https://doi.org/10.5194/esd-2019-70
Preprint. Discussion started: 29 November 2019

Earth System Dynamics
Discussions

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 temperature, local surface temperature and precipitation and variability in the ENSO region and central America. 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 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 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 currently available for individual global coupled climate models largely differs. The single-model ensembles within the Coupled Model Intercomparison Project Phase 5 (CMIP5) 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 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 (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 where possible. This approach allows us to estimate the required ensemble size for a specific model even if no large ensemble is available for the model. This objective approach can also help to allocate

https://doi.org/10.5194/esd-2019-70
Preprint. Discussion started: 29 November 2019

Earth System Dynamics
Discussions

resources more efficiently (Ferro et al., 2012) and inform the modelling community how many ensemble members are desirable for CMIP models.

One of the most common applications of single-model large ensembles is to separate a forced response due to 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 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 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.

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. (2015) 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 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, but up to 40 members are necessary to identify a response at high northern latitudes. However, their target is to detect a change that is different from zero, but not to quantify it. Quantifying the magnitude of the forced response may require an even larger ensemble size.

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 is sufficient to quantify internal variability and no large ensemble is required. Thompson et al. (2015) argue that the pre-industrial control run can be used to 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 over time.

https://doi.org/10.5194/esd-2019-70
Preprint. Discussion started: 29 November 2019

Earth System Dynamics Discussions

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 a forced change in ENSO or a robust difference between two models.

These studies demonstrate that different applications require different ensemble sizes. But they also 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. The required or ideal ensemble size is not only dependent on the model used, but also 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) How many ensemble members are required to identify the response to external forcing? (Section 4.1) (ii) How many ensemble members are required to adequately sample the spectrum of internal variability? (Section 4.2) (iii) 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)?

2 The resampling problem

The main difficulty when determining the required ensemble size for a specific question is resampling: in this study, 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. 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 forced response and the random internal variability in a time-series. Each realisation from a large ensemble experiences 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 (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.

https://doi.org/10.5194/esd-2019-70
Preprint. Discussion started: 29 November 2019

Earth System Dynamics
Discussions

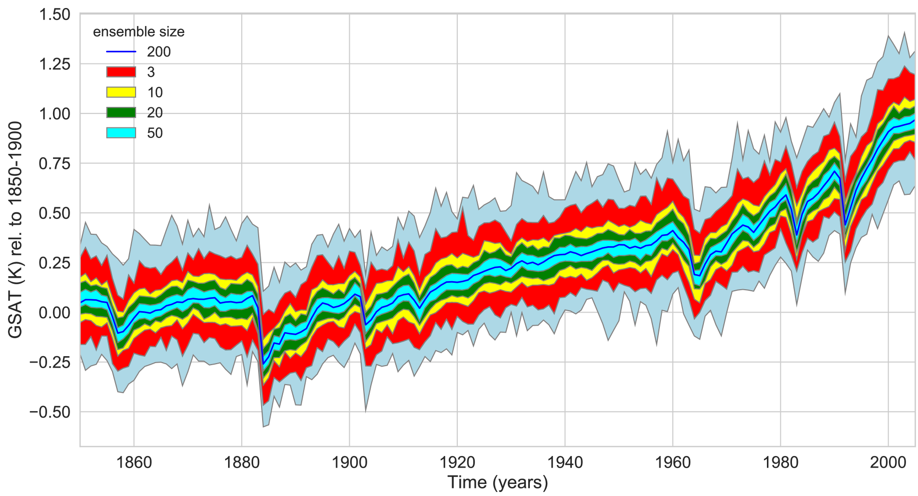

Figure 1. The forced response can be quantified using the ensemble mean in a large ensemble, while the ensemble mean of smaller ensembles is contaminated by internal variability. The figure is based on global and annual mean near-surface air temperature from the MPI-GE 200 member historical ensemble. 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, 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 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 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. While a reduction in the error with increasing ensemble size is expected, the vanishing error when using 200 members occurs by construction because we assume that the 200-member mean represents the true forced response. 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 2), 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

https://doi.org/10.5194/esd-2019-70
Preprint. Discussion started: 29 November 2019

