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 added 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.

We have updated the structure of the paper to introduce the method and the generalised recipe in section 3.

We have added a short paragraph to the conclusions section pointing out similarities to previous studies. However, the example applications in this study are not identical to the applications from previous work. Furthermore, model differences could contribute to different ensemble size requirements. Therefore, we use this section to emphasise that there is no

ensemble size that is sufficient for every model or application, but encourage our readers to estimate the required ensemble specifically for the combination of application and model(s) used.

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 have reduced the complexity of the GSAT example to focus on the steps to estimate the required ensemble size for this application. The 'recipe' is integrated in this section and directly builds on the GSAT example. The resampling problem is mentioned in a short paragraph. The detailed discussion of issues related to sampling, such as resampling and sampling with or without replacement, has been moved to a new appendix.

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. This can also be seen in the new figure A1:

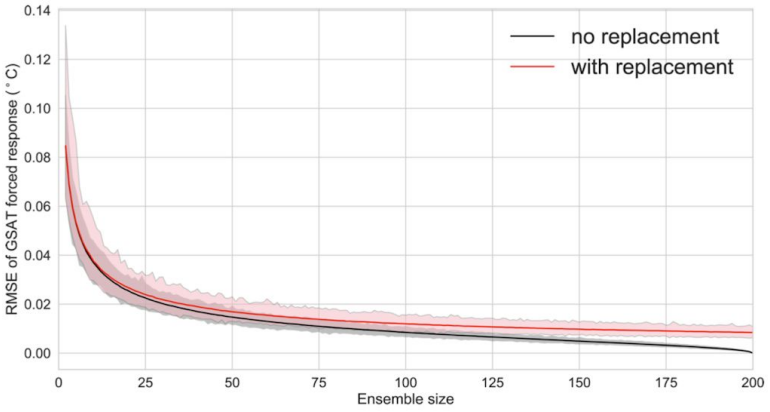


Figure A1. Sampling with or without replacement affects the error estimate and therefore the estimate for the required ensemble size. The black line shows the mean RMSE for GSAT for ensemble sizes from 2 to 200. The reference is the 200-member mean from figure 1 and the RMSE is computed for all 1000 samples. The shaded area shows the range of RMSE values for individual samples, the solid line shows the mean RMSE. The red line and shading show the RMSE for ensemble sizes from 2 to 200, but samples are generated by allowing sampling with replacement.

However, we do note that sampling with replacement would be an obvious solution to the problem we raise from a purely statistical perspective. We have therefore explained our choice to sample without replacement in appendix section A1.

Our reasoning for interpreting only up to 50% of the maximum available ensemble size is based on an empirical assessment of this threshold. We have considered an analytical derivation but concluded that this is too complex for the most applications in this study beyond the trivial case where the sampling uncertainty scales with 1/sqrt(n), which is only the case when estimating the mean of a sample generated by a stationary process (e.g. the mean of a pre-industrial control simulation). For higher order moments like the standard deviation or more complex error metrics applied here such as the RMSE, trends, or differences between time periods, estimating the theoretical sampling uncertainty is more complex. Our objective here is to introduce a simple framework that can be modified for a wide range of applications and that can easily be applied.

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 have added a subsection within 4.2 to explain our sampling choice and alternative options and hope this will provide additional value to our readers.

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.

We have removed the RP from the abstract and restructured the paper to put less emphasis on the RP.

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.

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 changed 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.

We have added an additional subsection within 4.2 to discuss the motivation for sampling from the pre-industrial control run and different sampling approaches.

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 added a short model section explaining the design of the MPI-GE and the runs used in this study. (pre-industrial control, CMIP5 historical, and 1% CO_2).

The method is central to this study, but we think that the approach is easier to understand when introduced with an example. This is what we do in the revised section 3.

- 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 added a new subsection 4.2.1 to discuss sampling in the pre-industrial control simulation.

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 followed 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. The final two paragraphs in the conclusions now address this in more detail.

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. We changed this to: "Therefore we differentiate three types of questions that represent questions typically addressed with large ensembles:"

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).

We moved most of the material on resampling to the appendix and start with a simple example to introduce and explain our method. The resampling is only briefly mentioned as a caveat in the main text.

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

Resampling with and without replacement is now discussed in appendix A1.

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.

It is true that sampling with replacement offers more degrees of freedom. However, this also produces synthetic ensembles that would be treated with suspicion when encountered in an existing large ensemble: an ensemble that contains two (or more) completely identical realisations would raise doubts about the correct initialisation rather than being treated as an ensemble containing fully independent members. We also discuss this in appendix A1.

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 changed the color as suggested. A version of figure 2 where we compare sampling with and without replacement is now shown and discussed in appendix A1.

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. We have made substantial changes to the structure in response to the 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.

We have moved this discussion to appendix A. 50% is not meant as a strict limit, but as a reminder that ensemble sizes around and beyond this point should be interpreted more carefully.

Our reason for sampling without replacement is explained in A1. (also see response to comment 11)

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 now reads:

> *Figure A3. 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 updated the colors to be consistent with figure A2.

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

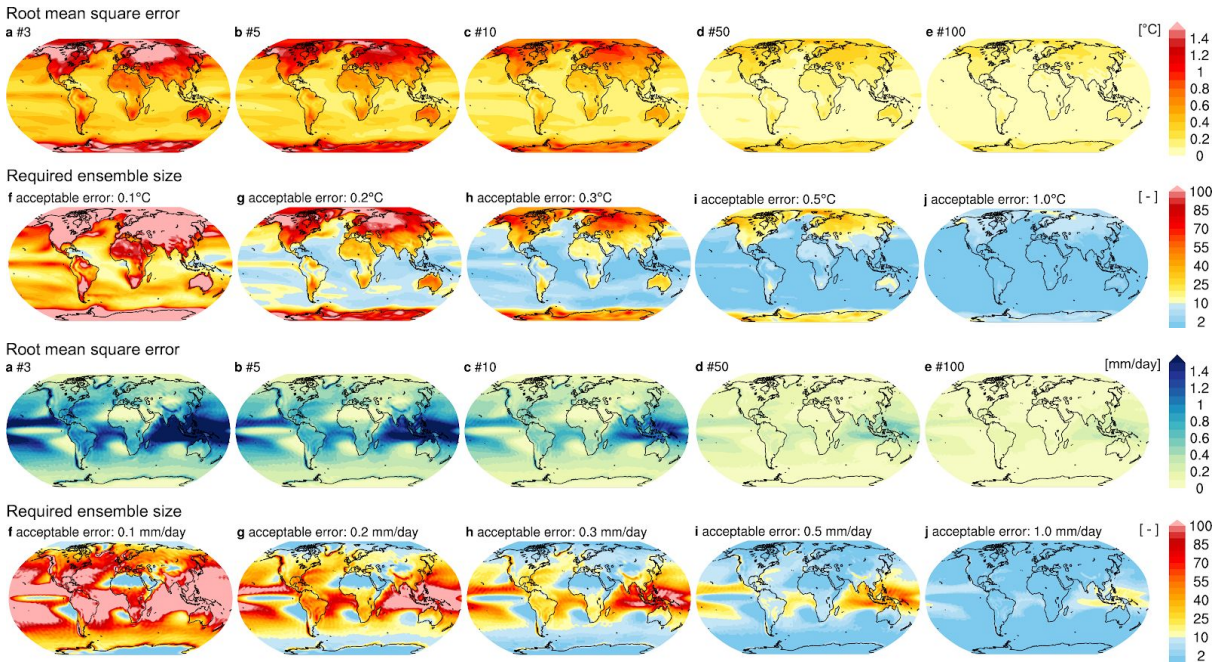
16. P6L1-2 Are the subsamples overlapping or completely independent ? It seems they are overlapping, which might lead to an underestimation of the standard deviation of the distribution due to the serial dependence of the time windows. Generating 50-year periods by randomly resampling individual years 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 explain this in more detail in the revised manuscript (also section 4.2.1). In the case where we quantify ENSO variability, random resampling would not be representative of real ensemble members because ENSO has a timescale longer 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 replace figure 4 and 5, now 3 and 4). The period is the full length of the historical simulations (1850–2005). 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).



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

Noted. See reply to 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.

Noted. See reply to 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.

Yes, sampling over several years in the control run and sampling over members in the MPI-GE does provide the same results, under the condition that the forcing in the MPI-GE has not changed the distribution. This direct comparison is therefore only possible in the first years of the historical simulations with negligible changes in GHG concentrations and prior to volcanic eruptions. Computing statistics across the ensemble does have both advantages and disadvantages. We mention this in section 4.2 and added references to previous studies (also in this special issue) that explore the use of the ensemble dimension in more detail.

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

Noted. We address the sampling options in 4.2.1

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

Noted. See reply to 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.

We address this in the final two paragraphs of the revised conclusion:

> *The examples in this study demonstrate that for some applications ensemble sizes around 5 members are sufficient while other applications require ensemble sizes well above 100 members. In section \ref{sec_introduction} we introduced several estimates for required ensemble sizes from the literature. While most of the applications from previous studies are not directly comparable to the examples we use here, the large range of required ensemble sizes emphasizes the need to systematically estimate the required ensemble size for each individual application. Furthermore, the required ensemble size may be model dependent. Therefore, the numbers derived in this and previous studies should only be used as approximate estimates and supported by a systematic model- and application-specific estimate following the approach outlined in this study.*

> *The information about the sufficient ensemble size is not only crucial when choosing or designing a large ensemble, but can also help to identify applications where a small number of ensemble members is sufficient and thereby inform the design of multi-model intercomparison studies. The method introduced in this study can add to the robustness of results both from single model large ensembles and multi-model large ensembles.*

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 have emphasised this more in the revised manuscript. Also see response to comment 23.

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). In the revised manuscript we have emphasised these objectives so that our results meet the readers' expectations.

We made substantial changes to the structure and hope that you will find that the focus of the paper is now clear.

We apologise for not removing the internal notes in the caption of figure 3 (now figure A3).

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 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 final two paragraphs of the conclusions now emphasise that the specific numbers provided in this and previous studies depend on the application and model. Therefore, we emphasise the method in this paper rather than specific suggestions for ensemble sizes.

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 restructured the paper to provide more background before mentioning the resampling problem. The updated structure is:

- *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 is much easier to follow.

We have added additional pointers in the introduction to emphasise that the examples in 4.1 to 4.3 follow the three questions:

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

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 is now explained in more detail in the revised manuscript to take the suggestions by reviewer 1 into account.

We apologise for the internal note in the caption of figure 3 (now A3. The figure itself has been updated, but we did not update the figure caption. The figure caption now reads:

> *Figure A3. 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 have updated the structure and moved most of the discussion of the resampling problem to the appendix. We now 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. This simple example is used to explain how we arrive at the generalised recipe.

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 was only mentioned later in the manuscript. We have updated the structure and moved this content to appendix A.

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 have added a sentence stating that the analysis for the maps is essentially the same as figure 2, but applied to each grid point individually. The RMSE for a grid point and ensemble size (e.g. figure 3a for 3 members) 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 3 and 4 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.

Note that we updated figures 3 and 4 (previously 4 and 5) and now 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 have added this information to the sentence:

> *"If the acceptable error (RMSE) is 0.1..."*

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

We added section 2 "Model" to explain the setup and simulations used.

Please specify somewhere what is GSAT and Nino3.4

Thank you, we have added the definition and boundaries for the Nino3.4 box.

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 could 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.

Being able to identify a signal in a single realisation has implications for detectability in observations, which are the single realisation we have for the real world.

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

Thank you, we have updated this.

