Geosci. Model Dev. Discuss., doi:10.5194/gmd-2016-291, 2017
Manuscript under review for journal Geosci. Model Dev.
Published: 4 January 2017
© Author(s) 2017. CC-BY 3.0 License.

Geoscientific Model Development Discussions
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, Australia
[2] Department of Atmospheric Sciences, Yonsei University, South Korea
[3]Climate Change Research Centre and ARC Centre of Excellence for Climate System Science, UNSW, Australia

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

Abstract.

We present a novel Bayesian statistical approach to computing model weights in climate change projection ensembles. 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, as well as 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 et al. (1989); Christensen et al. (2007); van der Linden et al. (2009); Evans et al. (2013); Evans et al. (2014); Mearns et al. (2013); Solman et al. (2013); Olson et al. (2016)). 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 et al. (2009); Prommel 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 et al. (1989); van der Linden et al. (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. (2016); 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 years) 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. Wang et al. (2016) method is applicable to relatively short time series, however convergence still influences model predictions.

Geosci. Model Dev. Discuss., doi:10.5194/gmd-2016-291, 2017
Manuscript under review for journal Geosci. Model Dev.
Published: 4 January 2017
© Author(s) 2017. CC-BY 3.0 License.

Geoscientific Model Development Discussions

Olson et al. (2016) 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. While this approach breaks important new ground, it 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 smoothing choice is not explicitly considered. Second, in the projection stage the Olson et al. (2016) implementation does not fully account for the uncertainty in model biases and in standard deviation of the model-data residuals.

In this article, we proposed 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 non-informative priors can be used, which makes implementation of the method much more straight forward. In Olson et al. (2016) 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. (2016)).

2 Posterior predictive weighting

In this section, we introduce the Bayesian methodology for weighting model output based on current day observations. 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

$$y_t = a_p + b_p(t - t_0) + \epsilon \tag{1}$$

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. 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

$$x_t^m = a_m + b_m(t - t_0) + \epsilon_t \tag{2}$$

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 Bayes rule,

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

Geosci. Model Dev. Discuss., doi:10.5194/gmd-2016-291, 2017
Manuscript under review for journal Geosci. Model Dev.
Published: 4 January 2017
© Author(s) 2017. CC-BY 3.0 License.

Geoscientific Model Development Discussions
Open Access
EGU

where $L(x^m|\cdot)$ denotes the likelihood of obtaining data x^m from model m. In this work, non-informative priors are used throughout.

We would like to weight the models based on the similarity of output x_t^m to the observation data, 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 is larger than σ_m, 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 range between 0.5 to 1 Kelvin. For our analyses of seasonally averaged temperature data in Section 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

$$w^m = \int L(y|a_m, b_m, \sigma_m + \delta) p(a_m, b_m, \sigma_m|x^m) da_m db_m d\sigma_m \tag{4}$$

where $L(y|a_m, b_m, \sigma_m + \delta)$ 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 Equation 4 will be larger if a_m, b_m and $\sigma_m + \delta$ 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 Equation 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

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

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}.

In order to understand this weight, we suppose for the moment that the data y comes from say, a $N(0,1)$. Suppose also that x^m comes from $N(\mu, \sigma)$, then if the posterior distribution of μ and σ are centered around 0 and 1, x^m should be assigned 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 $L(y)$ computed for $\mu = -2, \ldots, 2$ and $\sigma = 1$, and the right panel shows $L(y)$ computed for $\mu = 0$, $\sigma = 0.01, \ldots, 5$. The figure shows the changes in $L(y)$, and hence the weight, as parameter values move away from the true values of 0 and 1.

[Figure 1 about here.]

Geosci. Model Dev. Discuss., doi:10.5194/gmd-2016-291, 2017
Manuscript under review for journal Geosci. Model Dev.
Published: 4 January 2017
© Author(s) 2017. CC-BY 3.0 License.

Geoscientific Model Development Discussions
Open Access
EGU

2.1 Computation

In most cases, the posterior distributions $p(a_m, b_m, \sigma_m|x^m)$ in Equation 3 will be analytically intractable, however samples from this distribution can be easily obtained via Markov Chain Monte Carlo (MCMC). Many software packages performing MCMC are available, for the analysis in this paper, we used the **MCMCpack** library of the statistical package **R**, R Core Team (2013). 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. 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.

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

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

where $L(y|a_{m,i}, b_{m,i}, \sigma_{m,i} + \delta)$ denotes the likelihood of y under the ith sample of $a_{m,i}, b_{m,i}$ and $\sigma_{m,i}$ 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 Equation 6. Finally, the weights should be normalised by the constraint $\sum_{m=1}^{M} w^m = 1$.

To obtain the Bayesian model averaged posterior samples for Equation 5, we simply set for $i = 1, \ldots, N$,

$$a_{BMA,i} = \sum_{m=1}^{M} w^m a_{m,i}, \quad b_{BMA,i} = \sum_{m=1}^{M} w^m b_{m,i}, \quad \sigma_{BMA,i} = \sum_{m=1}^{M} w^m \sigma_{m,i},$$

where $a_{m,i}$, $b_{m,i}$ and $\sigma_{m,i}$ denotes the ith MCMC sample for model m.

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

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

2.2 Application

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

Geosci. Model Dev. Discuss., doi:10.5194/gmd-2016-291, 2017
Manuscript under review for journal Geosci. Model Dev.
Published: 4 January 2017
© Author(s) 2017. CC-BY 3.0 License.

Geoscientific Model Development Discussions

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. (2016), 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 modeled 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-east Australian regions that include New South Wales (NSW) planning regions, ACT, and Victoria, see Figure 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 Equation 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 Equation 4 to

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

or

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

where $w^{m,I}$ penalises models with large biases and wrong internal variability, and $w^{m,T}$ 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 $w^{m,I}$ and $w^{m,T}$ can be computed by fitting the observation data to the model in Equation 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 Equation 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 Figure 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 $w^{m,I}$ gives these models almost no weight. On the other hand these models simulated the trend well, and are preferred by $w^{m,T}$.

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

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

where a_m and b_m are taken as the posterior means of the MCMC samples, and $t = 0, \ldots, 18$. A similar calculation is done based on $w^{m,I}$, and $w^{m,T}$ shown in green and blue respectively. The plots here suggest the weights w^m are perhaps slightly better than $w^{m,I}$, both of which are better than $w^{m,T}$.

Geosci. Model Dev. Discuss., doi:10.5194/gmd-2016-291, 2017
Manuscript under review for journal Geosci. Model Dev.
Published: 4 January 2017
© Author(s) 2017. CC-BY 3.0 License.

