Geosci. Model Dev. Discuss.,
doi:10.5194/gmd-2016-285-AC1, 2017

Geoscientific
Model Development
Discussions

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

Benjamin Sanderson et al.

bsander@ucar.edu

Received and published: 4 April 2017

Thanks to the reviewer for their useful comments. We address each of the reviewer’s points below, and attach a revised version of the manuscript addressing the concerns:

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.

We agree that our weighting scheme is heuristic, but we also think that it could be potentially useful. Clearly, one could conceive of other weighting schemes which satisfy the desired characteristics laid out in the introduction, and we do not suggest that our proposed approach is the only possible or the best solution. We simply propose it as

a strategy, and would welcome other contributions from the community with alternative strategies which allowed for a simultaneous consideration of model skill and replication. Due to the lack of direct verification of climate projections, it is fundamentally impossible to decide what method or model is best, and choices in any such method are necessarily subjective to some extent. Different choices will also work better or worse for certain applications. We argue what is needed is not a justification of a method being correct or best, but traceability of what the choices were, and how they could impact the results.

I would however suggest that "relatively poor" would be more precise than the stated "demonstrably poor".

Changed as suggested

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?

We have added a section discussing BMA methods, and the REA method in the introduction. Notably, these methods are skill weights and do not easily allow for non-independent models. In BMA methods, a model’s projection is weighted by its posterior model probability, which is largely independent of other models in the archive (apart from in the weak sense that the probabilities in the archive as a whole are normalized). So - the technique doesn't satisfy one of our two requirements. This is true of REA as well - but REA also carries the rather unjustifiable assumption that a model which produces a projection which is an outlier from the rest of the ensemble should be downweighted, which would arguably increase the model interdependency issue rather than address it. REA also leads to overly narrow uncertainties in the presence of many models (Knutti et al. 2010 J. Climate).

We've added the following on the topic of interpretation of the scheme: "It should be noted that although our likelihood weighting function is empirical, the functional form satisfies in a simple way the required parameters of the weighting scheme. The structure of this functional form is not fundamental, it can simply be shown to have some desired features. The technique is presented in this paper in a form which maximises clarity and reproducibility, but its effect can be described in Bayesian language. The total model weight is the posterior likelihood of a given model representing truth. Each model's prior probability of representing truth is given by its independence weighting, and the likelihood function is defined for the multivariate dataset using an assumed Gaussian likelihood profile in a space defined by the the sum of the normalized RMSE differences over all variables between each model and the observations."

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?

We've tried to make it more clear in this version that the scheme is not intended to be *the* answer to weighting models. Yes, the functional form imposes some structural

limits on the weights one would obtain. By using a different power exponent, one could create a more or less polarized distinction between 'good' and 'bad' models - we could sample this dimension as another sensitivity study, but as you suggest, one could propose an infinite number of potential weighting functions, and we simply propose one which has some desirable characteristics, and we sample some useful parameters to sample a range of behavior - we claim no deeper interpretation than that. Given that Dq is chosen such that the method produces reliable uncertainties in the perfect model test, it is likely that a different exponent would lead to a different Dq but the overall mean and uncertainty would not change substantially.

However, there is precedent for using a Gaussian formulation for a likelihood function, we do not argue that our weighting scheme is not heuristic - our only requirement was to have a smooth, well behaved function which allocates maximum weight to a distance of zero, and no weight to a distance of infinity, without differentiating between two models which have distances « Dq. This actually leaves a rather limited set of choices for an appropriate functional form, for which a Gaussian structure is the simplest.

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.

Our argument for the representation of model interdependence is exactly that prior judgements of model interdependence are not required, because they are not always known - and this may be increasingly true in the future. As the reviewer points out, if the only problem was to downweight models from the same institution which are known to be similar, the problem would be simple - either giving each of these models a fractional weight, or by taking only one version of institution's model.

However, in some cases, there are model interdependencies which cross institutions (take NorESM and CESM, or ACCESS and HadGEM). Unless the researcher knows about these in advance - they would miss them, whereas our method is data-driven, and if inter-depedencies are evident from the data, they are de facto considered. Interdependence will also vary on the quantity considered, two models may show similar behaviour in sea ice if they share the sea ice model, but differ more in other parts where components are not shared, or where other uncertainties dominate. We demonstrate our selection of the independence parameter using known cases, because in these cases - we know approximately what the answer should be. The point is then that the method can be generalised to cases where we don't know a priori the degree to which two models are related.

The constraints of this application are such that we were obliged to produce a single set of weights - but for the methodology in general, it allows for models to be assessed for interdependency conditional on certain outputs of the model which are relevant to the question in hand.

C2 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.

We hope that the above arguments help justify our approach, we propose a structure which a) satisfies our original requirements (downweight replication, upweight skill) in a framework which b) allows sufficient free parameters to tune for increased skill

without risking an overly calibrated result which might increase the risk of the truth lying outside the weighted ensemble distribution, and c) produces a single sets of weights for each model to be used in climate impact assessments based on a method easy to understand and implement by non-statisticians. Note that this paper is written to address a narrowly defined set of boundary conditions required by the author team of the Climate Science Special Report - specifically for a single set of weights which could be readily applied to a wide variety of projections. The method is not presented as fundamental, rather it is presented as a model which is defensibly fit for this particular purpose of dealing with a multi model ensemble in a National Climate assessment..

A number of typos: 273-4 We briefly consider how the sensitivities of the method to different choices. Corrected, thanks.

322 taylor/tailor Corrected.

Fig 4 caption "1.5th percentile" really?

Sorry -this was a version mixup. Now reworded to be consistent with the definition of D_u in Figure 3.

Please also note the supplement to this comment:
http://www.geosci-model-dev-discuss.net/gmd-2016-285/gmd-2016-285-AC1-supplement.pdf

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