Earth Syst. Dynam. Discuss.,
https://doi.org/10.5194/esd-2019-70-RC1, 2020

Earth System Dynamics
Discussions

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

Anonymous Referee #1

Received and published: 22 January 2020

General comments:

In this paper, the authors study the impact of ensemble size on the estimation of different climate statistics using the MPI Grand Ensemble and a pre-industrial control simulation. They analyze the statistical error associated with different quantities as estimated from ensembles of varying sizes, such as the forced response in global surface air temperature, as well as in regional temperature and precipitation. They also assessed the required ensemble size for estimating ENSO variability, linear warming/cooling trends, and changes in internal variability for Arctic sea ice.

Overall, I think this study is highly relevant for guiding users on required ensemble sizes related to different applications, as well as to provide useful insights to climate modellers in the context of the production of upcoming large ensembles. The paper is generally well written and results are original, interesting and worth publishing. How-

ever, there are a few sections that would need to be revisited. For instance, I think a short additional section providing a basic description of the "Data and Methods" would make the paper much easier to understand. In addition, I have some concerns about the selected methods, whose details and implications should be discussed in more details. Finally, the conclusions should better put the original findings into a wider context, especially by comparing with other existing studies (as cited in the introduction) that also have estimated required ensemble sizes.

My main concern about the methodology used in this paper is the exaggerated importance of what the authors call the "resampling problem" (RP). If the aim of this paper is to provide robust estimates of the required ensemble size for different applications (as stated several times in the paper), the importance given to the RP is an obstacle to this goal. The RP is actually an artifact of the selected strategy of resampling the large ensemble without replacement and has profound impacts on the interpretation of the results. With this approach, the question of "How large does a large ensemble need to be?" becomes highly conditional to the size of the ensemble at hand, especially when 50% (here loosely estimated) of the maximum ensemble size is exceeded. If the author would replace their strategy by resampling WITH replacement, the RP would also become a limitation at some point, but for much larger sample sizes (probably even above than the actual maximum ensemble size of 200 members).

The previous comment mainly applies to the results based on MPI-GE, but the issue of the resampling strategy also applies to the results based on the pre-industrial control simulation. For this part, the authors do the resampling by generating synthetic members obtained by splitting the pre-industrial control into overlapping segments (e.g. 50 or 100 years). However, three resampling strategies were actually possible, without any explicit mention in the document: 1) overlapping segments (suffering from the serial dependence of the windows), 2) non-overlapping segments (leading to only 20 members from the 2000-year time series), and 3) random year selection to generate synthetic segments (either with or without replacement). Implications and interpreta-

tion of these possible approaches should be discussed in order to support the decision of selecting which one is better to apply in which context.

Specific comments:

1. p1l7-8 "First, we determine how much of an available ensemble size is interpretable without a substantial impact of resampling ensemble members" The RP is a limitation of the current approach and could be attenuated by changing the resampling approach. I don't think this issue should be mentioned in the abstract, and other similar comments in the paper should be revisited according to the above general comment on RP.
2. P2L13: "to to"
3. P2L22-24: I think the reference to Pausata et al. (2015) is not correct. Maybe another paper from the same author is cited ?
4. P1L24 "make use of a model's pre-industrial control run where possible." This is not that clear in the paper why sometimes we use MPI-GE and otherwise the preindustrial run. This should be clarified in the new Data and Methods section and supported by additional explanations regarding the resampling method.
5. P3 A basic description of data and methods is missing:
    - It would be welcome to provide a short description of the simulations used in this study, that is the control run and MPI-GE. Especially, it should be noted somewhere what RCP is used, and to mention the initialization method that was applied to produce MPI-GE.
    - It should be more clear why the analysis is sometimes applied to MPI-GE or to the preindustrial runs. The resampling methods used in the study should also be discussed.

6. P3L4-5 I would suggest rephrasing "When using a smaller ensemble, sampling uncertainty may be misinterpreted as a forced change in ENSO or a robust difference between two models." to something like: "When using a smaller ensemble, sampling uncertainty may lead to false detection of a forced change in ENSO or a robust difference between two models."
7. P3L8-10 The point that the required ensemble depends on the model (i.e. the magnitude of internal variability) is important and should be discussed further in conclusion.
8. P3L13 "Therefore we differentiate three types of questions that encompass the specific questions that are commonly addressed with a large ensemble and show examples for each type of question" – This sentence needs to be simplified.
9. P3L19-24 I think this section on the resampling problem should rather begin by justifying why one should in the first place resample to estimate the required ensemble size. Then, to describe the different possible resampling approaches in order to justify which one to use in which context (and according to either MPI-GE or the preindustrial runs).
10. P4L3 and P4L12: The choice of resampling without replacement is had hoc and this choice should have been discussed earlier.
11. P4L12-14 "At some point, the 1000 random subsamples are not independent anymore because they share many of the randomly drawn members from the full ensemble." I would highly suggest the authors to compare the number of possible ensembles that can be formed without and with replacement. The second approach offers much more degrees of freedom.
12. Fig. 1: Choose another color for the full envelope (1 member) as it is the same (light blue) as for the 50-member ensemble. Adjust the legend accordingly. A

version of this figure generated by resampling with replacement would add a non-zero uncertainty on the 200-member average.

13. P5L5-6 "For a smaller number of realisations in the full ensemble, the resampling starts to dominate the error convergence earlier than in a much larger ensemble." See general comment on the RP.

14. P5l11013 "The sample size for which the RMSE estimate in a smaller maximum ensemble size starts to diverge from the RMSE estimate based on a larger maximum ensemble size determines the threshold of where resampling substantially affects the error convergence." Here the 50% limit is estimated rather loosely. Comparing versions "with" and "without" replacement of Fig. 2 would give a good indication of where this limit could be. However, I'm not sure this is a very useful result since the alternative approach of resampling with replacement would attenuate the RP, at least for ensemble sizes smaller or equal to 200.

15. Fig. 3:

    - The caption should obviously be re-written and clarified.
    - Results would be more clear by inverting the order of plotting, that is red to light blue from top to bottom.
    - How can a standard deviation have negative values ?

16. P6L1-2 Are the subsamples overlapping or completely independent ? It seems they are overlapping, which might lead to an underestimation of the standard deviation of the distribution due to the serial dependence of the time windows. Generating 50-year periods by randomly resampling individual years could allow to circumvent this issue. The selection of the best approach for this problem should be discussed in the new Data and Methods section.

17. Fig. 4 and 5: Why not using all 200 members with replacement here ? This could allow to get rid of the saturation over the continents. In addition, it would be useful to know exactly over which period these maps are computed.

18. P7L21 "[. . .] while larger ensemble sizes are affected by resampling and therefore not shown." See general comment on the RP.

19. P7L27-28 "Beyond 50 members, the resampling problem inhibits reliable estimates of the sufficient ensemble size." See general comment on the RP.

20. P11L12-13 "The advantage of this approach, in contrast to the examples for the forced response, is that the required ensemble size can be estimated for any model without needing a large ensemble to be available." Yes – but is this approach (of splitting in overlapping windows) give similar results to a resampling over MPI-GE ? This should be verified by the authors and clarified in the methods section.

21. P11L18 (fig. 8) Same as previous comment about the overlapping windows.

22. p14L9-13 See general comment on the RP.

23. p15l17-18 It would be good to recall some examples from the introduction where other studies have assessed required ensembles for different applications, and compare with the results presented in the current paper.

24. Conclusion: Put important findings in the context of other studies cited in literature. Also discuss that ensemble sizes would likely be different with other models with different magnitude of internal variability.

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