, M., and Jacob, D.: Evaluation of an ensemble of regional climate model simulations over South America driven by the ERA-Interim reanalysis: model performance and uncertainties, Clim. Dynam., 41, 1139–1157, https://doi.org/10.1007/s00382-013-1667-2, 2013.

Terando, A., Keller, K., and Easterling, W. E.: Probabilistic projections of agro-climate indices in North America, J. Geophys. Res.-Atmos., 117, D08115, https://doi.org/10.1029/2012JD017436, 2012.

van der Linden, P. and Mitchell, J. F. B.: ENSEMBLES: Climate Change and its Impacts: Summary of Research and Results from the ENSEMBLES Project, Met Office Hadley Centre, Exeter, UK, 2009.

Walsh, K., White, C. J., McInnes, K., Holmes, J., Schuster, S., Richter, H., Evans, J. P., Di Luca, A., and Warren, R. A.: Natural hazards in Australia: storms, wind and hail, Climatic Change, 139, 55–67, https://doi.org/10.1007/s10584-016-1737-7, 2016.

Wang, X., Huang, G., and Baetz, B. W.: Dynamically-downscaled probabilistic projections of precipitation changes: A Canadian case study, Environ. Res., 148, 86–101, https://doi.org/10.1016/j.envres.2016.03.019, 2016.

Whetton, P., Hennessy, K., Clarke, J., McInnes, K., and Kent, D.: Use of Representative Climate Futures in impact and adaptation assessment, Climatic Change, 115, 433?442, https://doi.org/10.1007/s10584-012-0471-z, 2012.