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 extended 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 sufficient. 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…

In this case, 'lower latitudes' was lower in relation to the core of the maximum positive wind anomaly, i.e. still high latitudes in a global sense. We now use a formulation that avoids this ambiguity and is closer to the original text by Bittner et al.:

> *"...7 members are sufficient at the southward flank of the maximum positive wind anomaly, but up to 40 members are necessary to identify a response at high northern latitudes."*

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

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.

We have extended this paragraph to emphasise potential problems with this approach.

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 changed yellow to orange.

Dear editor,

We have resubmitted our manuscript 'How large does a large ensemble need to be?' for consideration in Earth System Dynamics. BWe have made substantial changes to the manuscript and have addressed all concerns raised by the two reviewers.

We note that the discussion of the resampling problem appeared out of context and with insufficient explanations in the previous manuscript. Based on the reviewer comments we have restructured the manuscript as follows:

- The introduction that includes the three types of questions that will be considered is followed by a short 'section 2: model' introducing the model simulations used in this study.
- 'section 3: A simple method to estimate the required ensemble size' uses a simple example to explain the steps that are summarised in a recipe at the end of this section. The previous figure 2 has been simplified. The resampling problem is briefly mentioned as a subsection in this section. Most of the material about resampling has been moved to a new appendix section that explains our choice of the sampling methods, including a discussion of alternative approaches. In the appendix, we also show a side-by-side comparison of sampling with and without replacement.
- Section 4 is then addressing the three questions mentioned in the introduction. This connection is emphasised by a short introduction paragraph in section 4.

By moving material that was not directly part of the main storyline to the supplementary information, we expect that our reasoning for the recipe and the connection between the introduction and the results in section 4 should be much clearer now.

Please note that in addition to the reviewer comments we have modified figures 2, 5, and 6 and start sampling consistently with a single realisation instead of 2 (as used previously in figure 2) to also explicitly include the error when using just one ensemble member.

We have updated our responses to the reviewer comments to reflect the changes implemented in the revised manuscript.

Kind regards,

Sebastian Milinski
Nicola Maher
Dirk Olonscheck

How large does a large ensemble need to be?

Sebastian Milinski[1], Nicola Maher[1], and Dirk Olonscheck[1]

[1]Max Planck Institute for Meteorology, Hamburg, Germany

Correspondence: Sebastian Milinski (sebastian.milinski@mpimet.mpg.de)

Abstract. Initial-condition large ensembles with ensemble sizes ranging from 30 to 100 members have become a commonly used tool to quantify the forced response and internal variability in various components of the climate system. However, there is no consensus on the ideal or even sufficient ensemble size for a large ensemble. Here, we introduce an objective method to estimate the required ensemble size that can be applied to any given application and demonstrate its use on the examples of global mean ~~surface~~ near-surface surface air temperature, local ~~surface~~ temperature and precipitation, and variability in the ENSO region and central ~~America.~~ United States for the Max Planck Institute Grand Ensemble (MPI-GE). Estimating the required ensemble size is relevant for designing or choosing a large ensemble, but also for designing targeted sensitivity experiments with a model. Where possible, we base our estimate of the required ensemble size on the pre-industrial control simulation, which is available for every model. ~~First, we determine how much of an available ensemble size is interpretable without a substantial impact of resampling ensemble members. Then, we~~ We show that more ensemble members are needed to quantify variability than the forced response, with the largest ensemble sizes needed to detect changes in internal variability itself. Finally, we highlight that the required ensemble size depends on both the acceptable error to the user and the studied quantity.

1 Introduction

Single model initial-condition large ensembles (SMILEs) are a valuable tool to cleanly separate a model's forced response from internal variability and to improve our understanding of the observed trajectory of the climate system in the past, and its projected future evolution (Zelle et al., 2005; Deser et al., 2012a; Rodgers et al., 2015; Kay et al., 2015; Maher et al., 2019; Branstator and Selten, 2009; von Känel et al., 2017; Kirchmeier-Young et al., 2017; Frankignoul et al., 2017; Stolpe et al., 2018).

The ensemble ~~size~~ sizes currently available for individual global coupled climate models largely differs. The single-model ensembles within the Coupled Model Intercomparison Project Phase 5 and 6 (CMIP5, CMIP6) are on the low end of available ensemble sizes, typically ranging from three to ten ensemble members for a model, with the majority of models having only one member available. In contrast, computationally expensive ~~single model initial-condition large ensembles~~ SMILEs position themselves on the top end of available ensemble sizes, providing up to 200 ensemble members for a single model and forcing scenario. While studies are beginning to compare multiple ~~large ensembles~~ SMILEs (Maher et al., 2018; Deser et al., 2019), there is still no clear consensus on how large such an ensemble should be for any given application.

We here introduce a new framework to objectively estimate the required ensemble size for different types of questions and make use of a model's pre-industrial control ~~run~~ simulation where possible. ~~This approach~~ Using the pre-industrial control simulation allows us to estimate the required ensemble size for a specific model even if no large ensemble is available~~for the model. This~~ . The objective approach can also help to allocate resources more efficiently (Ferro et al., 2012) and to inform the modelling community how many ensemble members are desirable for CMIP models.

One of the most common applications of ~~single-model large ensembles~~ SMILEs is to separate a forced response due to anthropogenic global warming from the noise of internal variability. In a sufficiently large ensemble the ensemble mean can be used as an estimator for the forced response (Frankcombe et al., 2018). This approach has been applied to study various regions and quantities.

On a global scale, Deser et al. (2012b) investigate the forced response in temperature and precipitation. They found that around 10 ensemble members are sufficient to detect changes in the global mean land temperature in the next decade, while more than 40 ensemble members are required to detect changes in precipitation. When going further into the future when the signal becomes larger, they find that fewer members are sufficient to detect a forced change. If the signal is large enough, a single ensemble member is sufficient to detect a significant change compared to ~~present day~~ present-day conditions. This happens when the trajectory of the single member emerges from the range of internal variability for present day conditions.

On both global and regional scales, Olonscheck and Notz (2017) used both the CMIP5 multi-model ensemble and the MPI-GE to conclude that multiple small ensembles from different models are useful to ~~to~~ quantify the response uncertainty across different models.

While a forced response in global mean temperature only requires a relatively small ensembles size, forced changes on a smaller regional scale can be more difficult to detect because of the larger variability. Li and Ilyina (2018) investigated the ocean carbon sink and found that up to 79 ensemble members are required to isolate a forced decadal trend in the RCP4.5 scenario in the ~~southern ocean. Steinman et al. (2015) on the other hand~~ Southern Ocean, a region with large internal variability. Steinman et al. (2015) quantify the forced response in North Atlantic temperature and argue that for this region, more than four ensemble members are required for a robust estimate of the forced ~~signal from a single-model ensemble~~ response from a SMILE. Although the objective of the two studies is similar—identifying a forced response—the required ensemble size is very different, indicating that different regions and quantities can have very different requirements on the ensemble size.

In addition to investigating forced changes to anthropogenic forcing, large ensembles also allow an investigation of forced responses to other external forcings such as volcanic eruptions. For regional temperature changes, ~~Pausata et al. (2015a)~~ Pausata et al. (2015b) find that up to 40 ensemble members are necessary for a robust detection of a temperature response after a volcanic eruption. Bittner et al. (2016) investigate changes in atmospheric ~~dynamics~~ circulation after a volcanic eruption. They analyse the polar vortex and find that the required ensemble size to detect changes in the zonal wind after a strong volcanic eruption depends on the latitude: 7 members are sufficient ~~in lower latitudes~~at the southward flank of the maximum positive wind anomaly, but up to 40 members are necessary to identify a response at high northern latitudes. ~~However, their target is~~ The ratio of the signal to the noise from internal variability is different in different regions because both the signal but also the internal variability are different. The target of Bittner et al. (2016) was to detect a change in the circulation that

is different from zero, but not to quantify it. Quantifying the magnitude of the forced response may require an even larger ensemble size for this application.

Large ensembles have also been used to quantify internal variability, with some studies arguing that very large ensemble sizes are necessary: Daron and Stainforth (2013) conclude that an ensemble with several hundred members is required to characterise a model's climate, while Drótos et al. (2017) demonstrate that 100 members are sufficient. On the other hand, some studies argue that the pre-industrial control ~~run~~ simulation is sufficient to quantify internal variability and no large ensemble is required. Thompson et al. (2015) argue that the pre-industrial control ~~run~~ simulation can be used to provide a robust estimate of internal variability and represent future internal variability, implying that a single ensemble member for each model may be sufficient. However, this approach only works if the internal variability does not ~~changes~~ change over time. In addition, a single realisation for a transient scenario does not allow a clean separation of the forced response and internal variability, even if the magnitude of the internal variability is quantified using a pre-industrial control simulation.

ENSO variability and its potential changes under global warming have been investigated in several studies and widely different future changes have been identified (Stevenson et al., 2012; Bellenger et al., 2013; Christensen et al., 2013). Maher et al. (2018) investigate ENSO variability and its potential changes under global warming in several large ensembles. They find that at least 30 ensemble members are required for a robust estimate of ENSO variability. When using a smaller ensemble, sampling uncertainty may ~~be misinterpreted as~~ lead to false detection of a forced change in ENSO or a robust difference between two models.

~~These~~ All of the aforementioned studies demonstrate that different applications require different ensemble sizes. ~~But they also~~ However, these studies suffer from two drawbacks. First, the required ensemble size can only be estimated once a signal has been identified in a large ensemble, which requires the large ensemble to exist and be large enough in the first place. Second, the result might be model dependent and may only provide a very rough estimate of the required ensemble size when addressing the same question with a different model.

In this paper, we introduce a basic recipe for estimating the required ensemble size in section 3. The required or ideal ensemble size ~~is not only dependent on the model used, but also on the~~ depends on the region and quantity that is investigated and the type of question. 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: (i)~~ represent questions typically addressed with large ensembles:

1. How many ensemble members are required to identify the response to a change in the external forcing? (Section ~~4.1) (ii)~~ 4.1)
2. How many ensemble members are required to adequately sample the spectrum of internal variability? (Section ~~4.2) (iii)~~ 4.2)
3. How many ensemble members are required to identify a forced change in internal variability (e.g., a mode of variability such as ENSO)~~(Section 4.3) ?~~ ? (Section 4.3)

An additional discussion of caveats associated with the choice of sampling method is discussed in Appendix A and is relevant for users of the approach proposed in this study.

2 ~~The resampling problem~~

~~The main difficulty when determining the required ensemble size for a specific question is resampling: in this study~~

2 Model

In this study, we are using simulations from the Max Planck Institute Grand Ensemble (MPI-GE). The MPI-GE consists of large initial-condition ensembles for several experiments with the Max Planck Institute Earth System Model (MPI-ESM) in its low-resolution configuration. Ensemble members are generated by sampling different years from a 2000-year pre-industrial control simulation for the initial conditions (macro-initialisation). The forcing for the experiments follows the protocol of the CMIP5 simulations (Taylor et al., 2012). The model configuration and experiments are described in more detail in Maher et al. (2019).

In this study, we use three experiments from the MPI-GE:

- pre-industrial control simulation (2000 years)
- historical simulations (1850-2005, ~~we generate ensembles of different ensemble sizes by randomly sampling members from a 200-member ensemble. Samples generated in this way are not fully independent when approaching the full ensemble size. For example, two random samples of 190 out of the available~~ 200 members~~will share most of their members~~)
- 1% CO_2 simulations (156 years, 100 members)

Note that only the first 100 historical realisations are described in Maher et al. (2019). Realisations 101–200 were added later and use the same configuration as the first 100 realisations, but are initialised from different years of the pre-industrial control simulation.