Geoscientific Model Development Discussions

While for most cases, the weights given by $w^{m,I}$ provide similar weighted fits as w^m, Figure 4 (showing the FW region for the season DJF) demonstrates the instances where the weighted fit produced by $w^{m,I}$ is clearly worse than w^m, the green line in the final plot shows that $w^{m,I}$ produces a fit which is very close to the observation at the intercept, but fails to capture trend. This is unsurprising since this weight penalises deviations of a_m to a_p. Similarly, the blue line $w^{m,T}$ 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 weights plots in the first row, the models that have non-negligible weights under $w^{m,I}$ are 4,5,7,8,10 and 11, corresponding to models whose intercepts are closest to the intercept of the observation model. The weights $w^{m,T}$ are more spread out, giving high weights to models 1 and 2 which have large biases but capture the trend well. The last five models take less weight, this corresponds to models that have smaller trend values. The weights w^m allocates most weight to model 6 and 7, both models closely follow the shape of the observed data. In fact, in terms of trend, the weights $w^{m,T}$ generally perform similarly to w^m, but sometimes they 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 the seasons JJA and MAM, the weights w^m and $w^{m,I}$ were quite similar in all regions. These weights gave very close fits to the observation model, while $w^{m,T}$ captured 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 Figure 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 the season DJF over the different regions in south-east Australia are plotted in Figure 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 $p(a_m^f, b_m^f, \sigma_m^f | x^m)$. Here we obtained 5000 posterior samples of a_m^f, b_m^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 $a_m^f.b_m^f$ and σ_m^f, say $a_{m,i}^f.b_{m,i}^f$ and $\sigma_{m,i}^f$
2. simulate a predictive temperature series y_t^f according to

$$y_t^f \sim N(a_{m,i}^f + b_{m,i}^f(t - t_0), \sigma_{m,i}^f)$$

for $t = 2060, \ldots, 2079$ and $t_0 = 2060$. This process produces the posterior predictive samples y_t^f according to Equation 7.
3. compute current model estimate $\hat{y}_t^m = a_m + b_m * (t - t_0)$, for $t = 1990, \ldots, 2009$ and $t_0 = 1990$ 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 $\hat{y}_t^m$.

Geosci. Model Dev. Discuss., doi:10.5194/gmd-2016-291, 2017
Manuscript under review for journal Geosci. Model Dev.
Published: 4 January 2017
© Author(s) 2017. CC-BY 3.0 License.

Geoscientific Model Development Discussions

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 the season DJF in Figure 6. The black lines in Figure 6 correspond to the pdf given by w^m, the green lines correspond to $w^{m,I}$ and the blue lines correspond to $w^{m,T}$. The red circles indicate the difference between the means of $\hat{y}_t$ and $\hat{y}_t^f$ 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 pdf based on w^m and $w^{m,I}$ are similar to each other, while the ones given by $w^{m,T}$ deviate substantially from the other two. We also superimposed the pdf obtained in Olson et al. (2016) 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. (2016). Unlike our posterior predictive pdf, the pdf in Olson et al. (2016) 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. (2016) , we also show in Figure 7 the bootstrapped pdf, here the red line indicates the pdf using Olson et al. (2016) weights with 95% credible interval shown in red vertical lines, and here we can see that Olson et al. (2016) generally produces significantly larger credible intervals than our approach.

The incident of bimodality or multimodality is reduced in our approach compared to Olson et al. (2016), 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. (2016) considers the penalty for bias and internal variability separately.

[Figure 2 about here.]

[Figure 3 about here.]

[Figure 4 about here.]

[Figure 5 about here.]

[Figure 6 about here.]

[Figure 7 about here.]

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 one 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 the 12 models, setting $m_i = 1, \ldots, 12$. Figure 8 shows the weighted projections for the region CC in the season DJF, each plot correspond to using one 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

Geosci. Model Dev. Discuss., doi:10.5194/gmd-2016-291, 2017
Manuscript under review for journal Geosci. Model Dev.
Published: 4 January 2017
© Author(s) 2017. CC-BY 3.0 License.

Geoscientific Model Development Discussions

falls inside the 95% credibility intervals, taken as the 0.025th and 0.975th quantile value 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 $w^{m,I}$ perform quite close to the nominal level at 95%, but the pdf's given by the weight $w^{m,T}$ are too large, always producing coverages that are much higher than 0.95. Finally, we also computed the mean squared error for each season, this is calculated as the average squared differences between the posterior predictive sample and the true value, the sum over all regions and all cross-validation sets are reported in Table 1. Overall, the weights w^m performed consistently better in this respect, and as expected, w^m outperforms $w^{m,I}$ by a larger margin in the seasons DJF and SON. The poorer performance of $w^{m,T}$ is largely due to the large biases in the $w^{m,T}$ models, one possibility of making $w^{m,T}$ models more useful is to perform some kind of post-hoc bias correction to the weighted estimates.

[Figure 8 about here.]

[Table 1 about here.]

3 Conclusions

In this article we have introduced a new framework for computing Bayesian model weights. Our framework is entirely 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, rather than matching the intercept (at time zero), we might consider matching the estimates around the middle point of the time duration. 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 with our approach. Our posterior predictive distributions also provided narrower confidence intervals than previous approaches.

Finally, our model weighting framework is not restricted to data from Normal distributions, or linear models. This approach could be extended to non-linear and non-Normal models.

Code and Data Availability

Code and data for the analyses carried out in this article is available in the Supplementary Materials.

Geosci. Model Dev. Discuss., doi:10.5194/gmd-2016-291, 2017
Manuscript under review for journal Geosci. Model Dev.
Published: 4 January 2017
© Author(s) 2017. CC-BY 3.0 License.

References

Bhat, K. S., M. Haran, A. Terando, and K. Keller (2011), Climate Projections Using Bayesian Model Averaging and Space-Time Dependence, J. Agric. Biol. Environ. Stat., 16(4), 606?628, doi:10.1007/s13253-011-0069-3.

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

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

Cortés-Hernández V.E., F. Zheng, J.P. Evans, M. Lambert, A. Sharma, S. Westra (2015), Evaluating regional climate models for simulating sub-daily rainfall extremes. Climate Dynamics, doi: 10.1007/s00382-015-2923-4.

Di Luca, A., J.P. Evans, A. Pepler, L.V. Alexander and D. Arg´ueso (2016) Australian East Coast Lows in a Regional Climate Model ensemble. Journal of Southern Hemisphere Earth Systems Science, 66(2), 108-124.

Di Luca, A., D. Arg´ueso, J.P. Evans, R. de Elia and R. Laprise (2016) Quantifying the overall added value of dynamical downscaling and the contribution from different spatial scales. Journal of Geophysical Research ? Atmospheres, 121(4), 1575-1590, doi: 10.1002/2015JD024009.

[Duan, Q., N. K. Ajami, X. Gao, and S. Sorooshian (2007), Multi-model ensemble hydrologic prediction using Bayesian model averaging, Adv. Water Resour., 30(5), 1371?1386, doi:10.1016/j.advwatres.2006.11.014.

