Earth Syst. Dynam. Discuss.,
https://doi.org/10.5194/esd-2019-69-RC1, 2020

Earth System Dynamics
Discussions

Interactive comment on "A weighting scheme to incorporate large ensembles in multi-model ensemble projections" *by* Anna L. Merrifield et al.

Anonymous Referee #1

Received and published: 10 January 2020

Review of "A weighting scheme to incorporate large ensembles in multi-model ensemble projections" by Merrifield et al.

In this paper, the authors describe the extension of a weighting scheme for multi-model climate projections described in previous works to incorporate single model initial condition large ensembles (SMILE). This weighting scheme uses a performance metric, based on the similarity of a simulation with observations and an independence metric, based on the similarity between simulations. Several properties of two variables (surface air temperature and sea level pressure) in the present climate are used to measure similarity. The authors intend to demonstrate the applicability and the usefulness of this weighting scheme with SMILEs, focusing on surface air temperature change over Northern Europe and the Mediterranean. They also discuss different properties

of the weights and some practical issues that may arise in such applications.

The subject of the paper is interesting and important, and there are some interesting analyses in this paper. It is well written and generally easy to follow. But I also think that the use of the proposed weighting scheme with SMILEs raises fundamental questions that are not addressed. As the incorporation of SMILEs in the weighting scheme is the novelty of the paper compared to previous works, these issues must be properly dealt with before the publication of the paper could be considered.

I am not sure that the authors can address these issues properly, as they are really intrinsic to the chosen approach, but I want to give them the opportunity to prove me wrong. I therefore recommend major revisions to the paper, but I may still recommend rejection of the paper at the next round.

Major comments

The notion of "independence" is perfectly defined in statistics and probability theory, but it is very ill defined when applied to climate models (which is not really acknowledged and discussed by the authors). In this paper, as in previous works, two models are considered more or less independent depending on the similarity of their results. Two models are considered "weakly" independent if their results are very similar and "strongly" independent if their results are very different. This is a hypothesis, and it should be discussed. The results of two "independent models" cannot be similar? Two independent models cannot converge towards the truth (if the models are close to the truth they will also be close from each other)? Overall, to my opinion, this hypothesis can make sense when dealing with multiple different models, and in any case there is no perfect theoretical and practical way to characterize model independence.

But I'm really bothered with this approach when dealing with members from the same model (only differing by initial conditions). I think that the attempt to use this weighting scheme with SMILEs illustrates some difficulties of the definition of independence in terms of similarity.

The members of a SMILE are independent in the statistical sense of the term, the only sense of independence that is well defined. But they are not independent in the approach proposed by the authors, and they can be more or less "independent" according the similarity of their results. For me, it is very problematic. If you roll a dice two times, you don't decide that two outcomes are "more independent" if you get a 4 and a 6 than if you get two 3. But it is basically what is done in the proposed method with SMILEs.

As an illustration of this issue: Imagine the particular case where we only have a single SMILE, and that we are interested by the distribution. Using the weighting scheme described in this paper is not correct in this case, right? We know that the SMILES members are independent and that each member should receive the same weight. It is what is done is all the studies based on a single SMILE. But the weighting scheme described in the paper would give different weights to different members. I think that the weighting scheme proposed by the authors (any weighting scheme) should hold seamlessly in a particular case like this one.

-Giving different performance weights to different members of the same climate model is also problematic, at a fundamental level, I think. The skill is intrinsic to the model, and not specific to a member of the model (once the memory due to initial conditions has disappeared). Whether a particular member of a SMILE is closer to the observations than another is purely accidental and says absolutely nothing on the realism of this particular member in the future climate.

-The baseline approach to which the weighting scheme is compared in this paper consists in giving an equal weight to all the members of the multi-member, multi-model ensemble (independently of the existence of other members of the same model in the ensemble). Obviously, it is a very bad approach, and nobody would do that, I think.

If we consider the models as independent and equally skilful, SMILE members can be easily incorporated in a multi-model ensemble, as it has been done for years, by

giving a weight to each member of a given model inversely proportional to the number of members of this model in the full ensemble. This approach is perfectly justified from a statistical standpoint (within the hypotheses made). (i) I think that the authors should use this approach as a baseline, to which they can compare their weighting scheme, and show the results obtained with this approach for example in Figure 3. (ii) Logically, the weights of an appropriate weighting scheme should tend towards the ones described above when the "hypothesis" of inter-model dependence and unequal realism is relaxed, I think. It is not the case with the weighting scheme described in the paper.