3 A simple method to estimate the required ensemble size

In this section, we use a simple example to design a generic recipe for estimating the required ensemble size for any given application. In the following sections 4.1 to 4.3, we then apply this recipe to various examples. ~~This resampling introduces a problem when the signal is defined by using the full ensemble. Any subsample that is close to the full ensemble size will then indicate that the ensemble size is sufficient by construction. In this section, we illustrate the resampling problem and propose how we can ensure that our result is not dominated by resampling.~~

One of the most common applications of a large ensemble is ~~to separate the~~ the separation the forced response and the random internal variability in a time-series. Each realisation from a large ensemble ~~experiences~~ is subject to the same external forcing. Due to different initial conditions, each realisation is a combination of the forced response due to this external forcing and a unique trajectory of quasi-random internal variability. By averaging over a large number of realisations, internal variability cancels out and the forced response remains (Frankcombe et al., 2015). Therefore, the ensemble mean of a large ensemble is often referred to as the forced response. Figure 1 shows the ensemble mean ~~GSAT (~~global mean near-surface air temperature (GSAT, blue line) of 200 realisations with CMIP5 historical forcing from the MPI-GE (Maher et al., 2019). Because of the large ensemble size and the use of a globally averaged quantity, the 200-member mean is a clean estimate of the forced response.

~~The forced response can be quantified using the ensemble mean in a large ensemble, while the ensemble mean of smaller ensembles is contaminated by internal variability. The figure is based on global and annual mean near-surface air temperature from the MPI-GE 200 member historical ensemble. The dark blue line shows the 200-member ensemble mean time series. Shaded regions show the range of forced responses estimated by resampling 1000 times for various ensemble sizes. The light blue shading shows the range of the full ensemble, i.e. the minimum to maximum of all 200 realisations for every single year.~~

Assuming that the 200-member mean provides a good estimate of ~~internal variability~~the forced response, we can then subset the large ensemble to investigate how well the ensemble mean of a smaller ensemble can isolate the forced response. We draw 1000 random samples of sets of 3 members from MPI-GE without replacement. For each of these samples, the 3-member ensemble mean is computed. The red envelope in figure 1 shows the range of these 1000 samples of a 3-member mean forced response. Compared to individual realisations (~~light blue~~ grey envelope), a 3-member mean reduces internal variability, but can deviate substantially from the 200-member mean. Repeating this analysis for 10, 20, and 50 members shows that a larger ensemble size can separate the forced response from internal variability more effectively.

To quantify how effective the separation of forced response and internal variability is, we show the ~~RMSE~~ root-mean-square error (RMSE) of ensemble means for different ensemble sizes compared to the 200-member mean. The solid black line in figure ~~??~~ 2 shows how the expected RMSE decreases with increasing ensemble size until reaching zero for 200 members. By choosing an acceptable error, we can then determine the required ensemble size. For example, an acceptable error of 0.02°C would mean that an ensemble with approximately 50 members is required. We will return to the discussion of what constitutes an acceptable error in the examples in sections 4.1 to 4.3.

While a reduction in the error with increasing ensemble size is expected and indicates that a larger ensemble allows a more accurate representation of the forced response, the vanishing error when using 200 members occurs by construction because we assume that the 200-member mean represents the true forced response. How fast the error is converging therefore depends on how the random samples are generated.

3.1 A cautionary note on resampling

One difficulty when determining the required ensemble size for a specific question is the chosen sampling approach: in this study, we generate synthetic ensembles of different ensemble sizes by randomly sampling members from a 200-member ensemble without replacement. Samples generated in this way are not fully independent when approaching the full ensemble

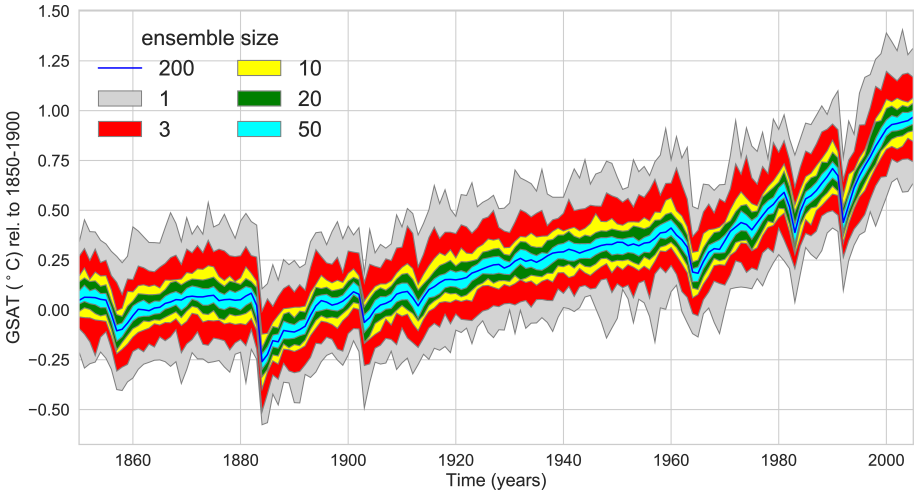


Figure 1. The forced response can be quantified using the ensemble mean in a large ensemble, while the ensemble mean of smaller ensembles still contains a contribution from internal variability. The figure is based on global and annual mean near-surface air temperature from the MPI-GE 200 member historical ensemble. The dark blue line shows the 200-member ensemble mean time series. Shaded regions show the range of forced responses estimated by resampling 1000 times for various ensemble sizes. The light grey shading shows the range of the full ensemble, i.e. the minimum to maximum of all 200 realisations for every single year.

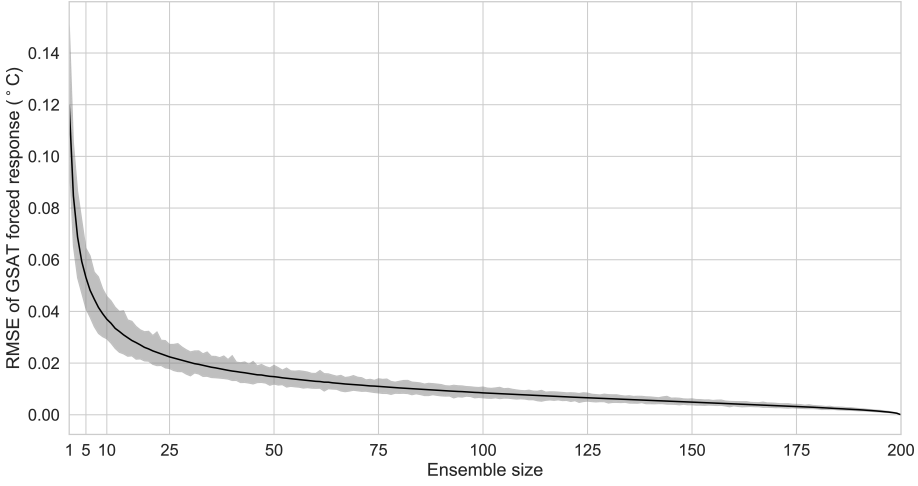


Figure 2. A larger ensemble allows a more accurate quantification of the forced response. The black line shows the mean RMSE for GSAT for ensemble sizes from 2 to 200. The reference is the 200-member mean from figure 1 and the RMSE is computed for all 1000 samples. The shaded area shows the range of RMSE values for individual samples, the solid line shows the mean RMSE.

size. For example, two random samples of 190 out of the available 200 members will share most of their members. This

resampling introduces a problem when the signal is defined by using the full ensemble. Any subsample that is close to the full ensemble size will then indicate that the ensemble size is sufficient by construction.

The resampling problem occurs with any limited sample. At some point, the 1000 random subsamples are not independent anymore because they share many of the randomly drawn members from the full ensemble. Therefore, they look more similar to each other, but also more similar to the 200-member mean. To demonstrate how this resampling affects our estimate of the error, we deliberately reduce the size of the ensemble. For instance, by only using the first 150 members and repeating the analysis(purple line in figure **??**), the random samples are subsets of these 150 members. Because the 150-member mean is now used as the best estimate, the RMSE is – by construction – zero at 150 members. Similar behavior can be seen when only using the first 100 (red), 75 (green), . In an empirical analysis, we find that samples using more than 50(blue), and first 20 members (yellow line).

In a smaller ensemble, the RMSE converges to zero earlier. This is caused by resampling and does not indicate that the error is small. The black line shows the mean RMSE for GSAT for ensemble sizes from 2 to 200. The reference is the 200-member mean from figure 1 and the RMSE is computed for all 1000 samples. The shaded area shows the range of RMSE values for individual samples, the solid line shows the mean RMSE. The other colors show the same analysis after excluding the last 50 members (purple), 100 members (red), 125 members (green), 150 members (blue), and 180 members (yellow) from the ensemble.

We investigate at which sample sizes the reduction of the error mainly occurs because of an increased ensemble size , or simply because of resampling that leads to an error convergence without additional information about a sufficient ensemble size. 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. Therefore, the comparison of the different maximum ensemble sizes in figure **??** indicates when the resampling begins to affect the error convergence. For ensemble sizes that are much smaller than the maximum ensemble size, the different random samples are largely independent and therefore hardly affected by resampling. When increasing the ensemble size in the subsamples, the resampling starts to affect the error estimate for a small maximum ensemble size (% of the available ensemble size to generate random samples lead to a substantial bias in the error estimate. We therefore recommend to treat results indicating that e.g. 20 members) whereas the samples are still independent when drawn from a much larger maximum ensemble size (e.g. more than 100 out of 200 members). 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. Beyond this sample size, the error estimate cannot be used to approximate the true error. are required with caution because the true required ensemble size might be much larger. A more detailed discussion is provided in Appendix A.

We find that the RMSE estimates for different maximum ensemble sizes in figure **??** always start to diverge when about 50% of the maximum ensemble size are used. This implies that up to 50% of the maximum ensemble size can be used to estimate the forced response of GSAT in a transient forcing scenario without inaccuracy caused by resampling.

The same resampling problem also occurs for other questions. To demonstrate this, we investigate how many members are necessary to sample ENSO variability. We use the 50-year standard deviation of the Niño3.4 box to quantify ENSO variability.

~~A single 50-year period is treated as one ensemble member. Random subsamples of 50-year periods from the 2000-year pre-industrial control run from the MPI-GE are used to generate a synthetic ensemble. In figure A3, the red envelope shows that by averaging the standard deviation from more members, a more accurate estimate of ENSO variability can be obtained.~~

~~Using the last 1000 years of the 2000 year control simulation pdfs of the standard deviation calculated over 50 years in the Niño3.4 box (Nic to check this is correct box) are created by resampling the control simulation 1000 times. The pdfs are shown for different ensemble sizes (red: 1000members, blue: 500 members, yellow: 200 members, green: 100 members and light blue: 50 members). 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). Nicola will create a better version of this next week.~~

~~We then reduce the maximum ensemble size by using only 500 (200, 100, and 50) years from the control run. Similar to the result in figure **??**, the error appears to converge when approaching the maximum ensemble size. By comparing the different maximum ensemble sizes in figure A3, we can see that the resampling begins to affect the error estimate when the ensemble size approaches 50% of the maximum ensemble size.~~

~~These two independent lines of evidence demonstrate that resampling affects the error estimate when using more than 50% of the available maximum sample size (either ensemble members or years in a pre-industrial control run). Beyond this ensemble size, the analysis does not provide a realistic estimate of the error and conclusions about the required ensemble size will be biased low.~~

4 ~~A recipe for estimating ensemble size~~

~~We~~

3.1 A recipe for estimating ensemble size

Based on the example introduced in this section, we suggest the following approach to ~~arrive at~~ derive a robust estimate of the required ensemble size for any application. This method can either be applied to one of the existing large ensembles, as shown above for the MPI-GE, or to a long control run, which is available for all models participating in CMIP. We summarise the method in ~~5~~ five steps before applying it to several examples in the next section:

1. Define the question to be addressed (isolate a forced response, quantify variability, detect a change in variability).
2. Choose an error metric (e.g. RMSE or variance across samples) and an upper threshold based on the maximum error that is acceptable in the specific application.
3. Estimate the error for different ensemble sizes by subsampling a long control run or a large ensemble of transient simulations.
4. Determine the minimum ensemble size that is required to reduce the error below the threshold chosen in step 2.

