Response letter to the reviewers of "A Markov chain method for weighting climate model ensembles".

Max Kulinich, Yanan Fan, Spiridon Penev, Jason P. Evan, Roman Olson

April 2021

1 General

Referees' original comments are in black font in Section 1 and Section 2, and our response is in blue font in Section 3.

2 Referee 1: Ben Sanderson sanderson@cerfacs.fr

Thanks to the authors for the detailed response, and for making efforts to address some of the concerns from the first submission. The addition of the leave-one-out tests, interdependency testing and parameter sensitivity study add robustness to the paper - and I'm now technically happy to see it published.

However, there are some points made in the original review which still require a caveat in the discussion. The sole consideration of globally integrated time-series as the outputs of a model ignores a lot of potentially useful data, and ultimately limits this analysis. For example - two models may have very similar global mean warming timeseries for the 20th Century, but with very different physical representations and regional climates. These two models should not, in practice, be considered to be highly related - but the method listed here would consider them to be so. The integrated response is a very low dimensional metric, and is unlikely to be informative about shared model assumptions. Other studies (Masson et al 2011, Sanderson 2017) have found that it is primarily complex spatial fields which are most informative on actual shared model components. I appreciate that implementing further analysis is out of scope, but this issue is a caveat to the current approach, and should be noted for consideration in future study.

Sanderson, B. M., Wehner, M., Knutti, R. (2017). Skill and independence weighting for multi-model assessments. Geoscientific Model Development, 10(6), 2379-2395.

Masson, David, and Reto Knutti. "Climate model genealogy." Geophysical Research Letters 38, no. 8 (2011).

3 Referee 2: Anonymous Referee 2

The paper is definitely improved. In particular, the cross-validation tests are a good and needed addition. And in terms of the comments I raised, the additional performance metrics listed in Section 2.5 (Trend Bias, Climatology monthly bias, Interannual variability, and Climatological monthly

RMSE) help to provide a much more salient and defensible measure of the method's performance in this context, at least as applied to the CMIP and NARCLiM data. With respect to the KMA case study, where the expanded set of performance metrics is not evaluated (since there is no seasonality in the heuristic), I understand that the idea is to show how the method performs in non-Gaussian or data-censored situations like the heat stress choice. I also understand that it's the same model ensemble being tested among the three methods, so it's a "fair fight" regardless of the physical meaning (or not) of the respective weights. But I still have a problem with presenting results from an ill-posed test without stating that these do not represent weights with real-world interpretability. So in reference to Section 3.3, there should be text added somewhere explicitly stating that a priori we would not expect these RMSE values to have any physical interpretation related to model skill. Or a short caveats section could be added to address these points. So while it's fine to use these case studies as toy problems to test the method's performance in different statistical contexts, the authors need to be clear that that's what this is.

4 Response to referees' comments

Thank you for your feedback and improvement suggestions. We acknowledge the remarks by both referees, that the physical interpretability of the weights is limited and can be improved in the future work by applying the MCE method on spatially distributed data.

We have now added the caveats in Section 2.1 (Data), lines 88-89:

"In this pilot study we use spatially averaged data, which limits physical interpretability of the model weights, but the method can be extended to spatially distributed data."

and in Section 2.2.3 (MCE method limitations), lines 173-174:

"Finally, as the MCE method does not consider spatial information, the resulting weights have limited physical interpretability. Extending the MCE method to utilize such information is a subject for future research."

and in Section 4 (Discussion), lines 291-293:

"However, as previous studies show (e.g. Masson and Knutti (2011); Sanderson et al. (2017)) and as discussed in Section 2.2.3, extending the MCE method to include spatial information would improve our ability to interpret the physical meaning of the resulting weights."