Response to interactive comment 1

Red is reviewer comments, Black response.

The authors report on a statistical method loosely connected to bayesian analysis to post-process ensemble simulations of regional climate models. they apply the method and report results for seasonal temperatures in australia over two 20 year periods. The authors claim that their approach is entirely novel, which I do not see but which is a matter of definition of ”new”.

We have removed the word ”entirely” from the conclusions.

More important i found a significant lack of theoretical background, rather the model is set up in a rather adhoc fashion: we need a weighted combination of regional climate models $->$ so why not taking bayesian model averaging with some weights; where to get the weights from $->$ why not taking the likelihood with some ”uninformed prior”; how to calculate the weights $->$ why not taking mcmc; where to get the ensemble from $->$ why not taking the regional climate simulations.

The content of the paper is certainly worth to be published but not in this very way as it is presented currently. Papers in GMD should not only report on the technical aspects but also on the theoretical background because this allows to draw conclusions about the assumptions made for the specific implementations.

By theoretical background it appears the reviewer is asking for more justification of the choices made in establishing the framework. First we add ”in order to create probabilistic projections” to the end of the first sentence in the abstract to emphasise that the calculation of model weights is not an end in itself. The introduction then discusses past attempts to

do this, their limitations, and that our proposed framework overcomes some of these limitations. A sentence is added on line 2 of page 3 to justify the use of uninformed priors "The use of non-informative allows the data to discriminate amongst models, whereas informative prior reflect the scientist's personal knowledge, and can lead to more subjective analyses. Non-Informative priors are sometimes considered preferable when data contains sufficient information". Further justification for the use of MCMC is added as the first sentence of section 2.1: "The procedure for the calculation of weights is designed to be applicable regardless of the distributional forms chosen to model the data." and after the first sentence of section 2 "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 to obtain robust estimators for our analyses, but we have to excluded them from presentation as the procedure produced similar results to that of the Normal error assumption (indicating no major violations from Normality)." The choice of regional climate projection ensemble is arbitrary but was made here due to ease of access and familiarity of the authors.

Furthermore even the technical aspect is only mildly covered because at no point except in the very last sentence it is said that the current implementation relies on (univariate! not mentioned!) normal distributed random variables. Additionally the use of a Bayesian approach is only marginally. Firstly, the modelling of uncertainty is rather adhoc (see my remark above) in the sense that the model parameters especially the precision values of the residuals are treated in a non common fashion, standard approaches eg in described in Gilks et al 1996 at least consider the normal-invers gamma model with a wide prior on the hyper-parameter of the invers gamma component , secondly the treatment of observations in the likelihood and the treatment of simulations in the prior does not consider the dependency between the residual components.

Text is added after the first sentence of section 2 "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 to obtain robust estimators for our analyses, but we have to excluded them from presentation as the procedure produced similar results to that of the Normal error assumption (indicating no major violations from Normality)."

It is not correct to say that we treat residual variance as non-random. We model the residual variance of model output in the standard way, with a standard prior. We account for the discrepancy between model output and observation with an additional term for the variance, this discrepancy comes from the fact that observations are subject to an additional source of measurement error.

We agree that if the errors are dependent, then one should use an appropriate model to account for correlated errors.

This holds for the actual analysis where the epsilons could be modeled as an AR process in time giving rise to a multivariate normal for the x_m which is then transformed by a linear projection into a bivariate normal for the trend and offset. The likelihood can be written as a function of the same projection jointly upon observations and simulations such that the negative loglikelihood looks like this

$$nll \sim (\theta_m - \theta_o)^t P^t (\Sigma_m + \Sigma_o)^{(-1)} P(\theta_m - \theta_o) + \log(\det(\Sigma_m + \Sigma_o))$$

where $\theta_{m,o}$ are bivariate vectors containing the offset and slope of the linear fitted function (or any amplitude of a generalized additive model) and P is a matrix containing in its columns the 1's for the offset and the $(t - t_o)$for the slope (or any function g_k in case of the generalized additive models). note that in this approach the uncertainty in the covariance

matrices $\Sigma_{m,o}$ is not yet treated, but this is possible if eg $\Sigma_o = \sigma^2 I$ (I identity matrix) and setting an inverss gamma prior for σ_o^2.

If we are not mistaken, the equation above is not a likelihood. But in general, the independent univariate normal distributional assumption on x_m and y is easily extended to the multivariate case by introducing a prior on a general covariance matrix, a standard prior would be the inverse-Wishart prior.

