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.

Thank you for your thoughtful review and suggestions for improving the manuscript. Our main objective is to suggest a conceptual recipe to estimate the required ensemble size, explain the reasoning for using the recipe, and discuss possible caveats in the interpretation of the results. We do provide the required ensemble sizes for several applications in the MPI-GE. These applications are meant to demonstrate how the method can be applied. The required ensemble sizes we find are likely dependent on the model used (its magnitude of internal variability and the relative magnitude of the investigated signal). We acknowledge that we need to ascertain that these objectives are clearly stated in the revised manuscript, so that our results meet the reader's expectations.

We apologise for not removing the internal notes in the caption of figure 3.

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.

Our intention was not to provide a conclusion about the numbers of ensemble members needed, because such a number does indeed depend on the specific question asked (region, variable) and the climate model used. Instead, we propose a generic method that can be used to estimate the required ensemble size for any given question and any climate model. The method can either be applied to an existing large ensemble to test if it is the right tool for the question at hand, but it can also be applied to a pre-industrial control run to estimate the required ensemble size before running a new large ensemble. In the revised manuscript, we will elaborate more on the option to use the pre-industrial control run of a model. (see our replies to reviewer 1 comments 4,5,7,16,20)

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.

We have realised that the resampling problem is mentioned too early and without the appropriate context. We will restructure the sections to provide more background before mentioning the resampling problem. The updated structure will be:

- *Introduction*
- *Model description*
- *The basic approach for estimating the required ensemble size (forced signal in GSAT and regional temperature and precipitation)*
    - *the resampling problem*
- *recipe*
- *applying the recipe to various typical problems*
- *…*

The applications of the recipe follow the three questions outlined in the introduction. We believe that the updated structure will be much easier to follow.

The applications follow the three questions in the introduction (currently sections 4.1 to 4.3)

1)response to external forcing: GSAT, regional temperature and precipitation, linear warming trend, cooling after volcanic eruption
2) quantify internal variability: ENSO and temperature variability over land
3) identify a forced change in variability: Arctic sea ice area

The updated structure of the revised manuscript will make it easier to link the examples to the three questions. We will also add a paragraph to the conclusions where we will link the examples to the respective question.

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 ~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).

The approach we take to resampling will be updated and explained in more detail in the revised manuscript to take the suggestions by reviewer 1 into account. We will also include more theoretical background to our choice of 50%, which is currently based on an empirical approach.

We apologise for the internal note in the caption of figure 3. The figure itself has been updated, but we did not update the figure caption. The figure caption should read:

> *PDF of ensemble-averaged Niño3.4 standard deviations possible in the MPI-GE pre-industrial control simulation for subsampling ensembles ranging from 50 to 1000 members (shown as different colors) for smaller ensemble sizes. Each PDF is shown relative to the corresponding ensemble mean value. We use the last 1000 years of the 2000 year control run to calculate the ranges. The Niño3.4 standard deviation is calculated over 50 year periods. The PDFs are created by resampling the control simulation 1000 times. For each PDF the entirety of the 1000 years are used (i.e. the blue 500 member pdf is the mean of 2 500 members PDFs).*

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…

We will solve this problem by the updated structure (see comment above). We use the forced response in historical GSAT as an example to illustrate how the question from the title can be approached and how resampling can become an issue. We will then use the examples of regional temperature and precipitation to demonstrate how different variables, regions, and acceptable errors influence the required ensemble size.

This will then be followed by the recipe and all remaining examples, all of which illustrate how the recipe can be applied to the three types of question mentioned in the introduction.

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?).

Here we take advantage of an existing very large 200-member ensemble. The advantage of using this ensemble is that the full ensemble is likely very close to the truth for many applications. In the case of GSAT, the 200-member mean provides a good reference for the true forced response in this model. We can then ask: how large is the error when using the ensemble mean of a smaller ensemble to estimate the model's forced response? We answer this question by subsampling the full ensemble.

When using the MPI-GE, one would certainly use all available members. In the context of this study, the 200 members from the MPI-GE allow us to explore how well our recipe works for other typical ensemble sizes of large ensembles (e.g. figure 2). Finally, we demonstrate how a pre-industrial control simulation can be used to estimate the required ensemble size for a given model and question. This approach can be used to determine which models from the CMIP5 or CMIP6 archive provide a sufficient number of realisations, or it can be used to determine the ensemble size required for a variety of questions before running a new large ensemble.

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”.

Thank you, we have realised that the necessary context for this paragraph is only mentioned later in the manuscript. We will improve this when restructuring the manuscript.

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.

We will extend the explanation. The RMSE for a grid point and ensemble size (e.g. figure 4a for 3 members) essentially contains the same information as the solid black line in figure 2 at an ensemble size of 3 (of course after recomputing this for the regional instead of global temperature). Note that the maps in figures 4 and 5 only represent the expected RMSE and not the uncertainty interval (shading in figure 2).

The sample size of 100 instead of 1000 was selected because this analysis is computationally expensive. We will consider reducing the sample size in figure 2 to be more consistent

Note that we updated figures 4 and 5 to use all 200 members instead of 100. (see response to reviewer 1, comment 17)

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.

Yes, acceptable error refers to RMSE in this context. We will clarify this in the revised manuscript.

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

We will add a section describing the model and experiments used after the introduction.

Please specify somewhere what is GSAT and Nino3.4

Thank you, we will add the definitions.

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

Thank you for this suggestion. In this paragraph, we want to introduce the concept of averaging over many ensemble members to eliminate the noise from internal variability. This approach is not directly evaluating the ratio between the signal from the forced response to the noise from internal variability.

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.

Here, we should have explicitly used the term signal-to-noise. We look at the question: how many members do we need to be certain that a signal exists. In the case where a single trajectory clearly emerges from the noise of internal variability, the presence of a signal can be detected in a single realisation. For example, a single RCP8.5 realisation is clearly sufficient to conclude that the end of the 21st century is warmer than pre-industrial conditions. In the introduction, we wanted to mention that not all applications require a large ensemble.

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

Thank you, this will be changed.

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.

Thank you, we will extend the description of these papers. Li and Ilyina investigate carbon uptake in the southern ocean. In this case, the signal-to-noise ratio is small because of the large variability in the southern ocean. Steinmann et al. investigate a region and quantity that is less variable and has a larger forced signal, therefore the signal-to-noise ratio is large and a small number of ensemble members is required. The only similarity between the two studies is that they try to identify a forced change. We choose these examples to illustrate the large differences in ensemble size requirements when the investigated quantity and region are different.

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

We will extend the description.

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

We will extend this paragraph. Here, we introduce a common strategy: instead of using a large ensemble, a long pre-industrial control run with no change in the external forcing is used to quantify internal variability. This estimate of internal variability can then be used to quantify the uncertainty due to internal variability in simulations where the external forcing is changing. The underlying assumption is that internal variability does not change when the external forcing is changing. While this is true for some quantities, it does not hold for other quantities such as the Arctic sea ice area as we show in section 4.3.

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

Thank you for the suggestion. We will revisit the choice of colors for this figure.