5. If the ensemble size determined in this way is less than 50% of the available sample size (e.g. 50 members when subsampling a 100-member ensemble), then the estimated required ensemble size provides a robust estimate for the specific question and model investigated. If the estimated required ensemble size is larger than 50% of the available sample size, then the estimate is biased low and the true required ensemble size could be substantially larger.

4 Estimating the required ensemble size: applications

In this section we use the pre-industrial control ~~run and the historical~~ simulation and transient forced simulations from the MPI-GE to estimate the required ensemble size for a variety of applications, ranging from global to regional quantities. We investigate the different aspects of quantifying the forced response or quantifying internal variability.

4.1 Quantifying the forced response

The forced response shown in figure 1 contains various signals. The most prominent signal is the ~~long term~~ long-term warming trend caused by anthropogenic greenhouse gas emissions. On shorter time scales, volcanic eruptions lead to a cooling of the global mean surface temperature.

In the first example, we continue to use the RMSE to quantify how well the entire forced response is estimated, but we move from the global mean to the regional forced response in near-surface air temperature in the historical runs from the MPI-GE. In figure 3 ~~a–c~~a–e, the expected RMSE for each grid point is shown for ensemble sizes of 3, 5, 10, ~~and~~ 50~~members.~~ , and 100 members. This analysis is equivalent to the computation of the mean RMSE for GSAT (black line in figure 2), but applied to each grid point separately. The RMSE is computed as the mean difference between 100 samples and the ~~100-member~~ 200-member mean. When the ensemble mean is based on just 3 members, the expected error in the estimated forced response is large over land regions, in particular in the northern hemisphere. Over the ocean, the RMSE is already small in many regions. Increasing the ensemble size reduces the error. At 50 members, the error is small in most regions of the globe. Because 50 members is smaller than 50% of the maximum ensemble size (200 members), the error estimate for this ensemble size is reliable~~, while larger ensemble sizes are affected by resampling and therefore not shown~~.

To estimate how many members are sufficient to reduce the error below a critical threshold, we first need to determine what is an acceptable error as outlined in step 2 of the recipe. This choice will depend on the region of interest and the accuracy to which the forced response needs to be quantified. In figure 3 ~~c–h~~f–j, we show how many members are necessary to estimate the forced response in near-surface air temperature for ~~four~~ five acceptable errors that were chosen for illustrative purpose. If the acceptable error (RMSE) is 0.1°C, 10-30 ensemble members are sufficient over the tropical ocean, while more than 50 ensemble members are required over most land regions. Beyond ~~50~~ 100 members, the resampling problem inhibits reliable estimates of the sufficient ensemble size. For an acceptable error of 0.25°C, less than 10 members are sufficient over most ocean regions, while more than 50 members are required over high northern latitude land regions. For an acceptable error of 0.5°C, only high-latitude land regions require a large ensemble while the forced response over ocean and land regions at lower latitudes can be estimated with less than 10 members.

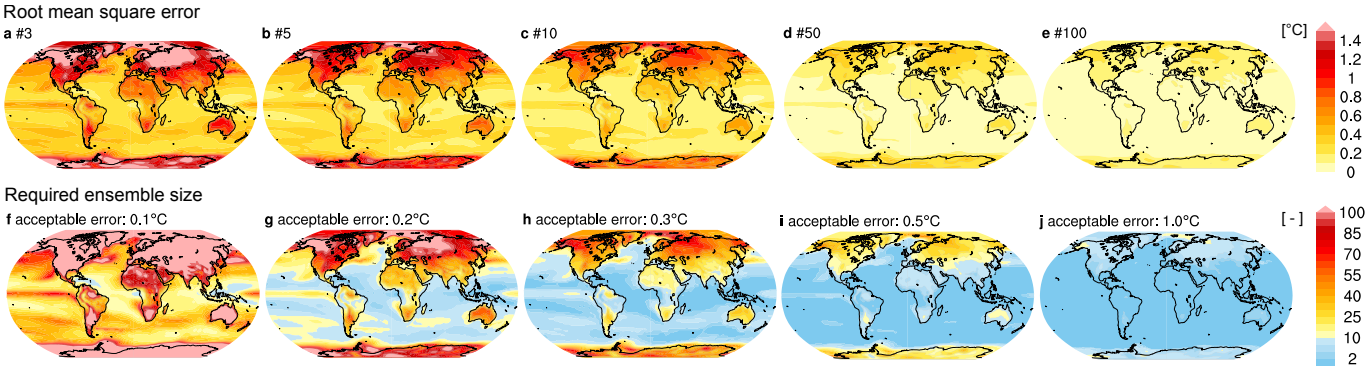


Figure 3. ~~**a–d**~~**a-e**, The mean RMSE for the forced response in historical monthly mean near-surface air temperature of MPI-GE for **a**, 3,**b**, 5, **c**, 10, ~~and~~ **d**, 50, and **e**, 100 ensemble members relative to the ~~100-member~~ 200-member mean, globally. ~~**e–h**~~The RMSE shown here is the mean from 100 random samples without replacement. **f-j**, Required ensemble size to capture the ~~100-member~~ 200-member mean forced response in historical monthly mean near-surface air temperature dependent on the acceptable error of ~~**a**~~**f**, 0.1, ~~**b**~~**g**, ~~0.25~~0.2, ~~**e**~~**h**, 0.3, **i**, 0.5, and ~~**d**~~**j**, 1.0°C.

Conversely for rainfall, the error in estimating the forced signal when using a small ensemble is larger over the tropics than over the higher latitudes (Figure 4 ~~a–d~~a–e). The largest errors can be found over the Indian ~~ocean~~ Ocean and western tropical Pacific. Similar to temperature, a 50-member ensemble shows very small errors across the globe.

In figure 4 ~~e–h~~ f–j we show how many members are necessary to estimate the forced response with an acceptable error of 0.1, 0.2, 0.3, 0.5, and 1 mm/day. For an acceptable error of 0.2 mm/day, ~~many~~ some ocean regions require more than ~~50~~ 100 members to capture the forced rainfall response with the required accuracy, while less than 20 members are sufficient over northern Africa and Eurasia. Over large parts of America, between 20 to 40 members are required to estimate the forced rainfall response. For an acceptable error of 0.5 mm/day, 20 to 40 members are required over the Indian ~~ocean~~ Ocean and western tropical Pacific, while less than 10 members are sufficient elsewhere.

For the example in figures 3 ~~–~~and 4, the objective was to isolate the full forced response in a time series, defined as the ~~100-member~~ 200-member ensemble mean time series at every grid point. The full forced response includes all external forcings, both natural and anthropogenic. In many applications, the objective might be to isolate a specific feature of the forced response rather than all components. In the following two examples, we will demonstrate how to estimate the required ensemble size needed to isolate the global warming trend in the 20th century and the global cooling after a major volcanic eruption~~can be estimated~~.

The global warming signal follows a much simpler trajectory than the forced response to all external forcings (cf. figure 1). Here, we fit a linear trend to the historical time series for 1920 to 2005 and define the 200-member mean as the true forced warming trend. Over the 68-year period from 1920 to 2005, the model warms by 0.65 K (figure 5). We acknowledge that a linear trend may not represent the anthropogenic warming accurately, but use this definition to illustrate how a specific aspect of the forced response can be investigated.

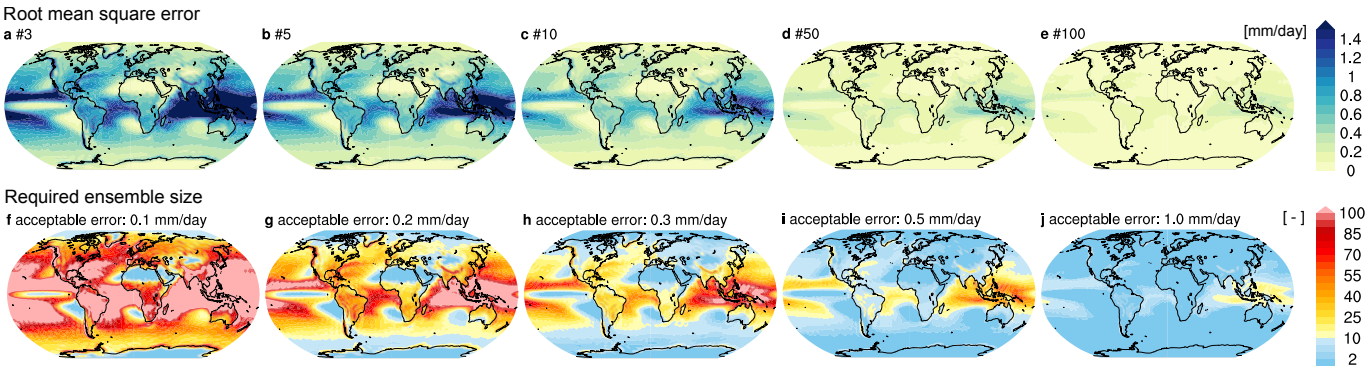


Figure 4. ~~**a-d**~~**a-e**, The mean RMSE for the forced response in historical monthly mean total precipitation of MPI-GE for **a**, 3,**b**, 5, **c**, 10, ~~and~~ **d**, 50, and **e**, 100 ensemble members relative to the ~~100-member~~ 200-member mean, globally. ~~**c-h**~~The RMSE shown here is the mean from 100 random samples without replacement. **f-j**, Required ensemble size to capture the ~~100-member~~ 200-member mean forced response in historical monthly mean total precipitation dependent on the acceptable error of ~~**a**~~**f**, 0.1, ~~**b**~~**g**, 0.2, ~~**c**~~**h**, 0.3, **i**, 0.5, and ~~**d**~~**j**, 1.0 mm day^{-1}.

We subsample the ensemble for smaller ensemble sizes to generate forced warming trends for smaller ensemble sizes. While the trends in a single realisation can be anywhere in the range from 0.4K to more than 0.8K warming over 68 years, increasing the ensemble size to 5 members already leads to a significant reduction in the error (figure 5). The warming trend in every 10-member ensemble is within the 20%-range (±10%, cyan dashed lines) of the true warming trend, indicating that ensembles with 5-10 members can provide a good estimate of the forced linear warming trend. While an error within the 20%-range of the true signal may be sufficient for some applications, the acceptable error for other applications might be larger or smaller and result in a smaller or larger acceptable ensemble size. For an acceptable error of ±15%, 5 ensemble members would be sufficient while for an acceptable error of ±5% at least 25 ensemble members are required. All of these error estimates are below 100 members and therefore not dominated by the resampling problem.

For signals on shorter time-scales, the required ensemble size can be quite different. In figure 6 we analyse the GSAT cooling after the Krakatoa eruption in 1883. The forced cooling is quantified as the difference between 1884, the year after the eruption, and 1882, the year before the eruption. The 200-member mean shows a forced cooling of -0.34K after the eruption. Due to internal variability, a single realisation can even show a warming after the volcanic eruption. ~~At least 5 members are~~ More than 1 member is required for the ensemble mean to capture a cooling in all samples~~, however,~~. However, the ensemble mean cooling ~~can still~~ for 5 members can still exceed the range from -0.2K to -0.5K. More than 50 ensemble members are necessary to estimate the forced cooling within ±15% of the true forced cooling, and approximately 100 members are required to reduce the error below ±10%. Due to the resampling problem, we cannot derive a robust estimate for the ensemble size required to reduce the error to less than ±5%. While the analysis in figure 6 suggests that 150 members would be sufficient for a ±5% error, this number is close to the full ensemble size of 200 members and therefore biased low. The true required ensemble size to reduce the error to ±5% is likely larger than 150 members.

