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.