The inclusion of the P matrix shows that the currently chosen way of trend fitting introduces a correlation between the error in offset and trend because the function $(t - t_o)$ does not sum to zero over the full time interval and therefore any error in the slope will produce an error in the average which has to be compensated by a negative deviation in the error of the mean.

We have now reparameterised the models to centre at the middle of the time series.

This shows that the analytical analysis of the approach using normal distributions firstly does not need a mcmc numerical solution and allows to draw important conclusions about the characteristics of the method.

MCMC is typically needed when the Bayesian analysis does not use specific forms of priors, leaving the posterior distributions intractable. We have left the discussion with the use of MCMC to allow for incorporations of general forms of the prior distribution, because conjugate priors can often be too restrictive. As noted above text has been added to the start of section 2 and 2.1 to clarify this, we add "Conjugate analyses for certain classes of models, including Gaussian error models are often possible, leading to analytical forms for posterior

distributions. In this work, we choose to present the results with non standard priors, and use MCMC for computation. The later approach is much easier when extending to more complex modelling scenarios".

Similar remarks can be made for the predictive probability densities, also here a lot can be learned from the (multivariate) normal densities. Note that the multivariate densities always include the univariate case but not the other way around.

The predictive density for observation should take the same form as model output.

Summary The paper is worth to be published in GMD but the authors should comment/ add modifications according to my general remarks. Detailed comments are found in the annotated pdf document. Please also note the supplement to this comment: http://www.geosci-model-dev-discuss.net/gmd-2016-291/gmd-2016-291-RC1- supplement.pdf

- p2, l 4, we have changed the text to "However, their approach still suffers from shortcomings."
- p2,, l 23, see responses above.
- p3, l 8, we do not treat δ as an unknown parameters, therefore we do not place a prior on δ.
- p3, l10, the term here is the standard deviation, not the precision as is used in many Bayesian text books. We do not in fact place a prior on this term, this is a likelihood term which determines the relative weights for us.
- p3, l 20, the weights are predefined, and following the Bayesian model averaging framework, the BMA model is the weighted sum of the competing models.

- p3, l 25, the use of the word centering refers to the location, i.e., about the mean or median of the distribution, it does not imply whether the distribution is skewed or not.

- p3, l 27, we have added some text here to make it clearer. New text now appear just before Sec 2.1 on page 4.

- p4, l1, by most cases, we mean that if we deviate away from the Normal or multivariate Normal distributions, tractable solutions do not exist. See our responses above for further discussions.

- p4, l 13, our weight calculations involve a likelihood term, and a mixture of posteriors term. If the likelihood is Gaussian and the mixture of posteriors are mixtures of Gaussians, then one may expect to obtain analytic solutions. However, we do not have Gaussian posteriors unless conjugate priors were used.

- p4, l 17, this is now removed, as we do not require this calculation throughout the paper.

- p4, l 20, this takes the same form as Equation 5, only that the parameters are based on future model output, as this is the data that is used to make predictions for future observation.

- p5, l 30, we have updated our results by centering the regression on the bias parameter.

- p6, l 24, calculations with training sample do not involve mixtures, the weighted mixture model is used for prediction.

- p7, l 6, the 95% credibility internal, where the location of the interval is indicated by black lines. These are computed from 2.5 and 97.5 quantiles.

- p8, l 29, we have used a simple linear model in the demonstrations in this paper, but the approach can be used for non-linear models, or any generalised models.

Response to interactive comment 2

Red is reviewer comments, Black response.

General evaluation

As I have shown above, the method proposed here is not basically different from what is used e.g. in Buser et al. (2009), Climate Dynamics 33, 849- 868. The main difference comes from the prior distribution.

We believe that Buser et al. (2009) also do not use a Bayesian Model averaging approach, whereas our approach provides a weight for each model that has a useful interpretation. Our approach for projection future climate is to down weight the models that are performing poorly. In addition, we do not need to make any independence assumptions between all the data related to the observed time series and model outputs.

For me, the assumption that one model is perfect is not natural. I prefer the idea that all models have strengths and weaknesses and therefore deviations are rather on a continuum. But since it is unknown which model is perfect, in the end the analysis still uses all models and thus the results are presumably not that different.

Our posterior predictive distribution in Equation (7), we say that the future prediction is obtained from a weighted sum of the model outputs, i.e., each model contribute w_i towards the prediction. So while we do not explicitly put a prior on model probability, the starting point for our analyses is the same – we believe that all models should contribute towards prediction, and there isn't a single best model.

The only time we assume one model is truth is when we perform the cross-validation checks. Since we do not have future observations to perform this evaluation, we instead

perform it 12 times, each time assuming that one model represents the "truth", our proxy observations. This allows us to apply the framework as described and test the implications for the probabilistic projections compared to a known future.

