Geosci. Model Dev. Discuss.,
doi:10.5194/gmd-2016-285-RC1, 2017

Geoscientific Model Development Discussions

Interactive comment on "Skill and independence weighting for multi-model assessments" *by* Benjamin Sanderson et al.

Anonymous Referee #1

Received and published: 27 February 2017

The authors present a weighting scheme which, in their words, "considers both skill in the climatological performance of models over North America as well as the interdependency of models arising from common parameterizations or tuning practises". The two components of the weighting scheme are presented and developed separately, and are found to compensate to some extent for each other: models that are weighted higher for performance tend to be down weighted for replication.

The main limitation of the manuscript in my view is the primarily heuristic nature of the weighting schemes, which are at best partially justified. The introduction l73-78 sets out "two fundamental characteristics" of the scheme which are probably uncontroversial but which are not sufficient to narrow down the nature of the weighting scheme very much. I would however suggest that "relatively poor" would be more precise than the stated "demonstrably poor".

Taking performance weighting first, there is a substantial literature on this, albeit perhaps with limited results. Methods based on Bayesian Model Averaging (e.g. Hoeting et al 1999) have perhaps the strongest theoretical justification, but other approaches have also been presented (such as the "reliability ensemble averaging" approach of Giorgi and Means 2002). Olson et al 2016(a,b) present some recent applications of BMA to regional projections which seem highly relevant. I would ask the authors to consider whether their performance weights can be considered as Bayesian likelihoods, that is to say, is there an underlying statistical model which would result in this weighting scheme? If not, would it be worth changing to a more transparently presented and explained model, perhaps one which has been more widely applied and tested? Of course any statistical method will necessarily rest on a number of assumptions and simplifications which may not be easily justified, but at least these could be presented explicitly. For example, while the distance factor Dq is considered as a tunable factor here, there is also the use of an exponential function which defines the weights, for which no explanation is given. Even without changing the overall structure of the weighting function, increasing the exponent from its value of 2 would result in a sharper cliff-edge at which weights drop from 1 to 0, and alternatively a lower exponent would result in a much more gradual change with weights more similar across the models. Is there a particular reason for the choices made here?

Now moving on to the question of model independence, which here seems to be used to mean model output difference (as measured by a metric on output fields). The functional choice for the weighting again seems rather arbitrary. Since the goal of the parameter tuning seems to be to match the authors' beliefs that various models are replicated a particular numbers of times, is there a reason to use a function - which can only provide an approximation to this prior belief - rather than just use the authors' own judgements instead? For example a weight of 1/4 say could be applied to the GISS models directly, rather than trying to obtain a value close to this by tuning a single parameter. The choice of a fitted function seems to provide only a very thin veneer of objectivity to this subjective choice.

Despite these comments, I have no particular beef with the framework that has been presented - it does not look wrong or silly in any obvious way - but I also don't feel like I have been given any particular reason for using it. As outlined above, several of the numerous choices made don't appear to be that well justified. The tuning parameters do appear to have been selected sensibly, but this is only the last step after the creation of a structure that doesn't seem well supported.

A number of typos:

273-4 We briefly consider how the sensitivities of the method to different choices.

322 taylor/tailor

Fig 4 caption "1.5th percentile" really?

Giorgi, F., & Mearns, L. (2002). Calculation of average, uncertainty range, and reliability of regional climate changes from AOGCM simulations via the "reliability ensemble averaging"(REA) method, J. Climate 15(10), 1141–1158. Hoeting, J., Madigan, D., Raftery, A., & Volinsky, C. (1999). Bayesian model averaging: A tutorial. Statistical Science, 14(4), 382–401. 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 Olson, R., J. P. Evans, A. Di Luca and D. ArguÌĹeso (2016) The NARCliM project: model agreement and significance of climate projections. Climate Research, 69, 209-227.

Interactive comment on Geosci. Model Dev. Discuss., doi:10.5194/gmd-2016-285, 2016.