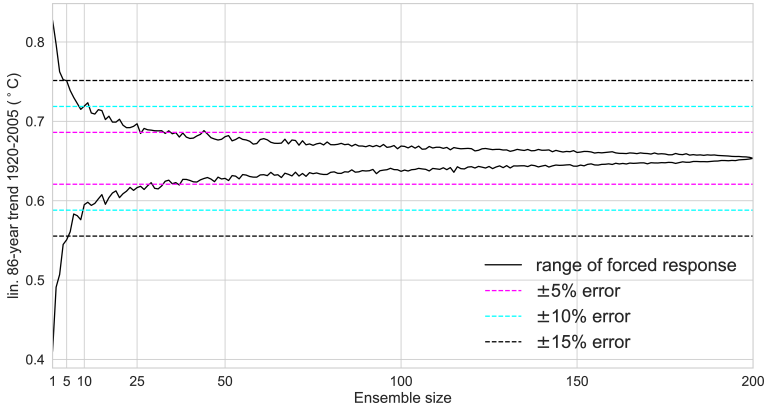


Figure 5. Linear warming trend from 1920 to 2005 for different ensemble sizes shown as a linear trend fitted to the ensemble mean. Black lines show maximum and minimum 86-year ensemble mean temperature trend from 1000 random samples. Errors are shown as percentage of the 200-member ensemble mean temperature trend.

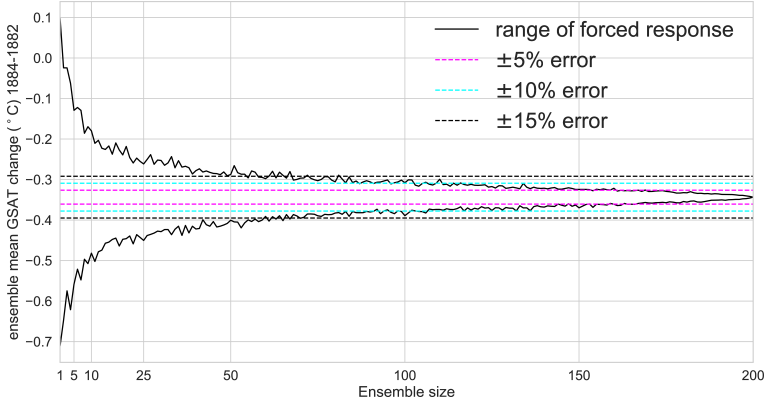


Figure 6. GSAT cooling after Krakatoa eruption for different ensemble sizes shown as the ensemble mean temperature difference between 1882 and 1884. Black lines show maximum and minimum temperature response from 1000 random samples. Errors are shown as percentage of the 200-member ensemble mean temperature response.

These examples demonstrate that the required sample size to estimate the forced response depends on the region and variable (figures 3 -and 4), as well as the feature of interest in the forced response (figures 5 and 6). Whereas for some applications 5 members are sufficient to reduce the error to an acceptable magnitude, other applications require at least 50 members. A robust estimate for the forced response is given by the ensemble mean when averaging over the ensemble attenuates internal variability sufficiently (Frankcombe et al., 2018). The number of members required for this depends both on the magnitude of the forced signal and the magnitude of internal variability, but also on the acceptable error for a specific application.

4.2 Quantifying internal variability

While quantifying the forced response only requires a robust estimate of the mean, quantifying internal variability requires more members because higher order moments of the distribution need to be estimated. In the following two examples, we use the second statistical moment of the distribution, the standard deviation, to quantify internal variability. We note that if the distribution deviates from a normal distribution, only using the standard deviation to quantify internal variability may not be sufficient.

Here, we investigate internal variability in two regions. The tropical Pacific, where the variability is primarily driven by the El-Niño Southern Oscillation (ENSO), and the central United States (34°N-46°N, 116°W-96°W). The tropical Pacific region shows substantial variability on interannual to decadal time scales. Previous work has demonstrated that large sample sizes are necessary to quantify ENSO variability (Maher et al., 2018; Wittenberg, 2009). As a second region, we analyse temperature variability over the central United States. We hypothesise that these two regions should have different requirements for the ensemble size, with a smaller required ensemble size for the central United States than the tropical Pacific to stay within an acceptable error range.

For the following examples we use the 2000-year pre-industrial control ~~integration~~ simulation from the MPI-GE. 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. The disadvantage is that when using the ~~control run~~pre-industrial control simulation, we assume that internal variability does not change under global warming.

We quantify ENSO variability by using the December, January, February (DJF) variability in the Niño3.4 box (5°N-5°S, 170°W-120°W). To ensure that ENSO variability on interannual to multi-decadal time scales is sampled, we use the Niño3.4 standard deviation for a 100-year period. The standard deviation, as computed for the full 2000-year time series is used as the truth in this context and indicated by the horizontal black line in figure 7a. To generate synthetic ensemble members, we split the pre-industrial control simulation into overlapping 100-year segments. Each segment is used as one ensemble member and the temporal standard deviation over the 100-year segment represents ENSO variability for this member. For an ensemble size of one, the spread in ENSO variability seen in figure 7a indicates that individual 100-year periods can have substantially more or less variability than the reference value based on the full control run.

To account for this centennial modulation of ENSO variability, the ENSO variability in multiple ensemble members can be averaged to get a more accurate estimate of the average ENSO variability. We simulate different ensemble sizes by averaging over randomly chosen members for a given ensemble size and repeat this 1000 times. By using a 5-member mean, the error of the estimated variability in all samples is within $\pm 15\%$ of the true value. To reduce the error below $\pm 10\%$, 10 ensemble members are sufficient. To improve the accuracy so that the ENSO variability estimate is within $\pm 5\%$ of the truth, nearly 50 ensemble members are necessary.

For a region with less variability, much smaller ensemble sizes are sufficient to obtain a similar accuracy. For annual mean central US temperatures (figure 7b) any individual realisation is within $\pm 15\%$ of the truth and 10 members are sufficient to increase the accuracy to the $\pm 5\%$ range around the truth, whereas 50 members ~~where~~ are necessary for ENSO. This emphasises

that for some regions and quantities, a moderate ensemble size or even a single realisation can be sufficient to quantify internal variability.

In both examples, the long sampling period of 100 years increases the sample size and thereby improves the accuracy for individual realisations. This is useful if the objective is to quantify variability when stationarity can be assumed, but can be problematic if the objective is to identify a change in variability, such as changes in ENSO characteristics under global warming. A more detailed discussion of estimating ENSO variability, in particular using the ensemble dimension instead of the time dimension in transient simulations to quantify internal variability, can be found in Maher et al. (2018) and Haszpra et al. (2020).

4.3 ~~Quantifying changes in internal variability~~

4.2.1 Notes on sampling from a pre-industrial control simulation

Sampling from a pre-industrial control simulation to estimate the required ensemble size has two advantages: this can be done before producing a large ensemble for the model and is based on a simulation that is available for every climate model in CMIP5 and CMIP6. Different approaches can be used when sampling from a pre-industrial control simulation. In the following, we discuss different options and their advantages and disadvantages.

- *Overlapping segments (applied here):* We choose to use continuous 100-year and 30-year segments to keep temporal autocorrelation intact. From the 2000-year simulation, we can thus generate 20 independent, non-overlapping synthetic realisations (for 100-year segments). To increase the sample size, we allow overlapping segments. These samples are not independent, which leads to a biased estimate as discussed in appendix A, but enables estimates for ensemble sizes larger than 20.

- *Non-overlapping segments:* The advantage of this approach is that synthetic members can be assumed to be independent and temporal autocorrelation is kept intact. However, for long segments or a short pre-industrial control simulation, only a small number of synthetic members can be generated.

- *random year selection to generate synthetic segments or members:* The synthetic segments generated by random year selection allow for a wider variety of samples in a segment than continuous segments sampled from the pre-industrial control simulation. However, information about temporal autocorrelation is lost and synthetic segments could have larger variability than continuous segments in the presence of strong variability on time scales longer than the segment. If the time scale of variability is not the focus of a study, sampling random years to generate synthetic ensemble members can be informative to estimate how well statistics computed across ensemble members (e.g. Maher et al., 2018; Haszpra et al., 2020) capture the model characteristics.

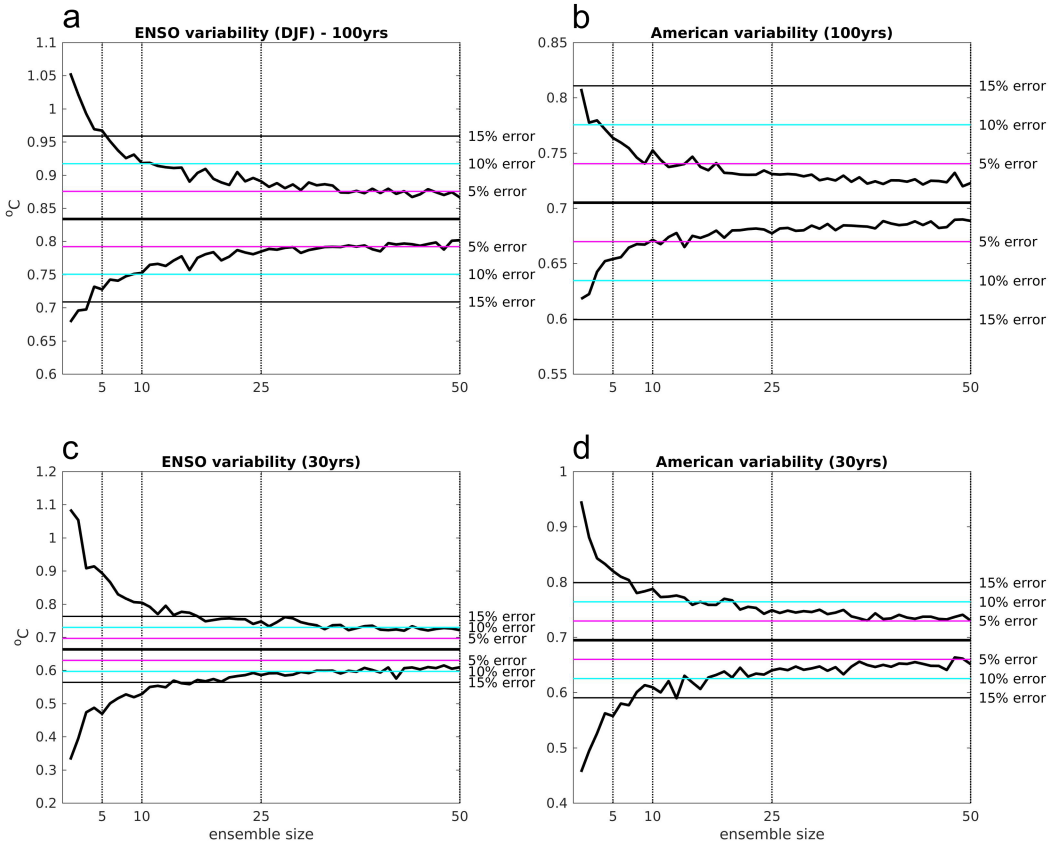


Figure 7. We show for increasing ensemble sizes the: a) ENSO variability in the Niño3.4 box (5°N-5°S, 170°W-120°W) calculated over 100 year periods, b) Central ~~American~~ United States variability (34°N-46°N, 116°W-96°W) calculated over 100 year periods, c) ENSO variability in the Niño3,4 box calculated over 30 year periods, d) Central ~~American~~ United States variability calculated over 30 year periods. All indices are calculated from the 2000 year MPI-GE control run. Each index is calculated as a running value at each time-step in the control. ENSO indices are calculated for DJF and ~~American~~ United States indices are calculated for the annual mean. Ensembles of 1 to 120 members are created by randomly sampling the control simulation without replacement. For each ensemble size we create 1000 artificial ensembles. The estimated true value is calculated by using the entire 2000 years of the control and is shown in the horizontal black line. The maximum and minimum values of each index from the 1000 samples are shown in the solid black lines. Varying error thresholds are shown in the horizontal coloured lines.

