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

Thank you for your thoughtful review and suggestions for improving the manuscript. We are happy that you are interested in our results and appreciate your suggestions for improving the manuscript. Please find our replies to your comments below.

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. However, 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.

We will add a short section describing the model and simulations used.

However, we would like to keep the description of the method connected to the applications. The primary goal of this study is to develop a method that can be applied to estimate the required ensemble size in any given context. The applications of this method are meant to demonstrate our reasoning for the chosen method and illustrate caveats in the interpretation.

In the conclusion section of our revised manuscript, we will discuss our results in the context of the previous studies.

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).

Thank you for this comment that has stimulated us to rethink how we address the resampling problem, and how we present it in the manuscript.

We have realised that the current structure is not ideal. The resampling problem is mentioned very prominently, but too early so that the relevant context is missing. In the revised manuscript, we intend to start with the forced response in GSAT example (fig. 1) and the associated estimate of the required ensemble size (fig. 2). Building on this, we will then integrate the discussion of the resampling problem. We will try to limit this to the extent that is necessary to follow the argument and provide additional information in the supplementary information.

Regarding the sampling approach, we made a conscious decision to resample without replacement. The reasoning behind this is that by subsampling for example 5 out of the 200 members, we try to imitate a situation where we only produced 5 members with our model. These could be any 5 out of the 200 members we actually have.
In the case where we resample with replacement, a single member could appear more than once in this sample of 5. We think this is unlikely to happen in reality because that would mean that two members produced by a climate model are (nearly) bit-identical despite a different initialisation. By allowing replacement, we would arrive at an arguably too conservative estimate of the required ensemble size.

However, we do note that sampling with replacement would be an obvious solution to the problem we raise from a purely statistical perspective. We will therefore include a better reasoning for our choice in the revised manuscript.
Our current reasoning for interpreting only up to 50% of the maximum available ensemble size is currently based on an empirical assessment of this threshold. We are working on an additional analytical reasoning to support the choice of this threshold. This will be limited to the effects of resampling without replacement for the mean (i.e. forced response). The 'standard error of the mean' is closely connected to the problem at hand and can be estimated. For higher order moments, this will be less straightforward.

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

Yes, the resampling does indeed have implications for the analysis based on the pre-industrial control simulation. We will elaborate on this in the revised manuscript and provide a reasoning for the strategy we used.

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.

As outlined above, we will make substantial changes to the treatment and presentation of the resampling problem in the manuscript.

However, we do think that the resampling problem is an important caveat that needs to be considered when determining the required ensemble size. Previous studies have concluded that X of N ensemble members are sufficient to detect a signal, with X/N being around 0.6-0.8. Therefore, we felt that this potential caveat should be highlighted in the abstract.

2. P2L13: "to to"

Noted, thank you.

3. P2L22-24: I think the reference to Pausata et al. (2015) is not correct. Maybe another paper from the same author is cited ?

Yes, this is indeed the wrong reference. We will change this to the correct reference: Pausata, F. S. R., Grini, A., Caballero, R., Hannachi, A. & Seland, Ø. High-latitude volcanic eruptions in the Norwegian Earth System Model: the effect of different initial conditions and of the ensemble size. Tellus B: Chemical and Physical Meteorology 67, 26728–17 (2015).

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.

Agreed, we will make sure to explain in more detail under which conditions the control run can be used, and how this can be done in practice.

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.

We will add a short section explaining the design of the MPI-GE and the runs used in this study. (pre-industrial control, historical, and 1% CO_2)

- 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.

Our objective is to use the preindustrial control run whenever possible because this simulation is readily available for every CMIP5/6 model, while a large ensemble is not. However, some of the applications require a different type of simulation where the external forcing is changing over time (increasing CO2, volcanic eruptions). In the revised manuscript, we will make this choice more clear. Our recommendation is to use a preindustrial control simulation when possible. As noted in response to comment 4, we will include more detailed recommendations for how the pre-industrial control simulation can be used for specific questions.

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."

Thank you, we will follow your suggestion.

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.

Thank you for this suggestion. It seems that this point was not clear enough and we will make sure to emphasise this more. Our motivation for introducing a method rather than recommended ensemble sizes is based on this point: analysis of a different model might result in a different required ensemble size. Therefore we suggest that this analysis is repeated with every model before using it, rather than assuming that the required ensemble size derived from the MPI-ESM in this study is valid for all models.

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.

Agreed. If this sentence is still in the manuscript after rewriting, we will make sure to simplify it.

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).

As mentioned above, we intend to 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*
- *…*

10. P4L3 and P4L12: The choice of resampling without replacement is had hoc and this choice should have been discussed earlier.

Yes, as stated above we will justify our resampling approach and discuss the alternative approach with replacement.

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.

As described in the responses to comments 4 and 5, we intend to extend the discussion of the use of the control run, including the sampling strategy for different types of question.

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.

We will change the color as suggested. Yes, resampling with replacement indeed adds a non-zero uncertainty for the 200-member average and increases the uncertainty for most other averages. As described above, we believe that this estimate would be too conservative. We choose the approach that provides a less conservative uncertainty estimate, but introduces the resampling problem. We will discuss this in more detail in the revised manuscript.

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.

Noted.

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.