Earth System Dynamics
Discussions

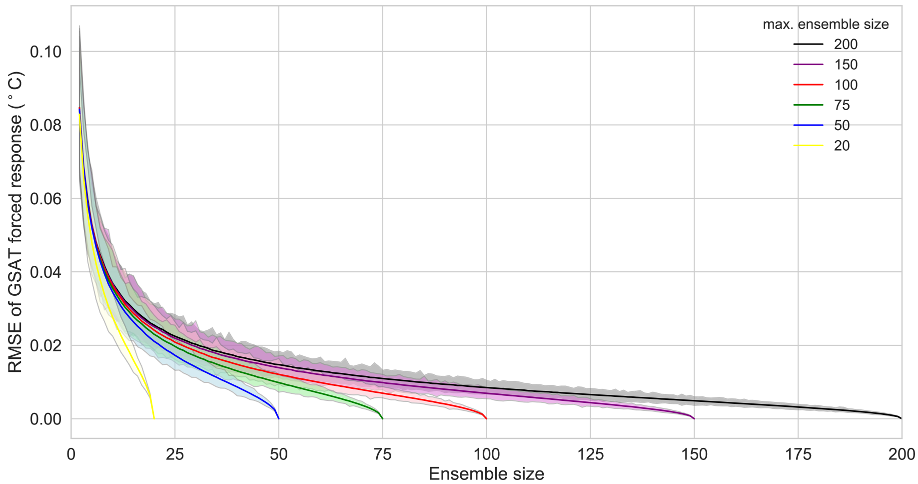

Figure 2. In a smaller ensemble, the RMSE converges to zero earlier. This is caused by resampling and does not indicate that the error is small. The black line shows the mean RMSE for GSAT for ensemble sizes from 2 to 200. The reference is the 200-member mean from figure 1 and the RMSE is computed for 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.

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 2 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 cannot be used to approximate the true error.

We find that the RMSE estimates for different maximum ensemble sizes in figure 2 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.

https://doi.org/10.5194/esd-2019-70
Preprint. Discussion started: 29 November 2019

Earth System Dynamics Discussions
Open Access
EGU

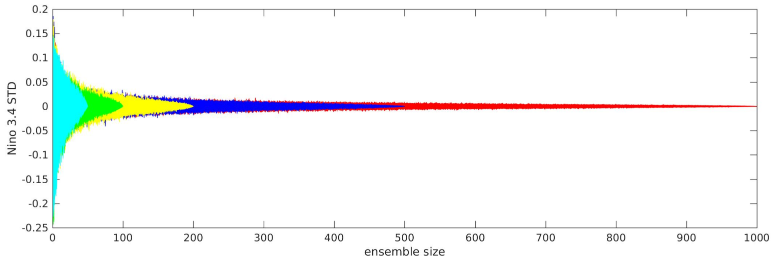

Figure 3. Using the last 1000 years of the 2000 year control simulation pdfs of the standard deviation calculated over 50 years in the Niño3.4 box (Nic to check this is correct box) are created by resampling the control simulation 1000 times. The pdfs are shown for different ensemble sizes (red: 1000members, 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.

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 3, the red 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 2, the error appears to converge when approaching the maximum ensemble size. By comparing the different maximum ensemble sizes in figure 3, 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.

3 A recipe for estimating ensemble size

We suggest the following approach to arrive at a robust estimate of the required ensemble size for any application. This method can either be applied to one of the existing large ensembles or to a long control run, which is available for all models participating in CMIP. We summarise the method in 5 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.

Earth System Dynamics Discussions

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 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 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 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 and the 100-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 50% of the maximum ensemble size, 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 4 e–h, we show how many members are necessary to estimate the forced response in near-surface air temperature for four acceptable errors that were chosen for illustrative purpose. If the acceptable error 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 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.

https://doi.org/10.5194/esd-2019-70
Preprint. Discussion started: 29 November 2019

