Geosci. Model Dev. Discuss.,
https://doi.org/10.5194/gmd-2020-253-RC1, 2020

Geoscientific Model Development Discussions

Interactive comment on "A Markov chain method for weighting climate model ensembles" *by* Max Kulinich et al.

Ben Sanderson (Referee)

sanderson@cerfacs.fr

Received and published: 2 November 2020

This study considers a novel application of Markov Chain methods to the problem of climate model weighting. The authors use a multi-model ensemble of climate model simulations, together with a range of different observation sources (global mean temperatures, regional averages and regional extreme temperature statistics). The novel "MCE" methodology is compared with two other approaches (a climatological weighted average score - Lambert and Boer 2001, AVE hereafter, and the ensemble transformation approach of Bishop and Abramowitz 2013, COE hereafter). The authors find desirable properties of their proposed approach in terms of out-of-sample skill and performance in non normally distributed quantities related to temperature extremes.

The approach shows promise, the potential for more robust error estimates in weights would be useful and the stochastic nature of the computation may offer additional benefits. However, these avenues are not explored in the current text and the authors have not yet fully addressed the basic requirements of an operational climate model weighting scheme (robust out of sample testing), and existing simpler approaches may outperform the approach presented here for the metrics considered (see below).

The introduction talks of the need to represent and consider model interdependency - but the actual study does not propose any mechanism for accounting for model interdependencies or common components in the weighting scheme. Example approaches for doing so are laid out in Sanderson (2017) and Lorenz (2018).

The study also generalizes existing literature by the two comparative cases considered (AVE and COE) as "linear" approaches, but there are more complex schemes in the literature which address issues not considered here, like optimal subselection (Sanderson 2015). In addition, given an RMSE or R^2 metric - methods which determine the global optimization of weighted scores (Herger 2018) are in theory unbeatable - so, although there may be tangential benefits to using MCE, it's unclear that the approach here could outperform a global weight optimization for these types of metric.

The paper attributes differences in weighting behavior between MCE, AVE and COE to differences in methodology, without sampling the subjective degrees of freedom in each of the approaches. As such, the observation that the MCE approach tends to rule out fewer models than COE is potentially a function of the chosen L and sigma parameters as well as the MCE approach itself. Any revision should present parameter sensitivities in the main text.

The paper also employs only a weak out of sample test, splitting the available data into a training and validation set. This is insufficient for the climate problem - where only the past is observable and the primary unknowns are climate projections in the future. Models with comparatively similar past trajectories might diverge significantly in

the future - and the validation scheme considered here (where random timesteps are withheld) does not capture this divergence. A stronger test is a perfect model study, where individual models are withheld from the ensemble and late century projected performance is considered.

The paper considers that, for a given output variable, that the climate models should be weighted only by their fidelity in producing that variable - but there is no clear reason why this should be the case. Climate projections are functions of the integrated climate system which determine global climate sensitivities and regional feedbacks. As such, fidelity in producing a given variable (such as a regional temperature timeseries) in the past is no guarantee of accuracy in the future (Sanderson 2012). Lorenz et al (2018), for example, discusses at length the relative utility of different variable types for constraining future model evolution.

Finally, the paper does not acknowledge the various reasons climate models may disagree - natural variability provides an absolute limit on the performance of an uninitialized climate model, and so some discussion is required on how different sources of error would impact the results in this case.

Given these issues, I suggest the following revisions:

1 - outline more clearly (and ideally quantify) the benefits of using MCE beyond absolute skill scores (where the method will be outperformed by construction by Herger 2018)

2 - present the parameter sensitivity of the method in the main text

3 - perform a rigorous leave-one-out test of weighting scheme performance in a perfect model projections of century-scale climate change

4 - Consider how to address model interdependency in this method (which would bias the leave-one-out test)- either by adaptation of existing approaches or otherwise

5 - Reconsider the relationship between variables used for weighting and those vari-

ables which are to be weighted (there is no reason to limit consideration in the weighting term to those variables which are themselves being weighted).

6 - consider the role of various sources of error and whether they should be represented within the scheme (natural variability, forcing errors, structural differences).

References:

Lorenz, R., Herger, N., Sedláček, J., Eyring, V., Fischer, E. M., & Knutti, R. (2018). Prospects and Caveats of Weighting Climate Models for Summer Maximum Temperature Projections Over North America. J. Geophys. Res. Atmos., 123(9), 4509–4526. doi: 10.1029/2017JD027992

Herger, N., Abramowitz, G., Knutti, R., Angélil, O., Lehmann, K., & Sanderson, B. M. (2018). Selecting a climate model subset to optimise key ensemble properties. Earth Syst. Dyn., 9(1), 135–151. doi: 10.5194/esd-9-135-2018

Bishop, C. H., & Abramowitz, G. (2013). Climate model dependence and the replicate Earth paradigm. Clim. Dyn., 41(3), 885–900. doi: 10.1007/s00382-012-1610-y

Lambert, S. J., & Boer, G. J. (2001). CMIP1 evaluation and intercomparison of coupled climate models. Clim. Dyn., 17(2), 83–106. doi: 10.1007/PL00013736

Sanderson, B. M., & Knutti, R. (2012). On the interpretation of constrained climate model ensembles. Geophys. Res. Lett., 39(16). doi: 10.1029/2012GL052665

Sanderson, B. M., Wehner, M., & Knutti, R. (2017). Skill and independence weighting for multi-model assessments. Geosci. Model Dev., 10(6), 2379–2395. doi: 10.5194/gmd-10-2379-2017

Sanderson, B. M., Knutti, R., & Caldwell, P. (2015). A Representative Democracy to Reduce Interdependency in a Multimodel Ensemble. J. Clim., 28(13), 5171–5194. doi: 10.1175/JCLI-D-14-00362.1

Interactive comment on Geosci. Model Dev. Discuss., https://doi.org/10.5194/gmd-2020-253,

2020.

C5