Geosci. Model Dev. Discuss.,
https://doi.org/10.5194/gmd-2017-317-RC2, 2018

Geoscientific
Model Development
Discussions
Open Access

Interactive comment on "Cluster-based ensemble means for climate model intercomparison" *by* Richard Hyde et al.

M.G. Schultz (Referee)

m.schultz@fz-juelich.de

Received and published: 13 February 2018

The manuscript by Hyde et al. presents a novel application of an efficient clustering method to the analysis of multi-model ensemble results. As the authors rightly point out, there have been relatively few applications of clustering to atmospheric data yet, and the clustering method allows for some deeper insights into large datasets, which do not reveal their secrets if only simple traditional methods, i.e. multi-model means are applied. The paper is well written and the details of the method and results are well presented. I also strongly support the idea to explore such methods in the context of climate data with an objective to improve the analyses of multi-model results and model evaluation in general. However, I must raise two fundamental concerns which I have with this study and which prevent me from recommending this article for publication.

These concerns are explained below. Because of the fundamental nature of these concerns I refrain from making detailed technical comments. I strongly encourage the authors to revise their manuscript and put the study on a solid foundation. As a community we clearly need this kind of method development! But we also have to make sure that we are not violating good practices when it comes to scrutinizing model results in confrontation with observations.

Concern #1: From the outset, the aims of the study are formulated as trying to reduce the bias between a multi-model ensemble and observations – in this case a climatology of tropospheric column ozone retrieved from a satellite instrument. For example, the abstract states "As a proof of concept, we show the proposed clustering technique can offer improvement in terms of reducing the absolute bias between the MMMs and observations." The problem I have with this is that the study objective therefore mixes two very different things: a clustering method, which has the sole purpose of classifying data into groups of somehow similar properties, and a model-data comparison, which should try to objectively describe how well or how poorly the models agree with observations and put this in relation to estimates how well the models could agree with the observations in theory, i.e. answering the question what degree of certainty a model-data comparison can give us. By mixing the objectives of these two steps and using the reduced bias as a proof of concept for applying cluster analysis to ensemble model results, the study loses the objectivity that is required for a meaningful model evaluation.

Concern #2: I never heard of a requirement to know the "predicted truth" in a cluster analysis, and this concept actually frightens me, because it suggests that the analysis is performed in such a way that the desired results are obtained by tweaking the method until it fits. First, there is no principle need to enforce selection of one cluster after the grouping is done – this actually defeats the purpose of clustering! Second, if the authors do wish to pick one (or two?) clusters from the resulting groups, then they should apply accepted, robust statistical methods for this. In the introduction a suggestion was made to select the cluster with most members. This would be an objective criterion. In terms of model evaluation it would convey the message that one selected a certain estimate of the truth from the majority of models. This must be clearly separated from the value that is represented by that cluster. In fact, it could be very valuable to define measures that describe how many clusters are formed from the model ensemble in each region or throughout time. For example, one could use the ratio of the number of members in the most populated cluster over the number of members in the next populated cluster – if this value is large, then this means that most models agree with each other; if the ratio approaches 1, the models differ a lot among themselves, and this indicates that some process is not well understood or well represented in all models. This conclusion can be reached regardless of any comparison with observations.

To illustrate these issues, let me present a simple example for a single point in the model evaluation (e.g. a single grid column value of a single month). Assume that the retrieval yields a TCO value of 30 DU. Five models produce values of 50 DU, and five models produce values of 10 DU. The traditional MMM will generate an average of 30 DU from these and claim a perfect match between model and observation. But the MMM hides the fact that none of the models was even close to the measurement, and it is by pure chance that the average of all models agrees with the data (had we had only nine models, we would have found a bias). Without additional analysis, we thus believe that our models are perfect. The cluster technique represents one way of taking the analysis a step further: in this example, it will identify two groups of model results, one at 10 DU and one at 50 DU. This is it. That is all that the cluster analysis does, and this is the information that should be used to describe the quality of the model results (see example of the ratio measure above). What the authors now suggest is to use a "predicted truth" to identify one of the clusters. In this hypothetical example, this would either fail (because both clusters are equally far away from the "truth"), or one of the clusters would be randomly selected. In the latter case, one might actually obtain an "evaluation" result which looks reasonable on the surface, because the large bias of 20 DU would be identified. However, with a slight modification of the example, i.e.

by adding a few models, say 2 with a result of 30 DU to the ensemble, the proposed method would identify these models as a third cluster and suggest perfect agreement with the observations. Clearly, this result is not meaningful in the context of evaluating the "quality" of the models or of the model ensemble.

Another issue which is not reflected in the manuscript is the error of the observations. A recent study by Gaudel eta al. (Elementa – Science of the Anthropocene, under review) shows substantial deviations between different TCO retrievals. Again, in order to make a robust statement about the quality of a model, one needs to know what a deviation between the model result and an observed value actually means. Would a perfect model even be expected to exactly reproduce the observation? (see also Solazzo et al., 2016/2017 on AQMEII evaluation)

Finally, I would like to emphasize again that I very much welcome the introduction of clustering as an analysis technique for model results, and that I see great potential for this method. As the authors describe, the technique is robust (or should I say graceful) against typical errors in the data sampling step of the model evaluation, i.e. if a feature is shifted by one grid box. The grouping of data within the model ensemble indeed offers several new insights to the "behavior" of models. My sole criticism is: what the authors claim is a "proof of concept" actually disproves the concept, because it forces the concept of clustering to achieve something it is not meant to do.

Interactive comment on Geosci. Model Dev. Discuss., https://doi.org/10.5194/gmd-2017-317, 2018.