Evans, J. P., L. Fita, D. Argüeso, and Liu, Y. (2013), 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.

Evans, J. P., F. Ji, C. Lee, P. Smith, D. Argüeso, and L. Fita (2014)], Design of a regional climate modelling projection ensemble experiment – NARCliM, Geosci Model Dev, 7(2), 621?629, doi:10.5194/gmd-7-621-2014.

Feser F, B. Rrockel, H. von Storch, J. Winterfeldt, and M. Zahn (2011) Regional climate models add value to global model data: a review and selected examples. Bulletin of American Meteorological Society, 92, 1181?1192.

Fischer, A. M., A. P. Weigel, C. M. Buser, R. Knutti, H. R. Künsch, M. A. Liniger, C. Schär, and C. Appenzeller (2012), Climate change projections for Switzerland based on a Bayesian multi-model approach, Int. J. Climatol., 32(15), 2348?2371, doi:10.1002/joc.3396.

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

Giorgi, F., and G. T. Bates (1989), The Climatological Skill of a Regional Model over Complex Terrain, Mon. Weather Rev., 117(11), 2325?2347, doi:10.1175/1520-0493(1989)117<2325:TCSOAR>2.0.CO;2.

Giorgi, F., C. Jones, and G. R. Asrar (2009), Addressing climate information needs at the regional level: the CORDEX framework, WMO Bull., 58(3), 175?183.

Goes, M., N. M. Urban, R. Tonkonojenkov, M. Haran, A. Schmittner, and K. Keller (2010), 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(12), doi:10.1029/2010JC006407.

Grose, M.R., J.Bhend, D. Arg´ueso, M. Ekstr´om, A. Dowdy, P. Hoffman, J.P. Evans, B. Timbal (2015), Comparison of various climate change projections of eastern Australian rainfall. Australian Meteorological and Oceanographic Journal, 65(1), 72-89.

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(4), 382?417, doi:10.1214/ss/1009212519.

Geosci. Model Dev. Discuss., doi:10.5194/gmd-2016-291, 2017
Manuscript under review for journal Geosci. Model Dev.
Published: 4 January 2017
© Author(s) 2017. CC-BY 3.0 License.

Ji, F., J.P. Evans, J. Teng, Y. Scorgie, D. Arg´ueso, A. Di Luca and R. Olson (2016), Evaluation of long-term precipitation and temperature WRF simulations for southeast Australia. Climate Research, 67, 99-115, doi:10.3354/cr01366.

Jones, D. A., W. Wang, and R. Fawcett (2009), High-quality spatial climate data-sets for Australia, Aust. Meteorol. Oceanogr. J., 58(4), 233?248.

Kerkhoff, C., H. R. Künsch, and C. Schär (2015), A Bayesian hierarchical model for heterogeneous RCM-GCM multimodel ensembles, J. Clim., 28(15), 6249?6266, doi:10.1175/JCLI-D-14-00606.1.

Kiem, A., F. Johnson, S. Westra, A. van Dijk, J.P. Evans, A. O'Donnell, A. Rouillard, C. Barr, J. Tyler, M. Thyer, D. Jakob, F. Woldemeskel, B. Sivakumar and R. Mehrotra (2016) Natural hazards in Australia: droughts. Climatic Change, accepted 26 August 2016.

Kirtman, B. et al. (2013), 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 T. F. Stocker, D. Qin, G.-K. Plattner, M. Tignor, S. K. Allen, J. Borshung, A. Nauels, Y. Xia, V. Bex, and P. M. Midgley, Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA.

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

Mearns, L. O. et al. (2013), Climate change projections of the North American Regional Climate Change Assessment Program (NARCCAP), Clim. Change, 120(4), 965?975, doi:10.1007/s10584-013-0831-3.

Mendoza, P. A., B. Rajagopalan, M. P. Clark, K. Ikeda, and R. M. Rasmussen (2015), Statistical postprocessing of high-resolution regional climate model output, Mon. Weather Rev., 143(5), 1533?1553, doi:10.1175/MWR-D-14-00159.1.

Montgomery, J. M., and B. Nyhan (2010), Bayesian Model Averaging: Theoretical Developments and Practical Applications, Polit. Anal., 18(2), 245?270, doi:10.1093/pan/mpq001.

Olson, R., Fan, Y. and J. P. Evans (2016), A simple method for Bayesian model averaging of regional climate model projections: Application to southeast Australian temperatures', Geophysical Research Letters, vol. 43, no. 14, pp. 7661-7669, http://dx.doi.org/10.1002/2016GL069704

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

Pepler, A.S., A. Di Luca, F. Ji, L.V. Alexander, J.P. Evans and S.C. Sherwood (2016) Projected changes in east Australian midlatitude cyclones during the 21st century. Geophysical Research Letters, 43(1), doi:10.1002/2015GL067267).

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

Prömmel K, B. Geyer, J. M. Jones, M. Widmann (2010) Evaluation of the skill and added value of a reanalysis-driven regional simulation for Alpine temperature. International Journal of Climatology, 30, 760?773.

Sharples, J.J., G. Cary, P. Fox-Hughes, S. Mooney, J.P. Evans, M. Fletcher, M. Fromm, P. Baker, P. Grierson and R. McRae (2016) Natural hazards in Australia: extreme bushfire. Climatic Change, accepted 3 September 2016.

R Core Team (2013), R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org/.

Raftery, A. E., T. Gneiting, F. Balabdaoui, and M. Polakowski (2005), Using Bayesian Model Averaging to Calibrate Forecast Ensembles, Mon. Weather Rev., 133(5), 1155 -1174, doi:10.1175/MWR2906.1.

Geosci. Model Dev. Discuss., doi:10.5194/gmd-2016-291, 2017
Manuscript under review for journal Geosci. Model Dev.
Published: 4 January 2017
© Author(s) 2017. CC-BY 3.0 License.

Sen, P. K. (1968), Estimates of the Regression Coefficient Based on Kendall?s Tau, J. Am. Stat. Assoc., 63(324), 1379?1389, doi:10.1080/01621459.1968.10480934.

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

Solman, S. A. et al. (2013), Evaluation of an ensemble of regional climate model simulations over South America driven by the ERA-Interim reanalysis: model performance and uncertainties, Clim. Dyn., 41(5-6), 1139?1157, doi:10.1007/s00382-013-1667-2.

Terando, A., K. Keller, and W. E. Easterling (2012), Probabilistic projections of agro-climate indices in North America, J. Geophys. Res. Atmospheres, 117(D8), D08115, doi:10.1029/2012JD017436.

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

Whetton, P., K. Hennessy, J. Clarke, K. McInnes, and D. Kent (2012), Use of Representative Climate Futures in impact and adaptation assessment, Clim. Change, 115(3-4), 433?442, doi:10.1007/s10584-012-0471-z.