We will support the empirically estimated 50% threshold with a more rigorous derivation of the sample size at which resampling affects the conclusion. As described above, this is more straightforward for the mean than for higher order moments, which is why we relied on the empirical approach in the submitted version of this manuscript.

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 ?

Apologies for including an old caption in the submitted manuscript. The figure was updated, but not the caption.The 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).*

We will invert the order of plotting as suggested.

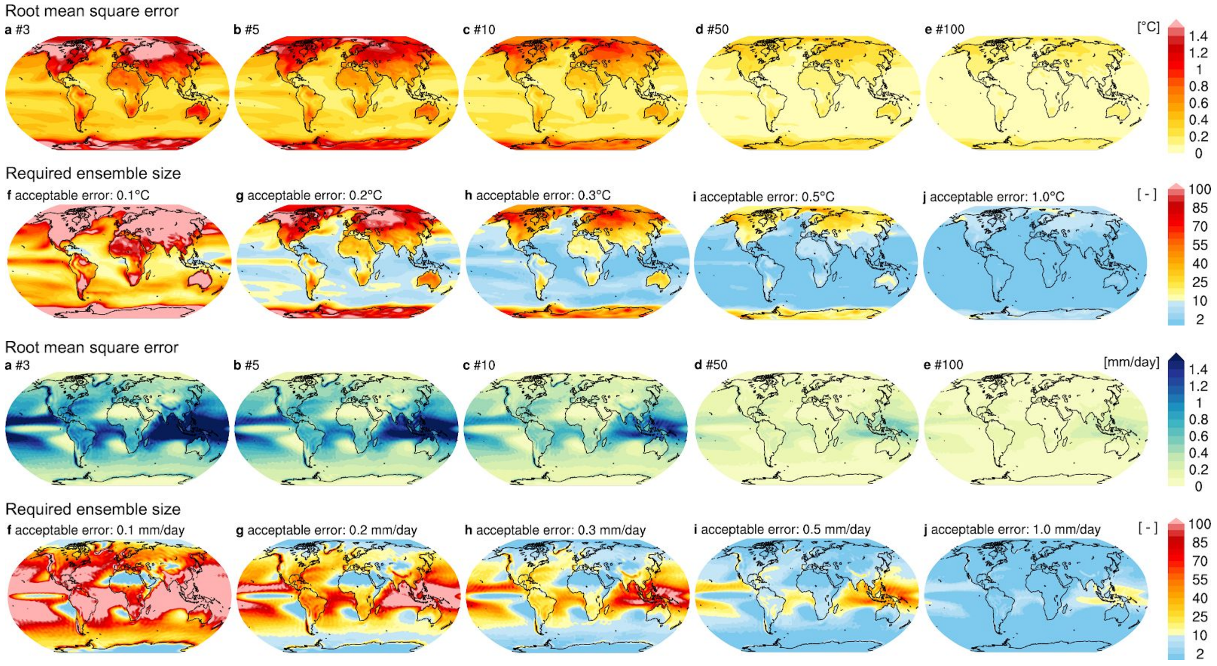
The standard deviation is relative to the mean value, this is now clarified in the caption.

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.

The subsamples are overlapping. We will explain this in more detail in the revised manuscript. In the case where we quantify ENSO variability, random resampling would not be representative of real ensemble members because ENSO has a timescale larger than 1 year. We selected consecutive years to retain the temporal characteristics of ENSO.

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.

We did repeat the analysis with all 200 members (figures below to replace figure 4 and 5). The period is the full length of the historical simulations (1850–2005). We will make this more clear in the revised text. This analysis is an extension of the analysis in figure 1 and 2. For each grid point, we show the expected RMSE at a specific ensemble size, which is equivalent to the value of the solid black line in figure 2 for that ensemble size (computed for a grid point instead of globally).


Root mean square error
a #3
b #5
c #10
d #50
e #100
[°C]
1.4
1.2
1
0.8
0.6
0.4
0.2
0
Required ensemble size
f acceptable error: 0.1°C
g acceptable error: 0.2°C
h acceptable error: 0.3°C
i acceptable error: 0.5°C
j acceptable error: 1.0°C
[-]
100
85
70
55
40
25
10
2
Root mean square error
a #3
b #5
c #10
d #50
e #100
[mm/day]
1.4
1.2
1
0.8
0.6
0.4
0.2
0
Required ensemble size
f acceptable error: 0.1 mm/day
g acceptable error: 0.2 mm/day
h acceptable error: 0.3 mm/day
i acceptable error: 0.5 mm/day
j acceptable error: 1.0 mm/day
[-]
100
85
70
55
40
25
10
2

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

Noted.

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

Noted.

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.

Yes, sampling over several years in the control run and sampling over members in the MPI-GE does provide results, under the condition that the forcing in the MPI-GE has not changed the distribution. We will clarify this in the revised manuscript.

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

Noted.

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

Noted.

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.

We will make sure that the discussion revisits questions raised in the introduction. However, we do not want to reproduce the analysis of previous studies in detail because we do not want to put too much focus on the actual numbers that we find in this study. Our main objective is to present a generic method that can be used to determine the required ensemble size, explain caveats, and how it can be applied in practice.

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.

We agree that the results are possibly highly model dependent. This is an important point and we will emphasise this in the revised conclusion.