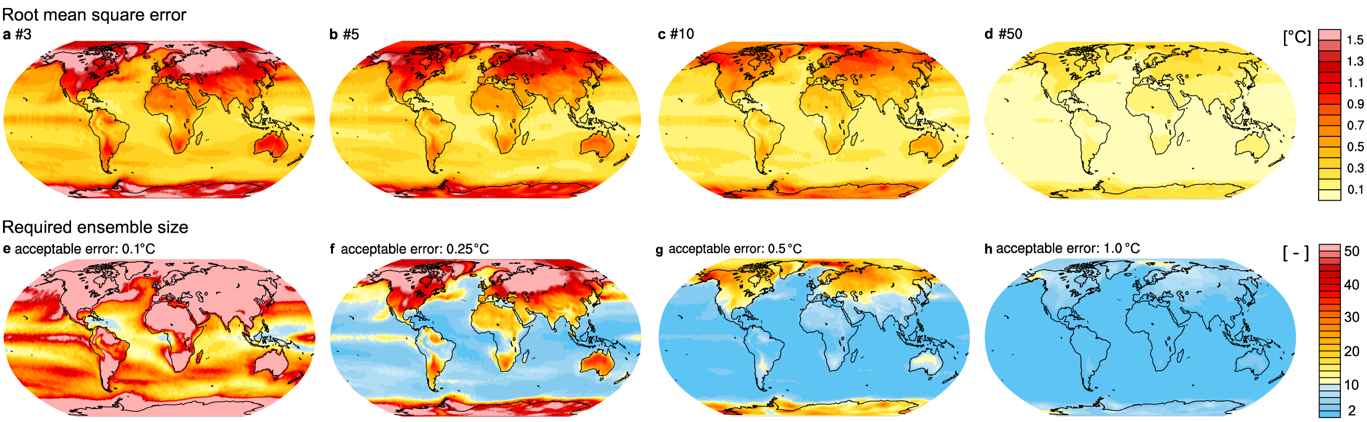

Figure 4. a-d, The mean RMSE for the forced response in historical monthly mean near-surface air temperature of MPI-GE for **a**, 3,**b**, 5, **c**, 10, and **d**, 50 ensemble members relative to the 100-member mean, globally. **e-h**, Required ensemble size to capture the 100-member mean forced response in historical monthly mean near-surface air temperature dependent on the acceptable error of **a**, 0.1, **b**, 0.25, **c**, 0.5, and **d**, 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 5 a–d). The largest errors can be found over the Indian ocean and western tropical Pacific. Similar to temperature, a 50-member ensemble shows very small errors across the globe.

In figure 5 e–h we show how many members are necessary to estimate the forced response with an acceptable error of 0.1, 0.2, 0.5, and 1 mm/day. For an acceptable error of 0.2 mm/day, many ocean regions require more than 50 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 and western tropical Pacific, while less than 10 members are sufficient elsewhere.

For the example in figures 4 - 5, the objective was to isolate the full forced response in a time series, defined as the 100-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 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 6). 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.

https://doi.org/10.5194/esd-2019-70
Preprint. Discussion started: 29 November 2019