Wang, X., G. Huang, and B. W. Baetz (2016) Dynamically-downscaled probabilistic projections of precipitation changes: A Canadian case study, Environmental Research, 148, 86?101, doi:10.1016/j.envres.2016.03.019.

Geosci. Model Dev. Discuss., doi:10.5194/gmd-2016-291, 2017
Manuscript under review for journal Geosci. Model Dev.
Published: 4 January 2017
© Author(s) 2017. CC-BY 3.0 License.

Geoscientific Model Development Discussions
Open Access
EGU

List of Figures

1 Pictorial representation of the weight distribution on μ and σ 13
2 New South Wales planning regions, the ACT and the state of Victoria. 14
3 Results for CC region of south-east Australia, in the DJF season. Top row, weights w^m of 12 models based on Equation 4 (L), Equation 8, $w^{m,I}$ (M) and Equation 9 $w^{m,T}$ (R). Each triplets represents a GCM (MIROC3.2, ECHAM5, CCCMA3.1, and CSIRO-Mk3.0). Middle row and first plot of last row: fitted observations according to Equation 1 (red dashed line) and fitted model output according to Equation 2 for 12 models. Last row: weighted fit based on w^m in solid black line (M) and weighted fit based on $w^{m,I}$ in solid green line and weighted fit based on $w^{m,T}$in solid blue lines (L). 15
4 Results for FW region of south-east Australia, in the DJF season. Top row, weights w^m of 12 models based on Equation 4 (L), Equation 8, $w^{m,I}$ (M) and Equation 9 $w^{m,T}$ (R). Each triplets represents a GCM (MIROC3.2, ECHAM5, CCCMA3.1, and CSIRO-Mk3.0). Middle row and first plot of last row: fitted observations according to Equation 1 (red dashed line) and fitted model output according to Equation 2 for 12 models. Last row: weighted fit based on w^m in solid black line (M) and weighted fit based on $w^{m,I}$ in solid green line and weighted fit based on $w^{m,T}$in solid blue lines (L). 16
5 Results for CWO region of south-east Australia, in the MAM season. Top row, weights w^m of 12 models based on Equation 4 (L), Equation 8, $w^{m,I}$ (M) and Equation 9 $w^{m,T}$ (R). Each triplets represents a GCM (MIROC3.2, ECHAM5, CCCMA3.1, and CSIRO-Mk3.0). Middle row and first plot of last row: fitted observations according to Equation 1 (red dashed line) and fitted model output according to Equation 2 for 12 models. Last row: weighted fit based on w^m in solid black line (M) and weighted fit based on $w^{m,I}$ in solid green line and weighted fit based on $w^{m,T}$in solid blue lines (L). 17
6 Posterior predictive projections of DJF temperature change in 2060-2079 compared to 1990-2009 for regions in south-east Australia. Black lines correspond to w^m weights, green lines correspond to $w^{m,I}$ weights and blue lines to $w^{m,T}$ weights. Red lines are results from Olson et al. (2016). Black vertical lines represent 95% credible intervals, and red vertical lines represent the 95% credible intervals obtained by Olson et al. (2016). Circles represent the difference between the changes in temperature using the individual models. Black cross indicates the simple ensemble mean of the changes in temperature. 18
7 Bootstrapped weighted projections of DJF temperature change in 2060-2079 compared to 1990-2009 for regions in south-east Australia. Black lines correspond to w^m weights, green lines correspond to $w^{m,I}$ weights and blue lines to $w^{m,T}$ weights. Red lines are results from Olson et al. (2016). Black vertical lines represent 95% credible intervals, and red vertical lines represent the 95% credible intervals obtained by Olson et al. (2016). Circles represent the difference between the changes in temperature using the individual models. Black cross indicates the simple ensemble mean of the changes in temperature. 19
8 Cross validation of weighted projections of DJF temperature change in 2060-2079 compared to 1990-2009 for region CC in south-east Australia. Black lines correspond to w^m weights, green lines correspond to $w^{m,I}$ weights and $w^{m,T}$ 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. 20

Geosci. Model Dev. Discuss., doi:10.5194/gmd-2016-291, 2017
Manuscript under review for journal Geosci. Model Dev.
Published: 4 January 2017
© Author(s) 2017. CC-BY 3.0 License.

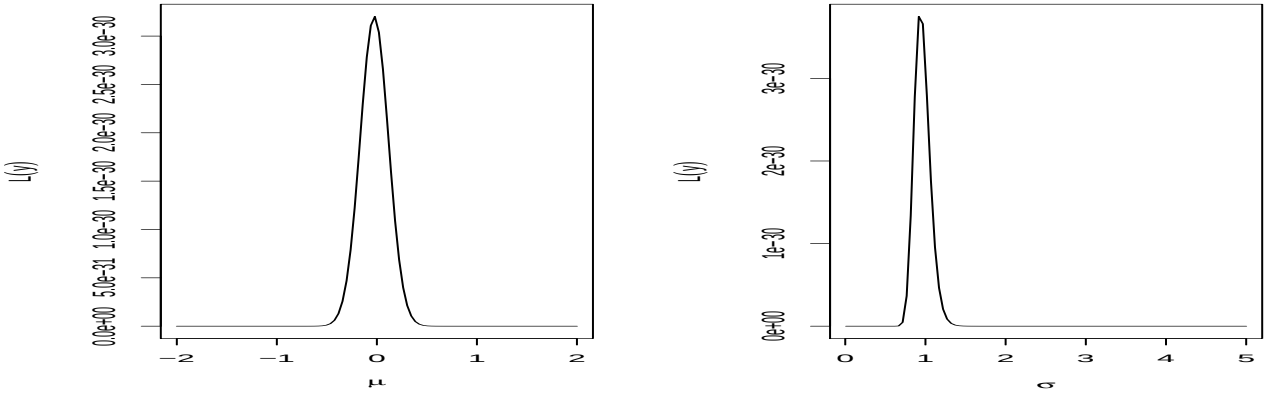

Figure 1. Picturial representation of the weight distribution on μ and σ .

Geosci. Model Dev. Discuss., doi:10.5194/gmd-2016-291, 2017
Manuscript under review for journal Geosci. Model Dev.
Published: 4 January 2017
© Author(s) 2017. CC-BY 3.0 License.