4.3 Quantifying changes in internal variability

To quantify changes in internal variability, we need a robust estimate of internal variability both for a reference period and for a period where we want to investigate a potential change in variability (e.g. a pre-industrial control state and a time period in a future scenario). This problem is more challenging than the previous examples because the errors for the variability estimates of the two time periods add up. To demonstrate this, we use the internal variability of September Arctic sea ice area as an example. Previous work has shown that the internal variability in Arctic sea ice area first increases under warming, before it approaches zero when most of the Arctic sea ice has melted (Goosse et al., 2009; Olonscheck and Notz, 2017). We analyse the 100 members from the 1% CO_2 scenario from the MPI-GE and use the ensemble standard deviation as an estimator of internal variability. After 120 years, nearly all ensemble members show a completely ice-free Arctic in September (figure B1a). The internal variability increases from model year 1 to year 80, before it sharply drops reaching zero around year 120 (figure B1b).

Here we focus on the increase in variability from the beginning of the simulation to year 80 and ask how many ensemble members are necessary to robustly quantify this change in internal variability. To increase the sample size, we use a decadal mean of the ensemble standard deviation rather than a single year. We then compute the difference in internal variability between the two time periods for ensemble sizes between 3 and 100 members. ~~In figure 8 we show~~ Figure 8 shows the range of this change in internal variability from 1000 random samples. To quantify the change in variability within $\pm 15\%$ of the true value (here defined as the internal variability change estimated with 100 members), 50 ensemble members are necessary. An error of less than $\pm 10\%$ and $\pm 5\%$ is only reached beyond 50 members. Due to the effect of resampling beyond 50 members, we cannot estimate the required ensemble size for these error thresholds from the 100-member ensemble used here. For very small ensemble sizes, the estimate of the variability change may even show the opposite sign of the true change~~.~~, i.e. a decrease in internal variability.

The large number of ensemble members required to robustly quantify this change in variability shows that identifying a change in internal variability requires the largest ensemble size of all examples shown in this study, even when using decadal averaging to increase the sample size. This is because a robust estimate of a change in internal variability requires a clean separation of internal variability from the forced response and a robust estimate of internal variability for two different time periods. Errors in any of these estimates will propagate to the estimated change in variability, thereby making it more challenging. A small forced change in internal variability will further complicate this analysis.

A first estimate for the magnitude of a detectable change in internal variability can be derived from the control run (as in figure 7). Any change in variability that is smaller than the uncertainty of the estimated internal variability for a given ensemble size is not detectable. We note that this method can also be used to add error bars to estimates of forced changes in internal variability under climate change in small ensembles or single realisations from CMIP and hence determine the robustness of results.

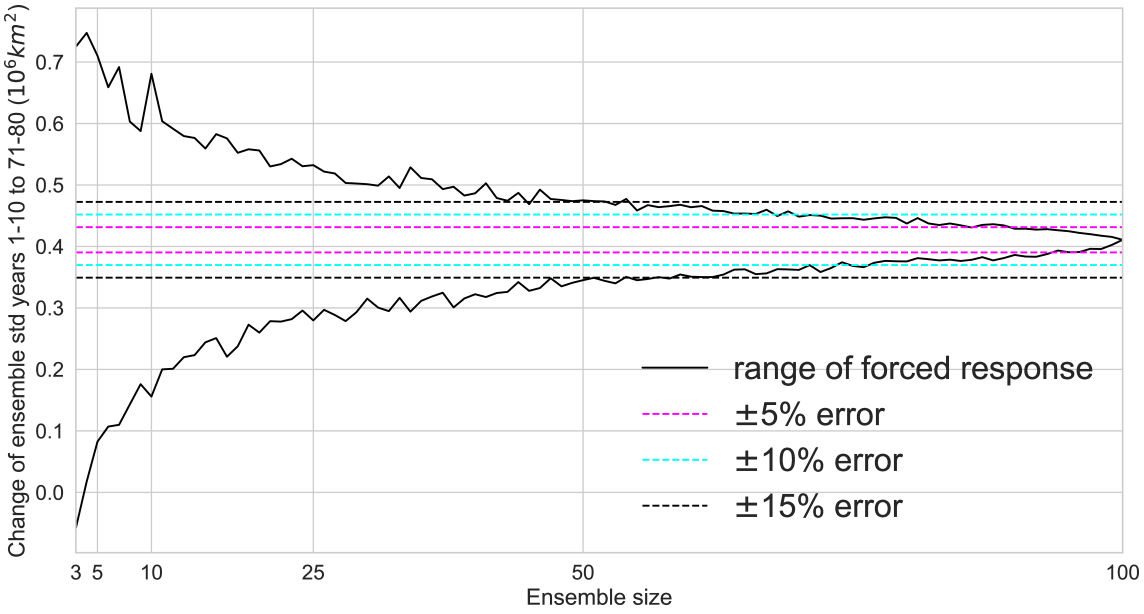


Figure 8. Change in internal variability of September Arctic sea ice are from the first decade to years 71–80 in a 1% CO_2 experiment. For different ensemble sizes, we compute the ensemble standard deviation and then average for the first decade and years 71–80 before computing the difference. Black lines show maximum and minimum change in variability from 1000 random samples. Errors are shown as percentage of the 100-member variability change.

5 Summary and conclusions

Multiple ensemble members for a single climate model are required for robustly estimating the model's forced response to an external forcing change and its internal variability. Without a robust characterisation of these model characteristics, differences between models or a model and observations can easily be misinterpreted as significant differences, while they could be simply caused by an insufficient sample size. Therefore it is important to use an ensemble size that is sufficiently large to allow a robust quantification of the model characteristic that is investigated.

Here we present a generalised approach to estimate the ensemble size that is required to robustly estimate a model's characteristics. While the focus of this study is on the generalised method, the example applications can provide some insight into the required ensemble size for a variety of applications in the MPI-GE. We differentiate three types of question: identifying a forced response, quantifying ~~variability,~~ internal variability, and identifying a change in internal variability. In a next step, an adequate error metric for quantifying the deviations from the true model characteristics is defined and an acceptable error suitable for the application is chosen. By subsampling a pre-industrial control ~~integration~~ simulation or a large ensemble of transient simulations, the error for different ensemble sizes can be estimated. By applying the previously selected acceptable error as a threshold to these error estimates for different ensemble sizes, the minimum required ensemble size for the given question and model can be determined. Because the subsampling of the full sample does not generate independent samples

when approaching the full ensemble size, the error estimate is biased for ensemble sizes close to the available ensemble size. We demonstrate that this resampling effect ~~dominates~~ substantially affects the error estimate when using more than 50% of the full ensemble. For example, a 50 member ensemble cannot be used to conclude that 50 members are sufficient for a given application, because all ensemble estimates beyond 25 members would be affected by resampling and therefore biased.

We apply the method to several examples and use the 200-member historical ensemble~~and~~ , a 2000-year pre-industrial control simulation, and a 100-member 1% CO_2 experiment from the MPI-GE to estimate required ensemble sizes for various applications for the MPI-ESM model.

To identify the externally forced temperature response from 1850–2005, most ocean regions require less than 10 members, while land regions at higher latitudes may require more than 50 members. To characterise rainfall changes over the same period, more ensemble members are required in the tropics than in higher latitudes. While regions that require more ensemble members can be objectively identified, the required number of members depends on a subjective choice of the acceptable error and can therefore vary substantially for different applications.

The analysis of the forced cooling after a volcanic eruption and the analysis of ENSO variability demonstrate that a small ensemble size can lead to a misinterpretation. For the example of the volcanic eruption, an ensemble consisting of ~~less than five~~ 2-3 members could show a warming after the volcanic eruption, while the true forced response of the model is a cooling. For ENSO, a too small ensemble still contains a large uncertainty in the estimate of ENSO variability. This may lead to a misinterpretation of a signal as a forced change in ENSO, whereas it might still be within sampling uncertainty. Wittenberg (2009) show that samples from different time periods in a pre-industrial control simulation can show substantially different ENSO characteristics. Cai et al. (2018) on the other hand use single realisations for different models to identify forced changes in ENSO in future projections. While the robustness of the results seems clear given most models show an increase in ENSO amplitude, we show that within a single model differences between realisations can be large due to internal variability alone. By using the method introduced in this study, we can add to the robustness of studies such as Cai et al. (2018) by adding error bars from the pre-industrial control simulation to each model to ~~see~~ test if changes in variability are indeed robust within each model.

The examples in this study ~~show~~ demonstrate that for some applications ensemble sizes around 5 members are sufficient while other applications require ensemble sizes well above 100 members. ~~This information~~ In section 1 we introduced several estimates for required ensemble sizes from the literature. While most of the applications from previous studies are not directly comparable to the examples we use here, the large range of required ensemble sizes emphasizes the need to systematically estimate the required ensemble size for each individual application. Furthermore, the required ensemble size may be model dependent. Therefore, the numbers derived in this and previous studies should only be used as approximate estimates and supported by a systematic model- and application-specific estimate following the approach outlined in this study.

Information about the sufficient ensemble size is not only crucial when choosing or designing a large ensemble, but can also help to identify applications where a small number of ensemble members is sufficient and thereby inform the design of multi-model intercomparison studies. The method introduced in this study can add to the robustness of results both from single model large ensembles and multi-model ~~large~~ ensembles.

References

Bellenger, H., Guilyardi, É., Leloup, J., Lengaigne, M., and Vialard, J.: ENSO representation in climate models: from CMIP3 to CMIP5, Climate Dynamics, 42, 1999–2018, 2013.

Bittner, M., Schmidt, H., Timmreck, C., and Sienz, F.: Using a large ensemble of simulations to assess the Northern Hemisphere stratospheric dynamical response to tropical volcanic eruptions and its uncertainty, Geophysical Research Letters, 43, 9324–9332, 2016.

Branstator, G. and Selten, F.: 'Modes of Variability' and Climate Change, Journal of Climate, 22, 2639–2658, https://doi.org/10.1175/2008JCLI2517.1, https://doi.org/10.1175/2008JCLI2517.1, 2009.

Cai, W., Wang, G., Dewitte, B., Wu, L., Santoso, A., Takahashi, K., Yang, Y., Carréric, A., and McPhaden, M. J.: Increased variability of eastern Pacific El Niño under greenhouse warming, Nature, 564, 1–18, 2018.

Christensen, J., H., K. K. K., Aldrian, E., An, S.-I., Cavalcanti, I., de Castro, M., Dong, W., Goswami, P., Hall, A., Kanyanga, J., Kitoh, A., Kossin, J., Lau, N.-C., Renwick, J., Stephenson, D., Xie, S.-P., and Zhou, T.: Climate Phenomena and their Relevance for Future Regional Climate Change, in: Climate Change 2013: The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change, edited by Stocker, T. F., Qin, D., Plattner, G.-K., Tignor, M., Allen, S., Boschung, J., Nauels, A., Xia, Y., Bex, V., and Midgley, P., pp. 1217–1308, Cambridge University Press, 2013.