Earth System Dynamics
Discussions

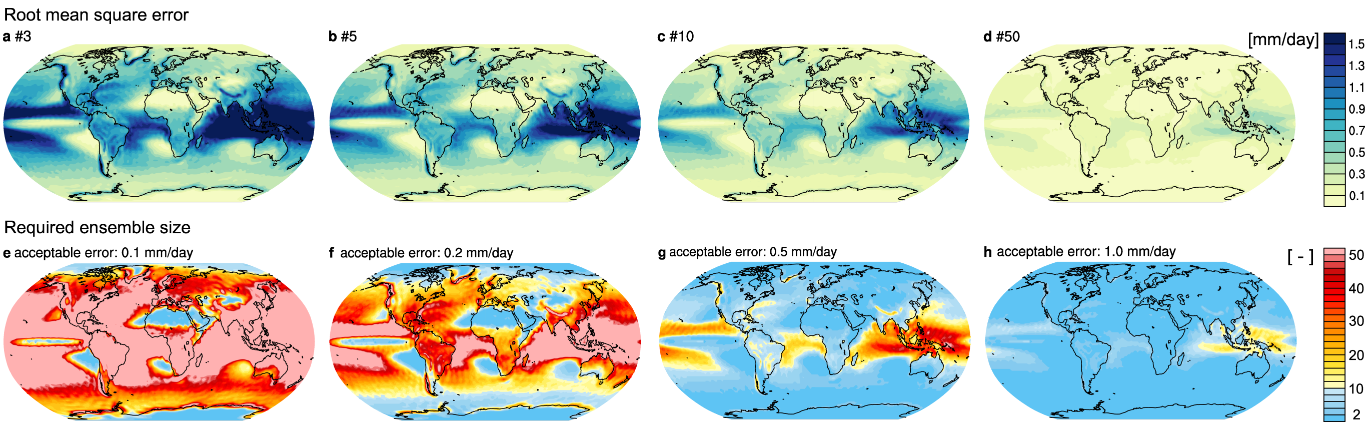

Figure 5. a-d, The mean RMSE for the forced response in historical monthly mean total precipitation of MPI-GE for **a**, 3, **b**, 5, **c**, 10, and **d**, 50 ensemble members relative to the 100-member mean, globally. **e-h**, Required ensemble size to capture the 100-member mean forced response in historical monthly mean total precipitation dependent on the acceptable error of **a**, 0.1, **b**, 0.2, **c**, 0.5, and **d**, 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 6). 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 7 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 required for the ensemble mean to capture a cooling in all samples, however, the ensemble mean cooling can still 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 7 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.

These examples demonstrate that the required sample size to estimate the forced response depends on the region and variable (figures 4-5), as well as the feature of interest in the forced response. Whereas for some applications 5 members are sufficient

https://doi.org/10.5194/esd-2019-70
Preprint. Discussion started: 29 November 2019

Earth System Dynamics
Discussions
Open Access

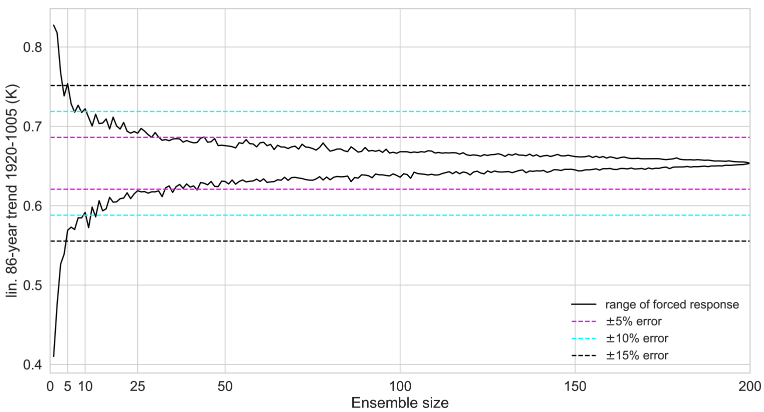

Figure 6. Linear warming trend from 1920 to 2005 for different ensemble sizes shown as a linear trend fitted to the ensemble mean. Black lines show maximum and minimum 86-year ensemble mean temperature trend from 1000 random samples. Errors are shown as percentage of the 200-member ensemble mean temperature trend.

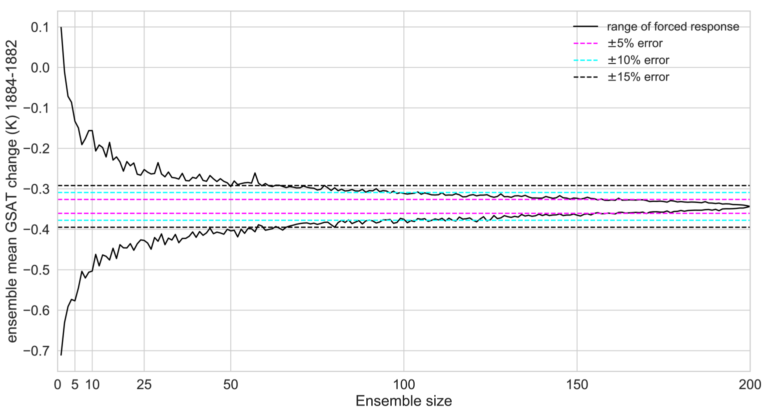

Figure 7. GSAT cooling after Krakatoa eruption for different ensemble sizes shown as the ensemble mean temperature difference between 1882 and 1884. Black lines show maximum and minimum temperature response from 1000 random samples. Errors are shown as percentage of the 200-member ensemble mean temperature response.