Geoscientific Model Development Discussions
Open Access
EGU

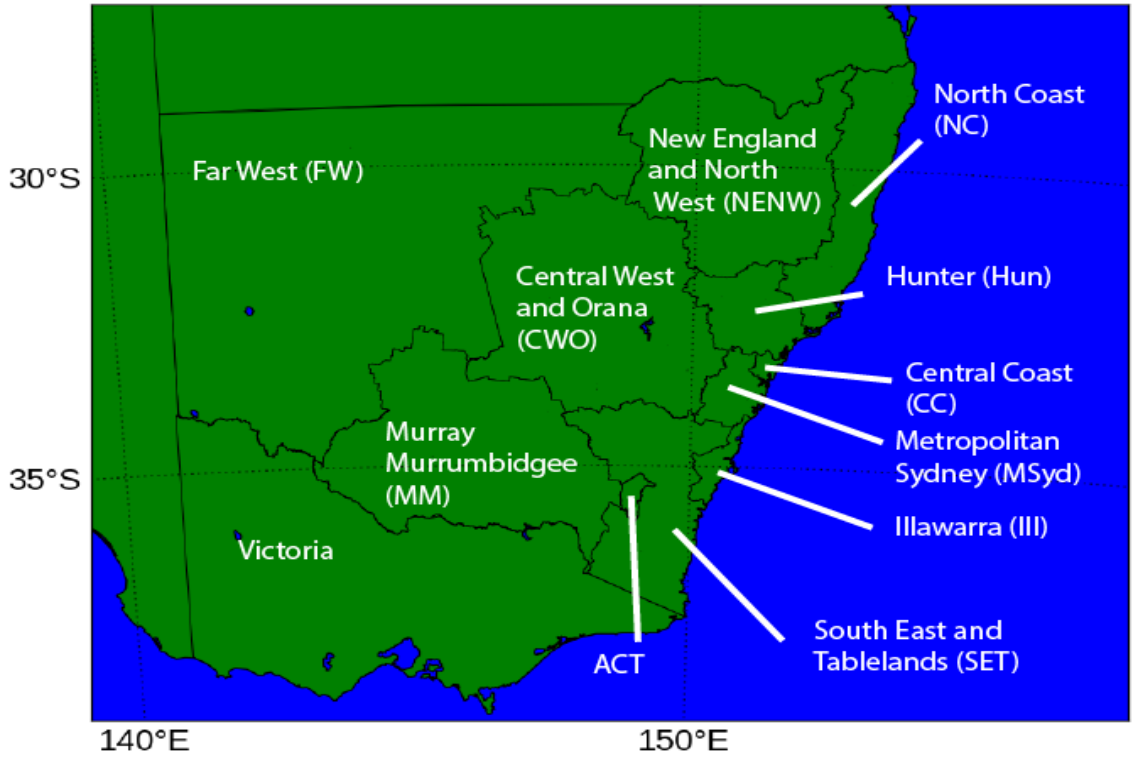

Figure 2. New South Wales planning regions, the ACT and the state of Victoria.

Geosci. Model Dev. Discuss., doi:10.5194/gmd-2016-291, 2017
Manuscript under review for journal Geosci. Model Dev.
Published: 4 January 2017
© Author(s) 2017. CC-BY 3.0 License.

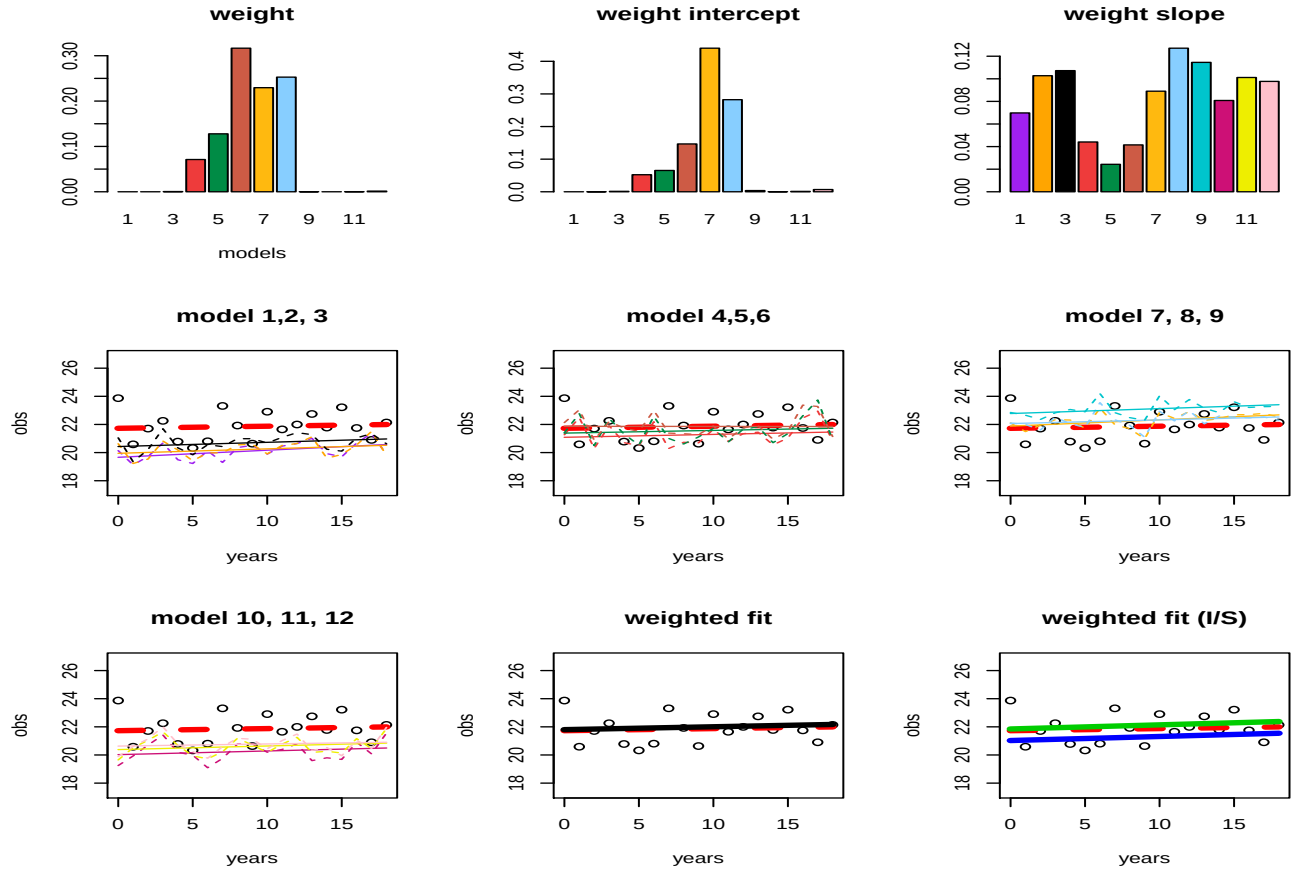

Figure 3. Results for CC region of south-east Australia, in the DJF season. Top row, weights w^m of 12 models based on Equation 4 (L), Equation 8, $w^{m,I}$ (M) and Equation 9 $w^{m,T}$ (R). Each triplets represents a GCM (MIROC3.2, ECHAM5, CCCMA3.1, and CSIRO-Mk3.0). Middle row and first plot of last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

Geosci. Model Dev. Discuss., doi:10.5194/gmd-2016-291, 2017
Manuscript under review for journal Geosci. Model Dev.
Published: 4 January 2017
© Author(s) 2017. CC-BY 3.0 License.