The second assumption, namely that the quality of a model can be judged on its behavior during the control alone, is harder to accept. I am not a climate scientist, but a model that agrees well with the observations in the control, but has a much slower or a much faster warming than all the other models seems doubtful to me. On the other hand, a model can be consistently too warm over the whole period from control until the end of the future period, but still give a good estimate of climate change. The authors point out that agreement between models can be due to common model errors. On the other hand, a good agreement between models and observations in the control can also be due to too much tuning, or it can be just a coincidence in case there are many models.

We agree that if a model performs well under recent climate conditions, this does not guarantee that it will perform well under future climate conditions. We do however argue that if a model is not able to perform well under recent climate conditions then it cannot be trusted to perform well under future conditions. That is, performing well compared to the control is a good indication, though not a sufficient condition for reliable performance in future climates. The model spread that remains after accounting for this necessary condition provides an indication of the future climate uncertainty. Text to this effect will be added to the beginning of section 2.

A different criticism concerns the fact that the dependence between different model chains is not taken into account. In my experience, there is non- negligible dependence between RCMs driven by the same GCM and this should be reflected in the likelihood. However, I guess that this would lead to complications.

We agree that there is non-negligible dependences between the RCMs driven by the same GCM. This translates to correlated weights. If the number of RCMs driven by the same GCM is different, this could lead to uneven weighting, Text to this affect can be added near the beginning of section 2.2, as a caveat of this approach.

Detailed comments

- p. 2, equation (1): I would use the parametrization

$$y_t = a_p + b_p(t - t_1) + \epsilon_t, \quad \text{where} t_1 = t_0 + T/2$$

  That is, the slope term is the same, but the intercept is the value in the middle instead of the beginning of the period. Keeping the intercept fixed as in equation (9) on p. 5 makes then much more sense to me.

  This parameterisation is easier to interpret, so we have now updated the paper using this parameterisation.

- p. 3, l. 4: In my experience with multimodel ensembles for Europe (PRUDENCE, ENSEMBLES and CORDEX) it is not true that models vary less than the observations from one year to the next. On the contrary, models often overestimate the variability by a factor up to 2. Also additive corrections of standard errors are strange. Instrumental and gridding errors should be independent of natural variability which would lead to $\sigma_p = \sqrt{\sigma_m^2 + \delta^2}$.

  We have changed the text from "In practice σ_p is larger than σ_m to "In practice, σ_p has additional terms". We have now changed σ_p to $\sqrt{\sigma_m^2 + \delta^2}$.

- p.3, l. 21: the weights w^m must be normalized to sum to 1, as stated on p. 4, l. 15

  We have added that the weights should be normalised in this section.

- Fig. 1: I don't understand what is shown here: The weight w for simulated x and y values (as suggested by the caption), or the likelihood for simulated y values (as suggested by the y-axis). In the latter case, it would be more interesting to show the bivariate likelihood (a function of μ and σ) with contour lines. But isn't the likelihood a well-known concept that doesn't need illustration?

  The y-axis label should be weight w. We have updated the Figure accordingly and added some text in Section 2 for clarification.

- p. 4, l. 17: To sample from a mixture distribution, you cannot take the weighted average of draws from the mixture components. You have to select first randomly a component and then draw from that component, as in the procedure described at the bottom of p. 6.

  We remove this section, as it is not used anywhere for the computation.

- Section 2.2: I miss the information about the chosen prior distributions.

  We added that the default priors from MCMCpack was used in Sec 2.1.

- p. 6, algorithm at the bottom: This can be simplified because the conditional distribution of $(T+1)^{-1}\sum_{t=0}^{T} y_t^f$ given $(a_b^f, b_b^f, \sigma_p^f)$ is normal with mean $a_p^f + b_p \frac{T}{2}$ and standard deviation $\sigma_p^f/\sqrt{T+1}$. Hence one can directly simulate $(T+1)^{-1}\sum_{t=1}^{T} y_t^f$, there is no need to simulate first the y_t^f. One can even use that the conditional distribution of $T^{-1}\sum_{t=1}^{T} y_t^f$ given that m is the perfect model is a Gaussian mixture with means $a_{m,i}^f + b_{m,i}\frac{T}{2}$, standard deviation $\sigma_{m,i}^f/\sqrt{T}$ and equal weights $1/N$. So we can directly compute its density or the quantiles.

  Yes indeed if we are only interested in the mean differences, we can simulated directly from the distributions of the mean. The algorithm we give at the bottom is more general for dealing with general distributions other than the normal, and the same procedure follows when we are interested in changes in quantities other than the mean.