Review of Manuscript

An information-theoretic approach to obtain ensemble averages from Earth system models'

By C. Sierra and E. Munoz

Dear Editor,

I have reviewed the manuscript. My conclusions and comments are as follows:

1. Scope

The article is within the scope of GMD.

2. Summary

In their manuscript, the authors propose a method for multi-model inference, in particular for the task of estimating quantities of interest as single values from ensembles of earth systems models (ESMs), with the particular challenges of the models i) being parameterized in non-MLE ways, ii) having a large but unknown number of parameters and iii) high dimensionality. The approach seeks to go beyond simple ensemble averaging by including a measure of model performance in the weighted combination. The method is based on information concepts in general and work by Akaike, Burnham and Anderson in particular, based on the key insight that the log-likelihood of a model trained on a set of observation data is equivalent to its expected Kullback-Leibler divergence, which is a universal measure of distance between a reference pdf (here: the observations) and a model pdf thereof. The log-likelihood therefore can be used to derive suitable weights when taking the average of the outputs of an ensemble of ESMs. The challenge is the estimation of the likelihoods. Under certain conditions they can be estimated by the variance of the residuals (= difference of model and reference truth). For ESM output these conditions are not met as the authors themselves state, but they use this approach for the lack of alternatives. The method then consists of calculating, grid-by-grid (to acknowledge spatial patterns of model performance) for each model the variance of the residuals against some observational product, selecting the best among the models and calculating the performance difference (delta) of the remaining models against this reference. Following Akaike, the exponential of delta is then used to obtain weights for each model, which then allows calculating a single-valued model ensemble average and related variances as a measure of estimation uncertainty. The authors then demonstrate their method at the example of 9 ESMs and two related observational products. A comparison to simple equal-weight ensemble averaging shows that i) they may differ substantially in the presence of poor-performing models and ii) that uncertainty estimates are generally much lower for the weighted average approach.

3. Evaluation

Overall, the manuscript provides a solution for an important task – averaging of ESM ensembles - that is both well-grounded in theory and easy to apply. The (sometimes strong) assumptions are clearly stated, and the method overall is well presented in terms of derivation and examples.

A few questions remain:

1) Calculation of models weights. In eqs (7) – (15), I take it that index "i" goes over all models except the best performing model ("m"), which serves as the reference. Does that mean the best performing model will not be included in the weighted average? Please clarify.
2) Patterns of model performance in space and time (see also line 180): The authors resolve spatial patterns of model performance by calculating model weights grid-by-grid over all points in time. This inherently assumes temporal invariance of relative model performance, which clearly is not the case. Therefore I wonder if it would not be more appropriate to

derive model residuals (and deltas) from space-time regions rather than time-regions alone. Please comment.

3) In Sect. 4, simple averaging is used as a benchmark to compare the weighted averaging proposed by the authors. While e.g. Fig. 5 clearly show differences among the methods, it is not clear if the weighted average really provides the better (in terms of smaller disagreement from observations) estimate that the simple average. I am quite sure this will be the case, but please add and discuss the related numbers.

My overall recommendation is to publish after these minor points have been suitably addressed.

Yours sincerely,

Uwe Ehret