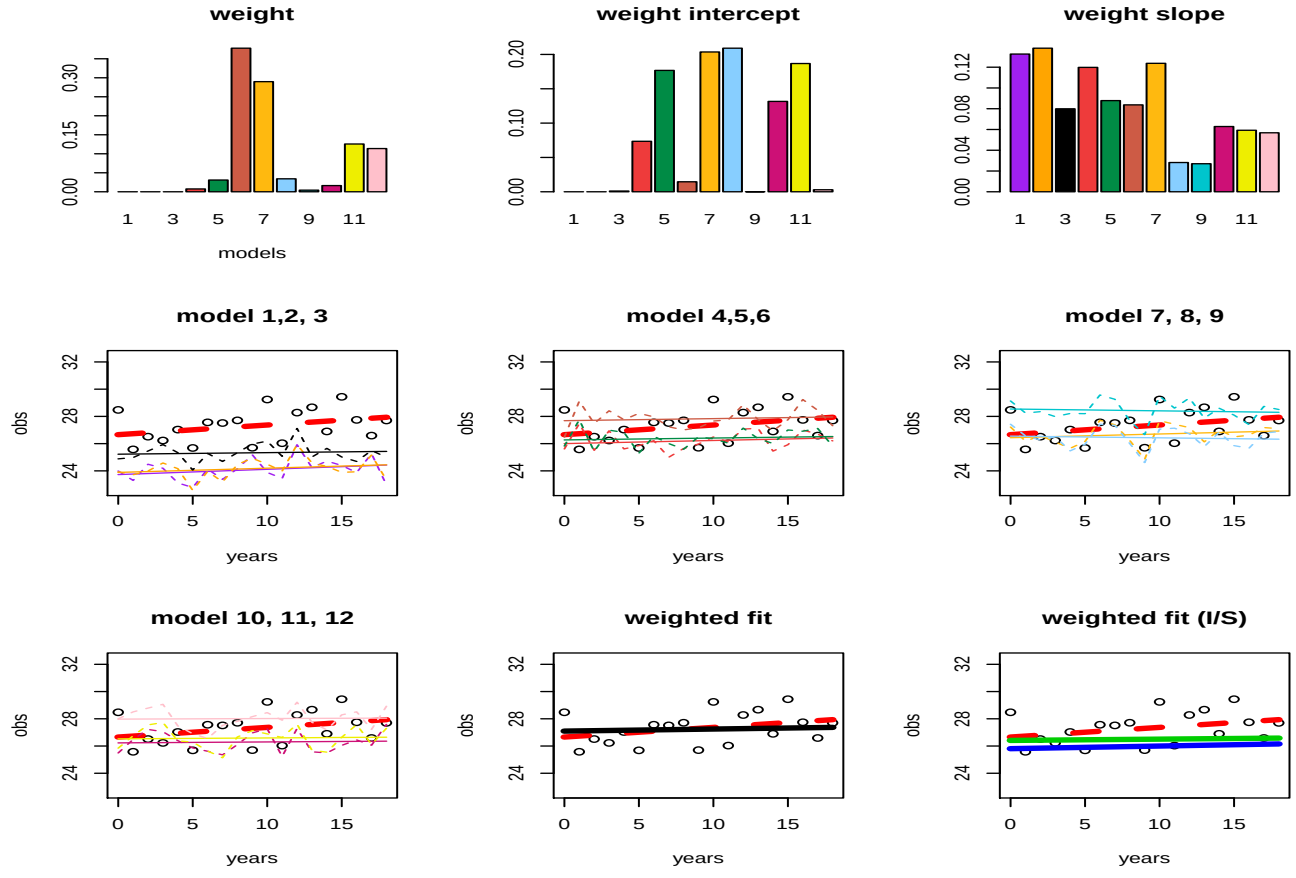

Figure 4. Results for FW region of south-east Australia, in the DJF season. Top row, weights w^m of 12 models based on Equation 4 (L), Equation 8, $w^{m,I}$ (M) and Equation 9 $w^{m,T}$ (R). Each triplets represents a GCM (MIROC3.2, ECHAM5, CCCMA3.1, and CSIRO-Mk3.0). Middle row and first plot of last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

Geosci. Model Dev. Discuss., doi:10.5194/gmd-2016-291, 2017
Manuscript under review for journal Geosci. Model Dev.
Published: 4 January 2017
© Author(s) 2017. CC-BY 3.0 License.