-I disagree with the interpretation of the results of dynamical adjustment in the paper. It is not possible to extract the "forced trend", even the "estimated forced trend" or the "radiatively-forced trend" with dynamical adjustment. Dynamical adjustment only allows separating the part of the trend that is due to large-scale atmospheric circulation from the part of the trend that is not due to large-scale atmospheric circulation. The "part of the trend that is due to atmospheric circulation" is not a correct estimation of the impact of internal variability, except in some particular cases. The variations in atmospheric circulation indeed can be forced, they are not necessarily of internal origin. There are quite a few papers on the detection and attribution of anthropogenic influences on large-scale atmospheric circulation, and there is a clear forced component (in the real sense of the term) in future circulation changes in many models. For this reason, the "part of the trend that is not due to large-scale atmospheric circulation" should not be named "forced trend", even "estimated forced trend". Additionally, it can bear the imprint of internal oceanic dynamics.

It is mainly a vocabulary issue here, as the interpretation of the results of dynamical adjustment does no really matter for the results discussed in the paper. Still, it is important to be correct.

Minor comments

l37. Parameterized processes are not the only reason for model uncertainty, I think. The dynamical cores can also be important in that context.

l53. As said in the major comments, dynamical adjustment cannot be used to quantify the impact of internal variability on climate variables. It can only be used to estimate the part of variability that is not driven by large-scale atmospheric circulation. It is completely different.

l55-56. You mean single "model" initial condition large ensemble and not single "member", right?

l85, data section I think it would be more logical to introduce the climate simulations (e.g. Table 1 etc.) before describing their result (Figure 1 etc.)

l96. ERA20C should not be used as observational reference for temperature. Only SLP and winds are assimilated in ERA20C, which leads to a sub-optimal representation of temperature variability. Not surprisingly, issues in regional temperature trends and low frequency variation exist in ERA20C. There are much better datasets to use for temperature. There is no need to use SLP and TAS from the same dataset to "assure consistency". Use the best dataset for each variable: normally, good observations from different sources are consistent. I also think that multiple observational datasets should be used in order to assess the impact of observational uncertainties.

l144. "that adds independent information". What is meant exactly by "independent information" (or "new" information, at some places)? It should be discussed, from a theoretical point of view.

l176. What "fit for purpose" means obviously depends on the purpose. I think it would be useful to state the purpose very precisely at this point (even if it can be inferred from other parts of the paper).

l185. I don't really understand how the RMSE distances are computed. You say that they are computed at each point before area averaging. RMSEs are not computed over

space but time? How do you compute the RMSE for a climatology at a given point? Please give the equations, it will be clearer.

l192-194. It is a reasonable idea when you consider two different models, but not when you consider two members of the same models. And in this paper two members of the same model are dealt with in the same way as two different models.

l200. There is no i (and ii) in Figure 2a. Please add the complete numbering of the sub-figures.

l207-213. The fact that the "estimated" forced trends are so different between members of the same model clearly shows that one should not talk of forced trends for the results of dynamical adjustment, preceded or not by "estimated". But I agree that independently of its name, it can be an interesting performance metric.

l214. "Internal variability": no, not necessarily (see major comments).

l228. Can you clarify what is meant by ""fair""?

l235 and Figure 3. I'm missing something: I don't understand how the weighted distributions (box-and-whiskers plots) are obtained, based on the weighting scheme described in the paper. It is not directly straightforward I think. Is it a parametric distribution, using the weighted variances and means and a Gaussian hypothesis? It does not seem to be the case as the whiskers are not symmetrical. Please explain how the percentiles are computed when using the weighting scheme.

l245. It would be interesting to add the results of the "classical" weighting scheme generally used when mixing SMILEs and multiple models (see major comments), that makes the hypothesis that the models are independent and equally skilful. It is a much better starting point for the comparison. Nobody in his right mind would add 200 members of the same climate model to the CMIP5 ensemble and compute the distribution without some basic weighting, right?

l254-255. This is rather obvious: see the previous comment.

l281-282. What criterion do you use to judge that the weighting is suitable? What is a suitable weighting scheme? It should be better discussed.

l285-320. I don't think that this analysis is that interesting. More important (and interesting) analyses are in Appendix.

l456-457. I don't see a test of the sensitivity to "sigma s" in Figure B2. You mean Figure B3? Should you not describe Figure B2 first?

Interactive comment on Earth Syst. Dynam. Discuss., https://doi.org/10.5194/esd-2019-69, 2019.