Earth Syst. Dynam. Discuss.,
https://doi.org/10.5194/esd-2019-70-RC2, 2020

Earth System Dynamics
Discussions

Interactive comment on "How large does a large ensemble need to be?" *by* Sebastian Milinski et al.

Anonymous Referee #2

Received and published: 10 February 2020

This manuscript is investigating the optimal number of members from single-model ensemble. To do so, they are suggesting a conceptual recipe which should provide the optimal number of members. They subdivide their investigation into three sections where they: 1) quantify the forced signal, 2) the internal variability and 3) the change in internal variability in order to provide the optimal number of members for each question using the MPI-Grand Ensemble. The study is showing some interesting results and is worth publishing. However, the writing could be improved (still some internal notes). Since the paper do not really fulfill its promises in a convincing way (providing the size of a large ensemble), the focus of the paper should be rethought. I will therefore suggest accepting the manuscript but only after a major revision. I hope that my comment will help the authors to improve the quality of their paper.

Major comments: Some of the results of this study are interesting and deserve to be published. However, I think the title is not representing the paper, since there is no

concrete conclusion about the number of members, the question remains still an open question which depends on where (regions), what (which variables), who (models) and when (periods), which is already shown in previous study about internal variability. I would suggest changing the whole structure of the paper.

The introduction does not match the rest of the paper. For example, there are three interesting questions at the end of the introduction, but then the paper since to be structured otherwise while suggesting that the recipe for estimating the ensemble size will be followed... It would greatly improve the clarity of the manuscript if the questions were explicitly addressed in the next sections (as subsection). I would suggest transferring this whole discussion of Sect.2 (but removing its main conclusion (see below)) into an Apendix section.

In Sect.2, the authors are investigating at which size the reduction of error is due to the increase of ensemble members and not to the resampling error (or the limits between those two). I fully appreciate the need for such an approach for your studies, however, I do not agree with your conclusion of lines 14 to 16. It may be true for the max ensemble size of 20, but not for the others...It is, at least, highly disputable. I do not see, and therefore not convinced, that the diverging point is $\sim$50% of the maximum ensemble size. I think that this is the weakest point of the manuscript, but quite important. However, I do not think that this is a deal breaker, since most of the text can me readjust (for example page 7, line 29; page 9 line 9; etc. . .). The following line seems to bring news proofs, but unfortunately I couldn't convince myself otherwise since the text was not clear and accompanied by still some internal notes shielding doubts about the figure (see captions of Fig.3). I would also suggest getting rid of the whole part of page 5 line 17 (or just mention it).

As written, the authors directly proposed a recipe for estimating the ensemble size, which (and I am sorry to say it) look like it is drawn from a hat. I do not understand why (and where) this comes up and why it is presented in that section. As presented, the recipe is stating the obvious and is presented as the center issues of the manuscript,

but is not anyway. I would first specifically answered the tree questions and then maybe proposed a recipe that could be tested in a small paragraph just before the conclusion. In that sense, I think that the manuscript is showing some interesting results, but not fulfilling his promises...

One more general comment, I often had the impression that the solution when choosing the size of the ensemble was to select subsample members of a large ensemble, which for me did not make sense since the whole ensemble should be used (otherwise, why running it?).

Minor comments:

Page 5 line 3-13: This whole paragraph was a bit obscure to me and could be clearer. It needed more details and terms should be explicitly mentioned (and maybe shown on Fig. 2 directly as an example) in the text, such as “the error convergence” in “the resampling start to dominate the error convergence”.

Page 7 line 16-20: Those few sentences are quite confusing, could you please add more explanations? In figure 4 a–c, the expected RMSE for each grid point is shown for ensemble sizes of 3, 5, 10, and 50 members. The RMSE is computed as the mean difference between 100 samples (of what of each ensemble size (like in Sect 2, 100 samples of sets of 3,5,10 and 50 members)? If yes, why not have chosen 1000 random samples as in Sect2) and the 100-member mean (which is the whole ensemble, right?). When the ensemble mean is based on just 3 members (so which one? The ensemble-mean of the 100 samples of set of 3 members?), the expected error in the estimated forced response is large over land regions, in particular in the northern hemisphere.

Page 7 line 25-27: ...the acceptable error is 0.1°C... do you mean the number of members needed to restrain the RSME to 0.1°C? If yes, please keep RSME instead of error. Otherwise, please clarify.

The manuscript should have a section explaining the MPI-LA set-up, so the paper can

stand by himself.

Please specify somewhere what is GSAT and Nino3.4

Page 2, line 3-5: I would explicitly mention the term signal-to-noise ratio in that paragraph.

Page 2, line 9-10-11 “If the signal...present-day conditions” I suggest getting rid of that line. I do not like this statement imply that there is enough members to quantify IV, so why would you look only one member. It is irrelevant.

Page 2, line 16: ..of the large regional variability...

Page 2 line 16 to 20: I think this is not correctly cited. One the reason that Li and Ilyina (2018) required so many members are most likely due to the week(er) overall forced signals from RCP4.5. As written, it seems that the two studies are comparable (Li and Ilyina (2018) and Steinman et al. (2015)), but their differences should be explicitly mentioned.

Page 2 line 24-28: Please reformulate, not clear. For example, they analyze the polar cortex but concluded about the lower latitude...

Page 2, line 33-34: Could you elaborate a little on that?

Page 5 Figure2: I would change to yellow color for another one...I do not see it well when printed...

Interactive comment on Earth Syst. Dynam. Discuss., https://doi.org/10.5194/esd-2019-70, 2019.