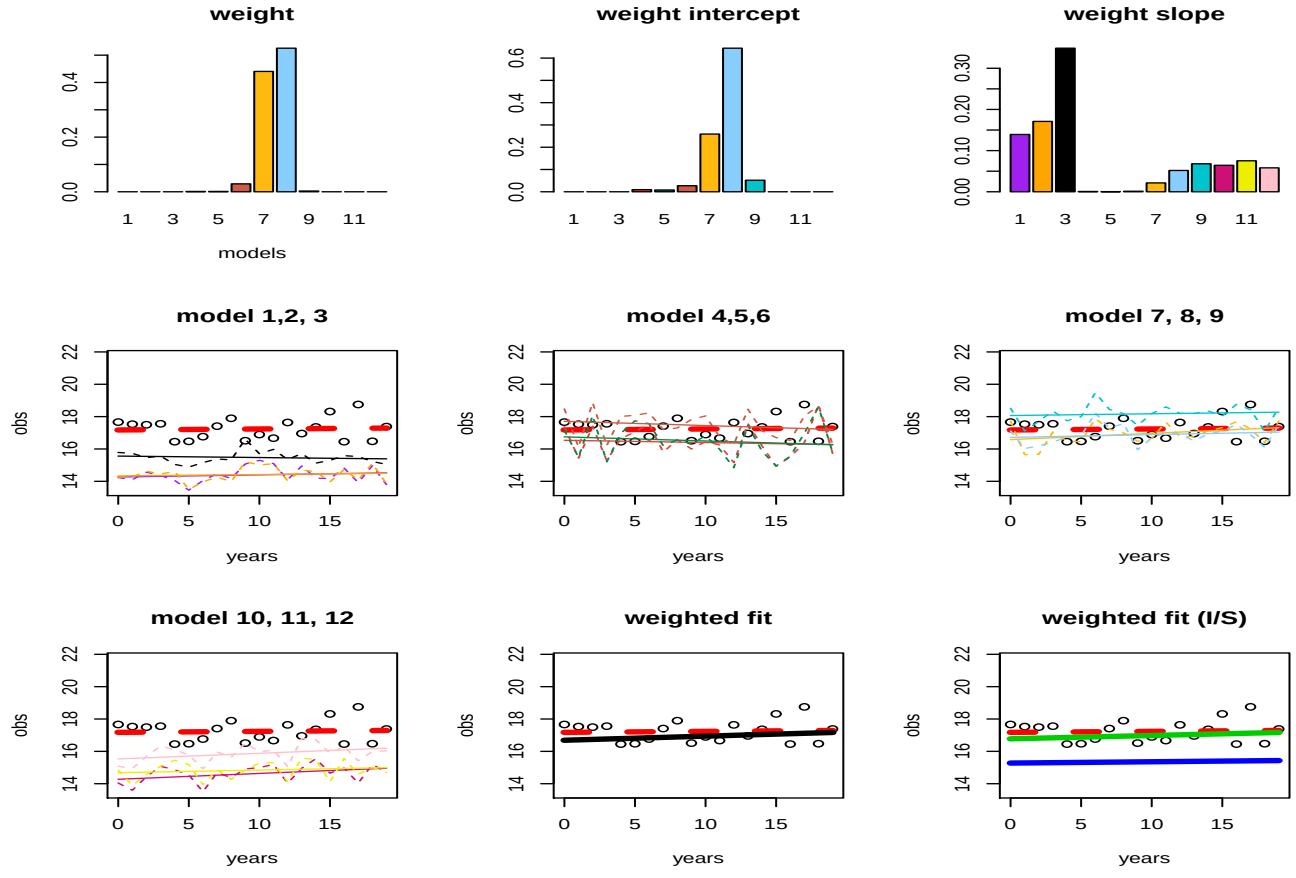

Figure 5. Results for CWO region of south-east Australia, in the MAM season. Top row, weights w^m of 12 models based on Equation 4 (L), Equation 8, $w^{m,I}$ (M) and Equation 9 $w^{m,T}$ (R). Each triplets represents a GCM (MIROC3.2, ECHAM5, CCCMA3.1, and CSIRO-Mk3.0). Middle row and first plot of last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

Geosci. Model Dev. Discuss., doi:10.5194/gmd-2016-291, 2017
Manuscript under review for journal Geosci. Model Dev.
Published: 4 January 2017
© Author(s) 2017. CC-BY 3.0 License.

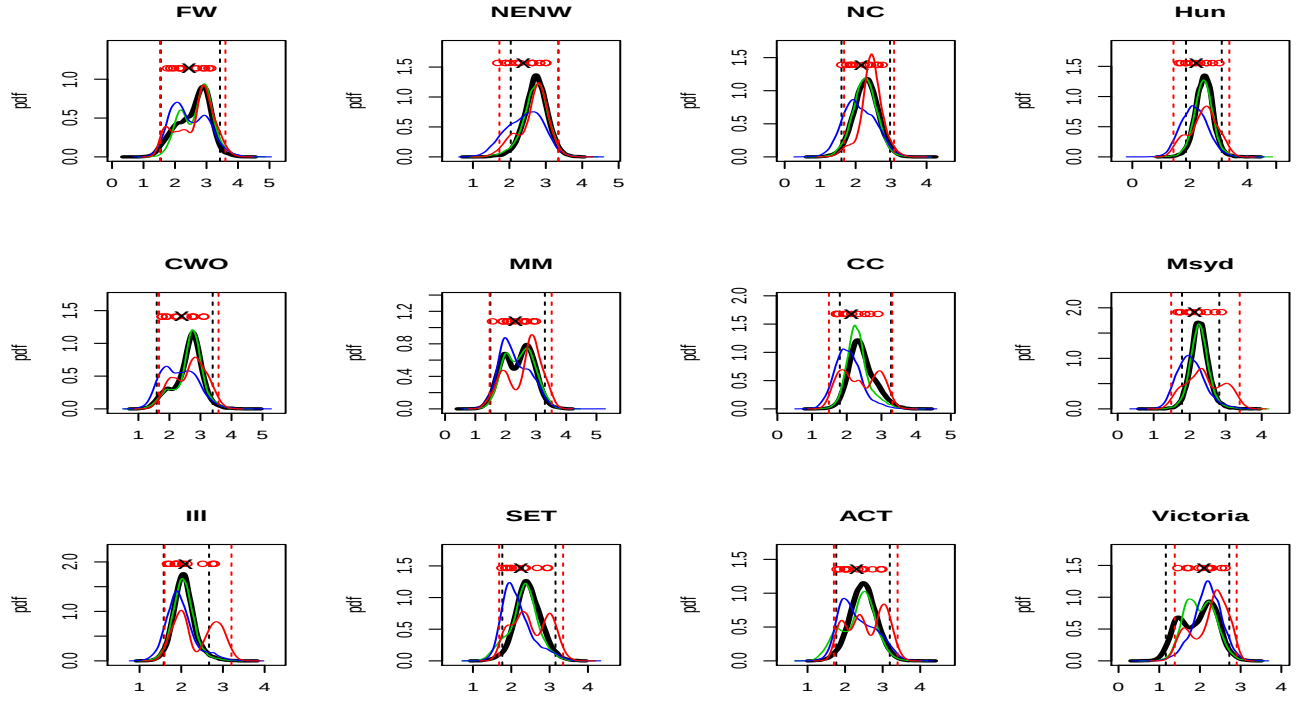

Figure 6. Posterior predictive projections of DJF temperature change in 2060-2079 compared to 1990-2009 for regions in south-east Australia. Black lines correspond to w^m weights, green lines correspond to $w^{m,I}$ weights and blue lines to $w^{m,T}$ weights. Red lines are results from Olson et al.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

Geosci. Model Dev. Discuss., doi:10.5194/gmd-2016-291, 2017
Manuscript under review for journal Geosci. Model Dev.
Published: 4 January 2017
© Author(s) 2017. CC-BY 3.0 License.