Daron, J. D. and Stainforth, D. A.: On predicting climate under climate change, Environmental Research Letters, 8, 034 021, https://doi.org/10.1088/1748-9326/8/3/034021, https://doi.org/10.1088%2F1748-9326%2F8%2F3%2F034021, 2013.

Deser, C., Knutti, R., Solomon, S., and Phillips, A. S.: Communication of the role of natural variability in future North American climate, Nature Climate Change, 2, 775–779, 2012a.

Deser, C., Phillips, A., Bourdette, V., and Teng, H.: Uncertainty in climate change projections: the role of internal variability, Climate Dynamics, 38, 527–546, 2012b.

Deser, C., Lehner, F., Rodgers, K. B., Ault, T. R., Delworth, T. L., diNezio, P., Fiore, A., Frankignoul, C., Fyfe, J. C., Horton, D. E., Kay, J. E., Knutti, R., Lovenduski, N. S., Marotzke, J., McKinnon, K. A., Minobe, S., Randerson, J., Screen, J. A., Simpson, I. R., and Ting, M.: Strength in Numbers: The Utility of Large Ensembles with Multiple Earth System Models, submitted to Nature Climate Change, 2019.

Drótos, G., Bódai, T., and Tél, T.: On the importance of the convergence to climate attractors, The European Physical Journal Special Topics, 226, 2031–2038, https://doi.org/10.1140/epjst/e2017-70045-7, https://doi.org/10.1140/epjst/e2017-70045-7, 2017.

Ferro, C. A. T., Jupp, T. E., Lambert, F. H., Huntingford, C., and Cox, P. M.: Model complexity versus ensemble size: allocating resources for climate prediction, Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, 370, 1087–1099, 2012.

Frankcombe, L. M., England, M. H., Mann, M. E., and Steinman, B. A.: Separating Internal Variability from the Externally Forced Climate Response, Journal of Climate, 28, 8184–8202, 2015.

Frankcombe, L. M., England, M. H., Kajtar, J. B., Mann, M. E., and Steinman, B. A.: On the Choice of Ensemble Mean for Estimating the Forced Signal in the Presence of Internal Variability, Journal of Climate, 31, 5681–5693, 2018.

Frankignoul, C., Gastineau, G., and Kwon, Y.-O.: Estimation of the SST Response to Anthropogenic and External Forcing and Its Impact on the Atlantic Multidecadal Oscillation and the Pacific Decadal Oscillation, Journal of Climate, 30, 9871–9895, https://doi.org/10.1175/JCLI-D-17-0009.1, https://doi.org/10.1175/JCLI-D-17-0009.1, 2017.

Goosse, H., Arzel, O., Bitz, C. M., de Montety, A., and Vancoppenolle, M.: Increased variability of the Arctic summer ice extent in a warmer climate, Geophysical Research Letters, 36, 401–5, 2009.

Haszpra, T., Herein, M., and Bódai, T.: Investigating ENSO and its teleconnections under climate change in an ensemble view – a new perspective, Earth System Dynamics, 11, 267–280, 2020.

Kay, J. E., Deser, C., Phillips, A., Mai, A., Hannay, C., Strand, G., Arblaster, J. M., Bates, S. C., Danabasoglu, G., Edwards, J., Holland, M., Kushner, P., Lamarque, J. F., Lawrence, D., Lindsay, K., Middleton, A., Munoz, E., Neale, R., Oleson, K., Polvani, L., and Vertenstein, M.: The Community Earth System Model (CESM) Large Ensemble Project: A Community Resource for Studying Climate Change in the Presence of Internal Climate Variability, Bulletin of the American Meteorological Society, 96, 1333–1349, 2015.

Kirchmeier-Young, M. C., Zwiers, F. W., and Gillett, N. P.: Attribution of Extreme Events in Arctic Sea Ice Extent, Journal of Climate, 30, 553–571, https://doi.org/10.1175/JCLI-D-16-0412.1, https://doi.org/10.1175/JCLI-D-16-0412.1, 2017.

Li, H. and Ilyina, T.: Current and Future Decadal Trends in the Oceanic Carbon Uptake Are Dominated by Internal Variability, Geophysical Research Letters, 45, 916–925, 2018.

Maher, N., Matei, D., Milinski, S., and Marotzke, J.: ENSO change in climate projections: forced response or internal variability?, Geophysical Research Letters, pp. 1–27, 2018.

Maher, N., Milinski, S., Suárez-Gutiérrez, L., Botzet, M., Dobrynin, M., Kornblueh, L., Kröger, J., Takano, Y., Ghosh, R., Hedemann, C., Li, C., Li, H., Manzini, E., Notz, D., Putrasahan, D., Boysen, L., Claussen, M., Ilyina, T., Olonscheck, D., Raddatz, T., Stevens, B., and Marotzke, J.: The Max Planck Institute Grand Ensemble: Enabling the Exploration of Climate System Variability, Journal of Advances in Modeling Earth Systems, 28, 867–20, 2019.

Olonscheck, D. and Notz, D.: Consistently Estimating Internal Climate Variability from Climate Model Simulations, Journal of Climate, 30, 9555–9573, https://doi.org/10.1175/JCLI-D-16-0428.1, https://doi.org/10.1175/JCLI-D-16-0428.1, 2017.

Pausata, F. S. R., Chafik, L., Caballero, R., and Battisti, D. S.: Impacts of high-latitude volcanic eruptions on ENSO and AMOC, Proceedings of the National Academy of Sciences, 112, 13 784–13 788, 2015a.

Pausata, F. S. R., Grini, A., Caballero, R., Hannachi, A., and 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, 26 728–17, 2015b.

Rodgers, K. B., Lin, J., and Frölicher, T. L.: Emergence of multiple ocean ecosystem drivers in a large ensemble suite with an Earth system model, Biogeosciences, 12, 3301–3320, 2015.

Steinman, B. A., Frankcombe, L. M., Mann, M. E., Miller, S. K., and England, M. H.: Response to Comment on "Atlantic and Pacific multidecadal oscillations and Northern Hemisphere temperatures", Science, 350, 1326–1326, 2015.

Stevenson, S., Fox-Kemper, B., Jochum, M., Neale, R., Deser, C., and Meehl, G.: Will There Be a Significant Change to El Niño in the Twenty-First Century?, Journal of Climate, 25, 2129–2145, 2012.

Stolpe, M. B., Medhaug, I., Sedláček, J., and Knutti, R.: Multidecadal Variability in Global Surface Temperatures Related to the Atlantic Meridional Overturning Circulation, Journal of Climate, 31, 2889–2906, https://doi.org/10.1175/JCLI-D-17-0444.1, https://doi.org/10.1175/JCLI-D-17-0444.1, 2018.

Taylor, K. E., Stouffer, R. J., and Meehl, G. A.: An Overview of CMIP5 and the Experiment Design, Bulletin of the American Meteorological Society, 93, 485–498, 2012.

Thompson, D. W. J., Barnes, E. A., Deser, C., Foust, W. E., and Phillips, A. S.: Quantifying the Role of Internal Climate Variability in Future Climate Trends, Journal of Climate, 28, 6443–6456, 2015.

von Känel, L., Frölicher, T. L., and Gruber, N.: Hiatus-like decades in the absence of equatorial Pacific cooling and accelerated global ocean heat uptake, Geophysical Research Letters, 44, 7909–7918, https://doi.org/10.1002/2017GL073578, https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1002/2017GL073578, 2017.

Wittenberg, A. T.: Are historical records sufficient to constrain ENSO simulations?, Geophysical Research Letters, 36, 3–5, 2009.

Zelle, H., Jan van Oldenborgh, G., Burgers, G., and Dijkstra, H.: El Niño and Greenhouse Warming: Results from Ensemble Simulations with the NCAR CCSM, Journal of Climate, 18, 4669–4683, https://doi.org/10.1175/JCLI3574.1, https://doi.org/10.1175/JCLI3574.1, 2005.

Code availability. Primary data and scripts used in the analysis and other supporting information that may be useful in reproducing the author's work are archived by the Max Planck Institute for Meteorology and can be obtained by contacting publications@mpimet.mpg.de.

Data availability. Output from the MPI Grand Ensemble that was used in this study and additional output can be downloaded from https://www.mpimet.mpg.de/en/grand-ensemble/.

Author contributions. All authors conceptualised the study, and carried out the formal analysis. SM wrote the original draft with input from all authors.

Competing interests. The authors declare that they have no conflict of interest.

Acknowledgements. We thank Chao Li for conducting an internal review of the manuscript, Jin-Song von Storch for helpful comments, and the Max Planck Society for the Advancement of Science for funding all three authors. We thank Mikhail Dobrynin and Johanna Baehr from the University of Hamburg for completing the second hundred MPI-GE ensemble simulations and providing the data from these simulations for use in this paper. DO was supported by the European Union's Horizon 2020 Research and Innovation Programme under grant agreement number 820829 (CONSTRAIN).

Appendix A: Notes on sampling

In this study, we made several choices on how we sample from a large ensemble or pre-industrial control simulation. In this section, we discuss alternative sampling approaches and caveats.

A1 Resampling with and without replacement

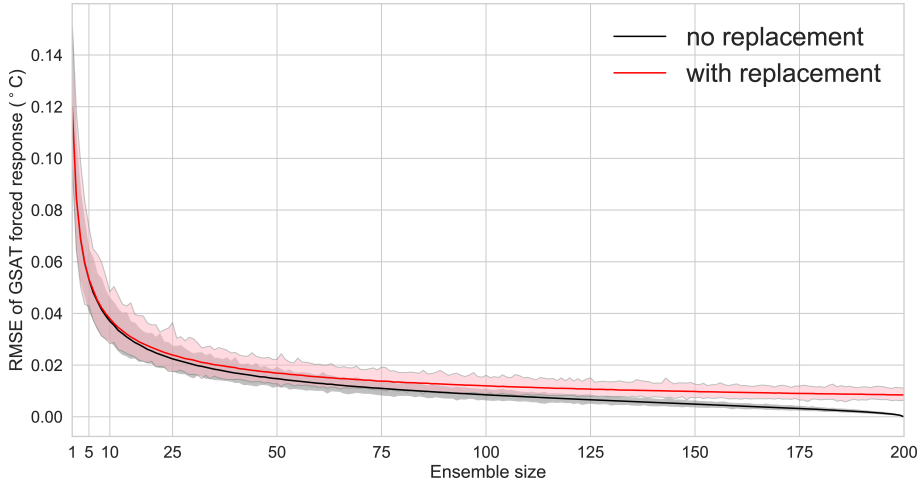


Figure A1. Sampling with or without replacement affects the error estimate and therefore the estimate for the required ensemble size. The black line shows the mean RMSE for GSAT for ensemble sizes from 2 to 200. The reference is the 200-member mean from figure 1 and the RMSE is computed for all 1000 samples. The shaded area shows the range of RMSE values for individual samples, the solid line shows the mean RMSE. The red line and shading show the RMSE for ensemble sizes from 2 to 200, but samples are generated by allowing sampling with replacement.

We choose to resample without replacement for all examples shown. While this choice leads to ambiguities in error convergence as discussed in the following section A2, we argue that sampling without replacement is a better proxy for what we try to imitate by resampling: a random set of members that we could have produced when running a given number of realisations. Sampling with replacement would mean that for example a randomly sampled 5-member ensemble could contain two (or more) identical realisations. Given how SMILEs are initialised, this is unlikely to happen and even if it would happen, such an ensemble would not be used as a set of independent realisations without careful investigation.