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

https://doi.org/10.5194/esd-2019-70
Preprint. Discussion started: 29 November 2019

Earth System Dynamics Discussions

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 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. 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 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, 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. 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 8a. To generate synthetic ensemble members, we split the pre-industrial control 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 8a 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 8b) 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 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

https://doi.org/10.5194/esd-2019-70
Preprint. Discussion started: 29 November 2019

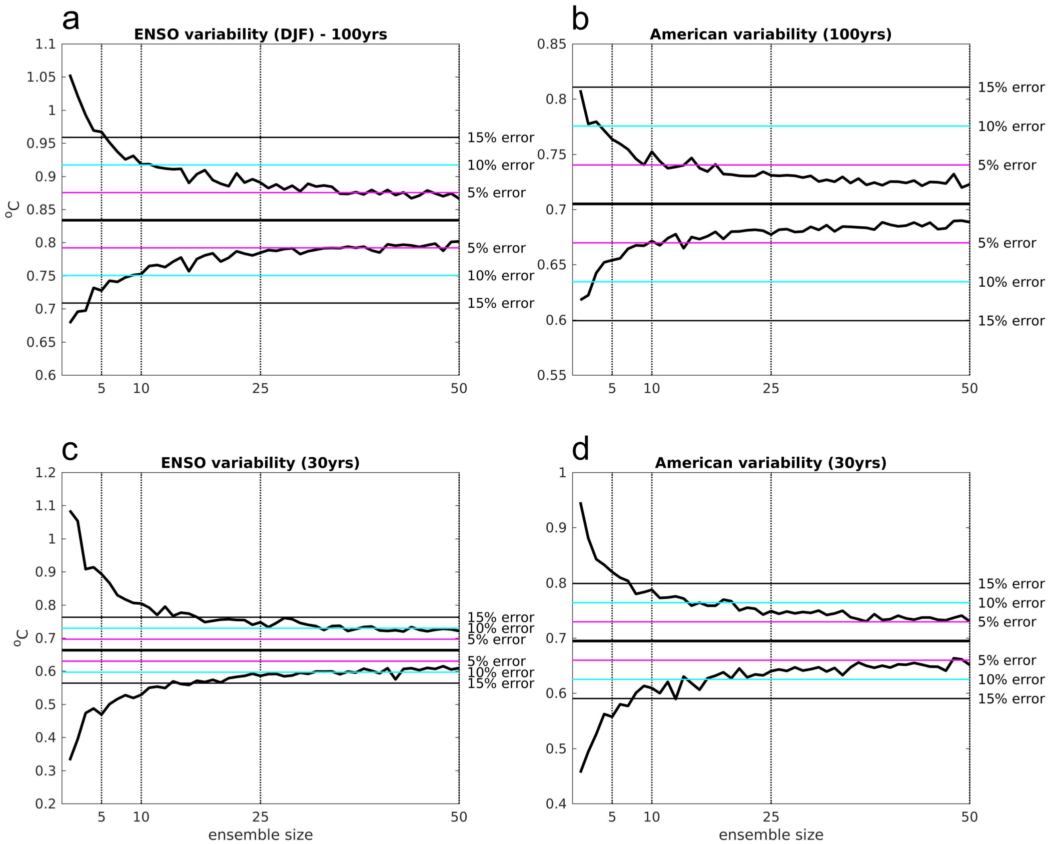

Figure 8. We show for increasing ensemble sizes the: a) ENSO variability in the Niño3.4 box calculated over 100 year periods, b) Central American variability calculated over 100 year periods, c) ENSO variability in the Niño3,4 box calculated over 30 year periods, d) Central American variability calculated 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 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.

be problematic if the objective is to identify a change in variability, such as changes in ENSO characteristics under global warming.

https://doi.org/10.5194/esd-2019-70
Preprint. Discussion started: 29 November 2019

Earth System Dynamics
Discussions

4.3 Quantifying changes in internal variability

To quantify changes in 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 A1a). The internal variability increases from model year 1 to year 80, before it sharply drops reaching zero around year 120 (figure A1b).