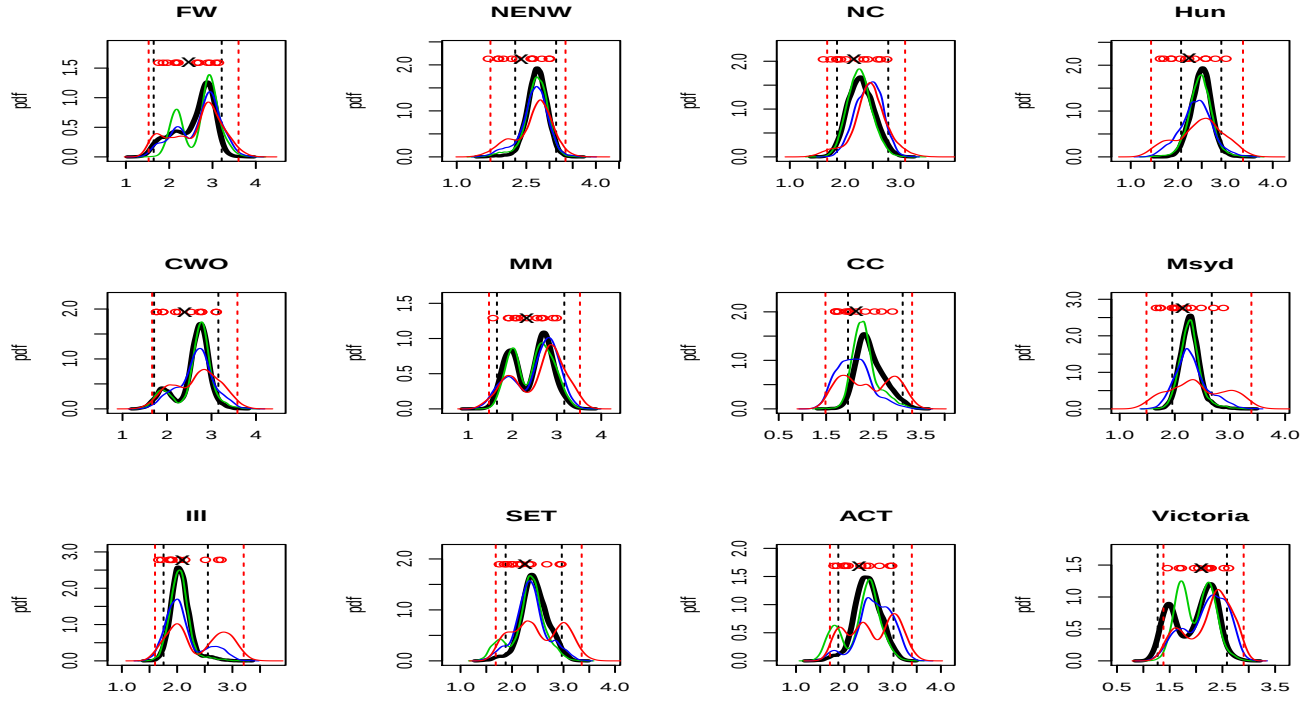

Figure 7. Bootstrapped weighted projections of DJF temperature change in 2060-2079 compared to 1990-2009 for regions in south-east Australia. Black lines correspond to w^m weights, green lines correspond to $w^{m,I}$ weights and blue lines to $w^{m,T}$ weights. Red lines are results from Olson et al.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

Geosci. Model Dev. Discuss., doi:10.5194/gmd-2016-291, 2017
Manuscript under review for journal Geosci. Model Dev.
Published: 4 January 2017
© Author(s) 2017. CC-BY 3.0 License.

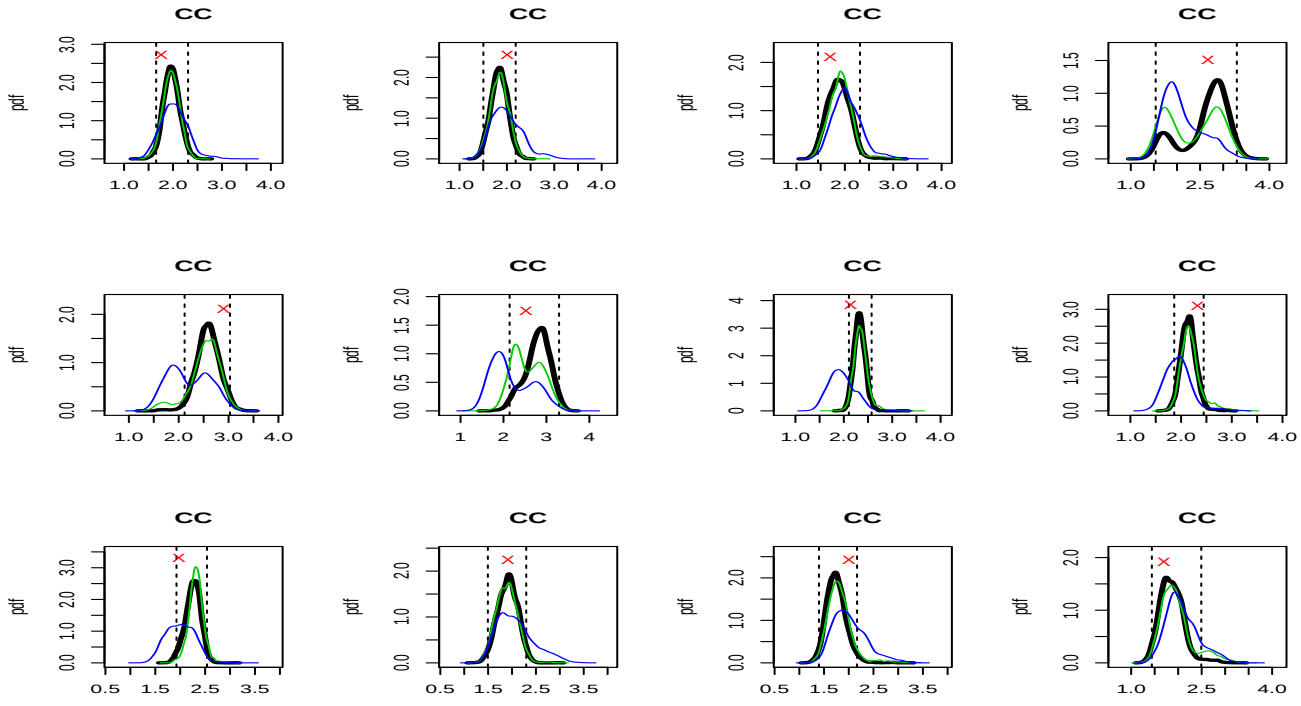

Figure 8. Cross validation of weighted projections of DJF temperature change in 2060-2079 compared to 1990-2009 for region CC in south-east Australia. Black lines correspond to w^m weights, green lines correspond to $w^{m,I}$ weights and $w^{m,T}$ weights. Each plot represents the weighted posterior 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.

Geosci. Model Dev. Discuss., doi:10.5194/gmd-2016-291, 2017
Manuscript under review for journal Geosci. Model Dev.
Published: 4 January 2017
© Author(s) 2017. CC-BY 3.0 License.

Geoscientific Model Development Discussions
Open Access
EGU

List of Tables

1 Mean squared error and 95% coverage probabilities for the three sets of weights. 22

Geosci. Model Dev. Discuss., doi:10.5194/gmd-2016-291, 2017
Manuscript under review for journal Geosci. Model Dev.
Published: 4 January 2017
© Author(s) 2017. CC-BY 3.0 License.

Geoscientific Model Development Discussions
Open Access
EGU

	DJF		MAM		JJA		SON	
	MSE	Cov	MSE	Cov	MSE	Cov	MSE	Cov
w^m	48.35	0.944	14.40	0.951	14.13	0.910	41.89	0.917
$w^{m,I}$	51.61	0.951	14.61	0.965	14.39	0.931	43.53	0.930
$w^{m,T}$	56.93	0.993	30.94	0.979	20.50	0.986	40.42	1.000

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