In figure A1, we repeat the analysis shown in figure 2 but allow replacement when resampling from the 200 members. We still use the 200-member mean as the reference for the forced response in historical GSAT. Sampling with replacement results in a consistently larger error estimate for the mean RMSE, resulting in a larger required ensemble size for a given error.

A2 How resampling from a small ensemble can bias the error estimate

Generating samples without replacement as applied in this study can bias the error estimate when approaching the full ensemble size. We use the distribution parameters of the full ensemble, e.g. the mean or standard deviation, as the 'truth' in many of the examples shown here. When the size of the sample approaches the size of the full ensemble, for example 190 members from a 200-member ensemble, the difference between these ensembles will be small because they share most of their members. This results in a small error estimate, but does not necessarily mean that 190 members are sufficient for a given application.

The resampling problem occurs with any limited sample. At some point, the 1000 random subsamples are not independent anymore because they share many of the randomly drawn members from the full ensemble. Therefore, they look more similar to each other, but also more similar to the 200-member mean. To demonstrate how this resampling affects our estimate of the error, we deliberately reduce the size of the ensemble. For instance, by only using the first 150 members and repeating the analysis (purple line in figure A2), the random samples are subsets of these 150 members. Because the 150-member mean is now used as the best estimate, the RMSE is—by construction—zero at 150 members. Similar behavior can be seen when only using the first 100 (red), 75 (green), 50 (blue), and first 20 members (yellow line).

We investigate at which sample sizes the reduction of the error mainly occurs because of an increased ensemble size, or simply because of resampling that leads to an error convergence without additional information about a sufficient ensemble size. 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. Therefore, the comparison of the different maximum ensemble sizes in figure A2 indicates when the resampling begins to affect the error convergence. For ensemble sizes that are much smaller than the maximum ensemble size, the different random samples are largely independent and therefore hardly affected by resampling. When increasing the ensemble size in the subsamples, the resampling starts to affect the error estimate for a small maximum ensemble size (e.g. 20 members) whereas the samples are still independent when drawn from a much larger maximum ensemble size (e.g. 200 members). 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. Beyond this sample size, the error estimate should not be used to approximate the true error.

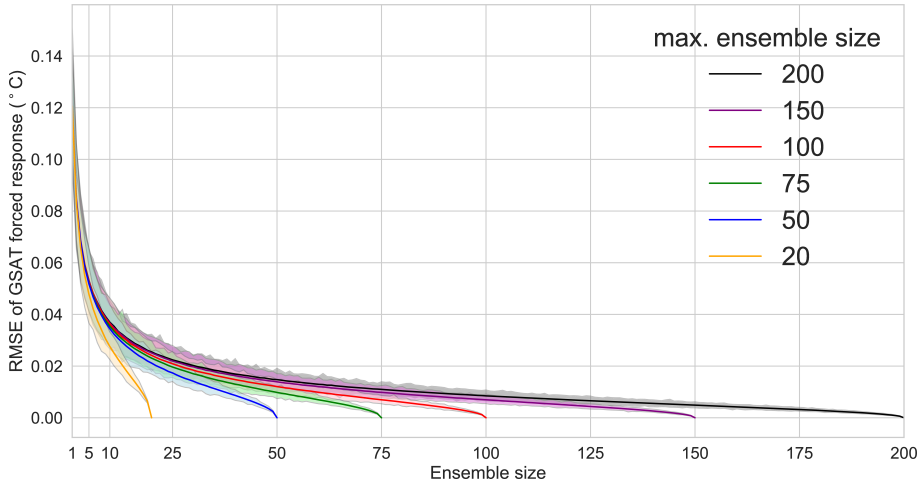


Figure A2. In a smaller ensemble, the RMSE converges to zero earlier. This is caused by resampling and does not indicate that the error is small. The black line shows the mean RMSE for GSAT for ensemble sizes from 2 to 200. The reference is the 200-member mean from figure 1 and the RMSE is computed for all 1000 samples. The shaded area shows the range of RMSE values for individual samples, the solid line shows the mean RMSE. The other colors show the same analysis after excluding the last 50 members (purple), 100 members (red), 125 members (green), 150 members (blue), and 180 members (yellow) from the ensemble.

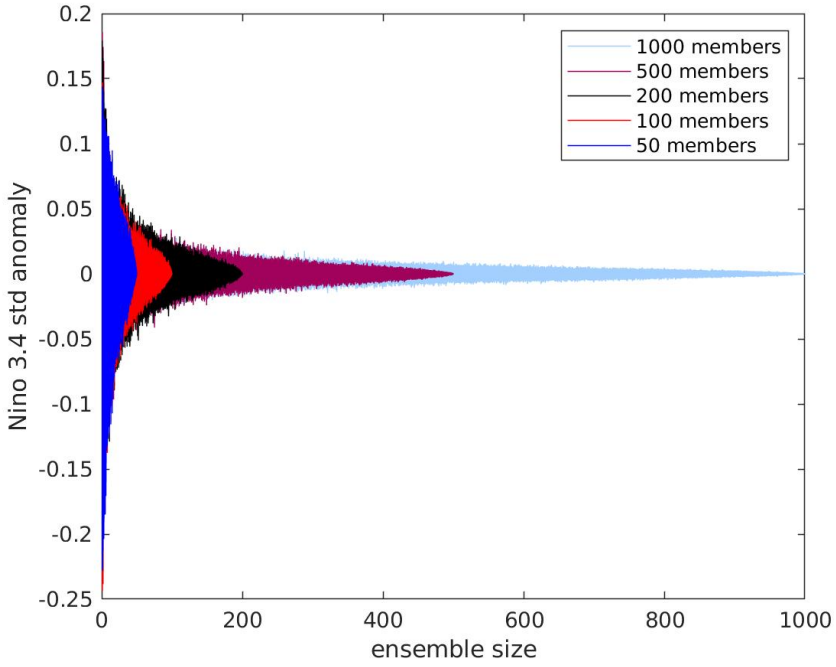


Figure A3. PDF of ensemble-averaged Niño3.4 standard deviations possible in the MPI-GE pre-industrial control simulation for subsampling ensembles ranging from 50 to 1000 members (shown as different colors) for smaller ensemble sizes. Each PDF is shown relative to the corresponding ensemble mean value. 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 find that the RMSE estimates for different maximum ensemble sizes in figure A2 always start to diverge when about 50% of the maximum ensemble size are used. This implies that up to 50% of the maximum ensemble size can be used to estimate the forced response of GSAT in a transient forcing scenario without a major impact from resampling.

The same resampling problem also occurs for other questions. To demonstrate this, we investigate how many members are necessary to sample ENSO variability. We use the 50-year standard deviation of the Niño3.4 box to quantify ENSO variability. A single 50-year period is treated as one ensemble member. Random subsamples of 50-year periods from the 2000-year pre-industrial control simulation from the MPI-GE are used to generate a synthetic ensemble. In figure A3, the light blue envelope shows that by averaging the standard deviation from more members, a more accurate estimate of ENSO variability can be obtained.

We then reduce the maximum ensemble size by using only 500 (200, 100, and 50) years from the control run. Similar to the result in figure A2, the error appears to converge when approaching the maximum ensemble size. By comparing the different

maximum ensemble sizes in figure A3, we can see that the resampling begins to affect the error estimate when the ensemble size approaches 50% of the maximum ensemble size.

These two independent lines of evidence demonstrate that resampling affects the error estimate when using more than 50% of the available maximum sample size (either ensemble members or years in a pre-industrial control simulation). Beyond this ensemble size, the analysis does not provide a realistic estimate of the error and conclusions about the required ensemble size will be biased low. We note that for very simple applications, such as the mean of a stationary time series, the error scales with $\frac{1}{\sqrt{n}}$. For more complex error estimates, such as the RMSE between non-stationary time series, the scaling law is not as simple, which is why we rely on the empirical analysis outlined above.

Appendix B: Arctic sea ice area under strong warming

The internal variability of September Arctic sea ice area is known to change under global warming. In this study, we use September Arctic sea ice area as an example for a quantitiy with a change in internal variability under global warming.

Previous work has shown that the internal variability in Arctic sea ice area first increases under warming, before it approaches zero when most of the Arctic sea ice has melted (Goosse et al., 2009; Olonscheck and Notz, 2017). We analyse the 100 members from the 1% CO_2 scenario from the MPI-GE and use the ensemble standard deviation as an estimator of internal variability. After 120 years, nearly all ensemble members show a completely ice-free Arctic in September (figure B1a). The internal variability increases from model year 1 to year 80, before it sharply drops reaching zero around year 120 when all sea ice is lost (figure B1b).

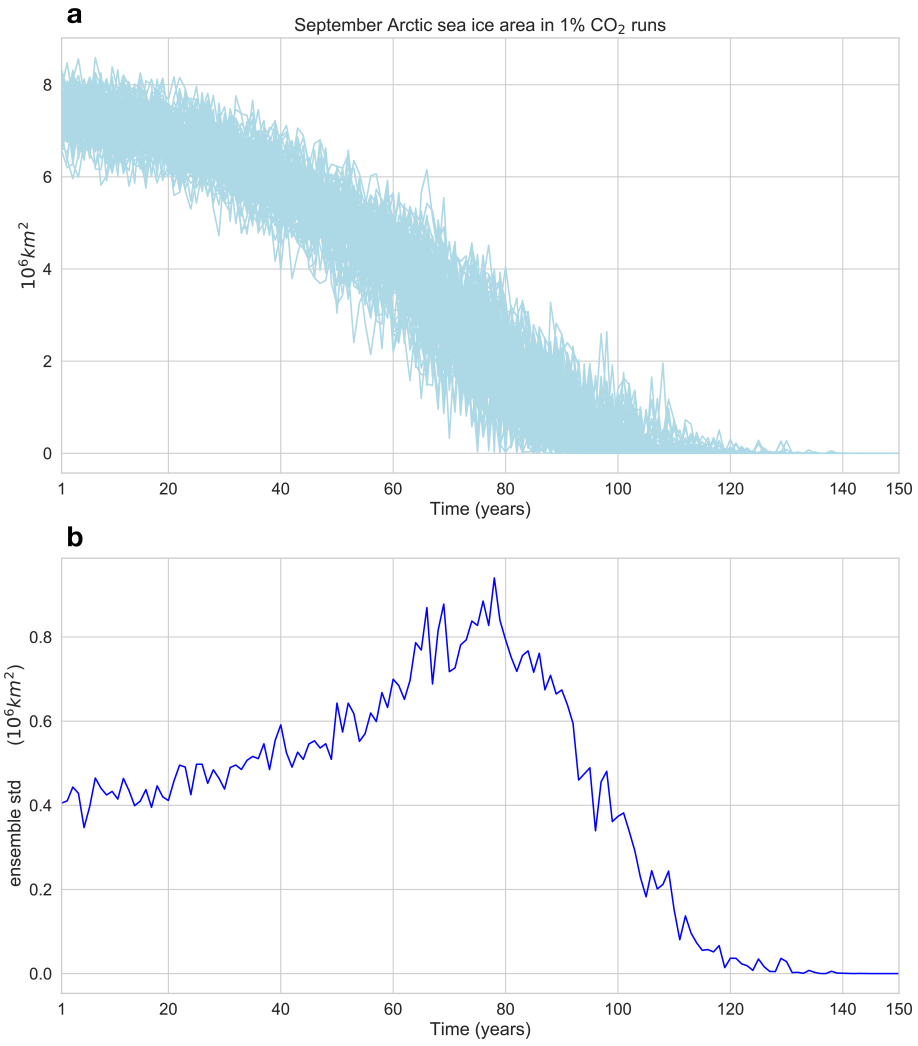


Figure B1. a) September Arctic sea ice area in the 100 realisations for the 1% CO_2 experiment. b) ensemble standard deviation for the 100 realisations.