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 9 we show 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.

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 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 8). 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.

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

https://doi.org/10.5194/esd-2019-70
Preprint. Discussion started: 29 November 2019

Earth System Dynamics
Discussions

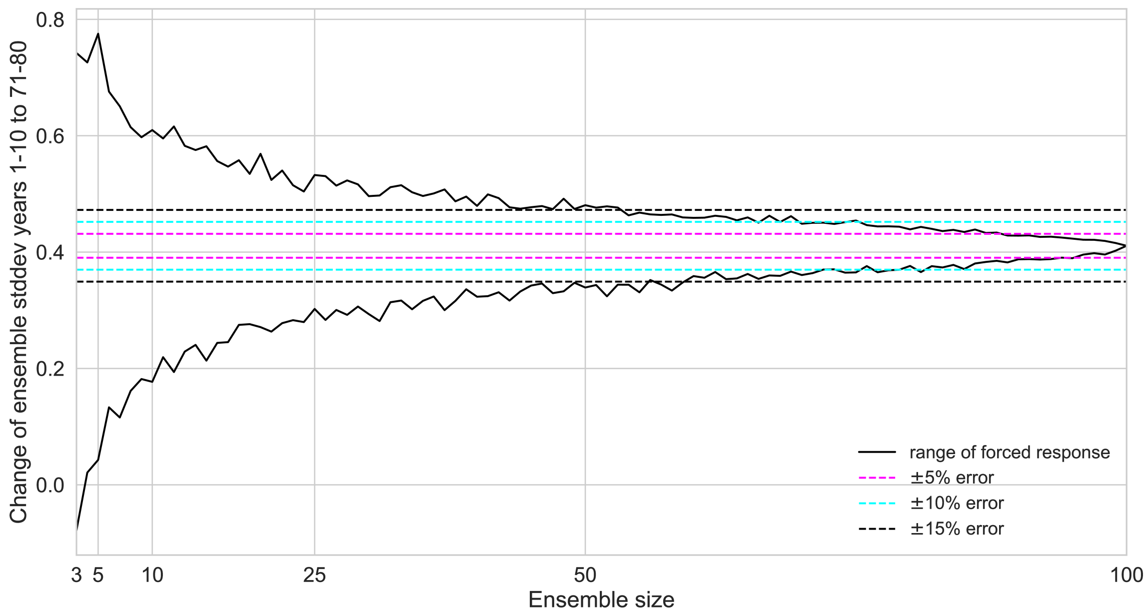

Figure 9. Change in internal variability of September Arctic sea ice are from the first decade to years 71–80 in a 1% CO_2 experiment. For different ensemble sizes, we compute the ensemble standard deviation and then average for the first decade and years 71–80 before computing the difference. Black lines show maximum and minimum change in variability from 1000 random samples. Errors are shown as percentage of the 100-member variability change.

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.

Here we present a generalised approach to estimate the ensemble size that is required to robustly estimate a model's characteristics. We differentiate three types of question: identifying a forced response, quantifying variability, identifying a change in 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 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 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 2000-year pre-industrial control simulation from the MPI-GE to estimate required ensemble sizes for the MPI-ESM model.

https://doi.org/10.5194/esd-2019-70
Preprint. Discussion started: 29 November 2019

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 members could show a warming after the volcanic eruption, while the true 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 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 if changes in variability are indeed robust within each model.

The examples in this study show that for some applications ensemble sizes around 5 members are sufficient while other applications require ensemble sizes well above 100 members. This information 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.

https://doi.org/10.5194/esd-2019-70
Preprint. Discussion started: 29 November 2019

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.

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, 2015.

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.

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 and the Max Planck Society for the Advancement of Science for funding all three authors. DO was supported by the European Union's Horizon 2020 Research and Innovation Programme under grant agreement number 820829 (CONSTRAIN).

Appendix A

https://doi.org/10.5194/esd-2019-70
Preprint. Discussion started: 29 November 2019

Earth System Dynamics
Discussions
Open Access

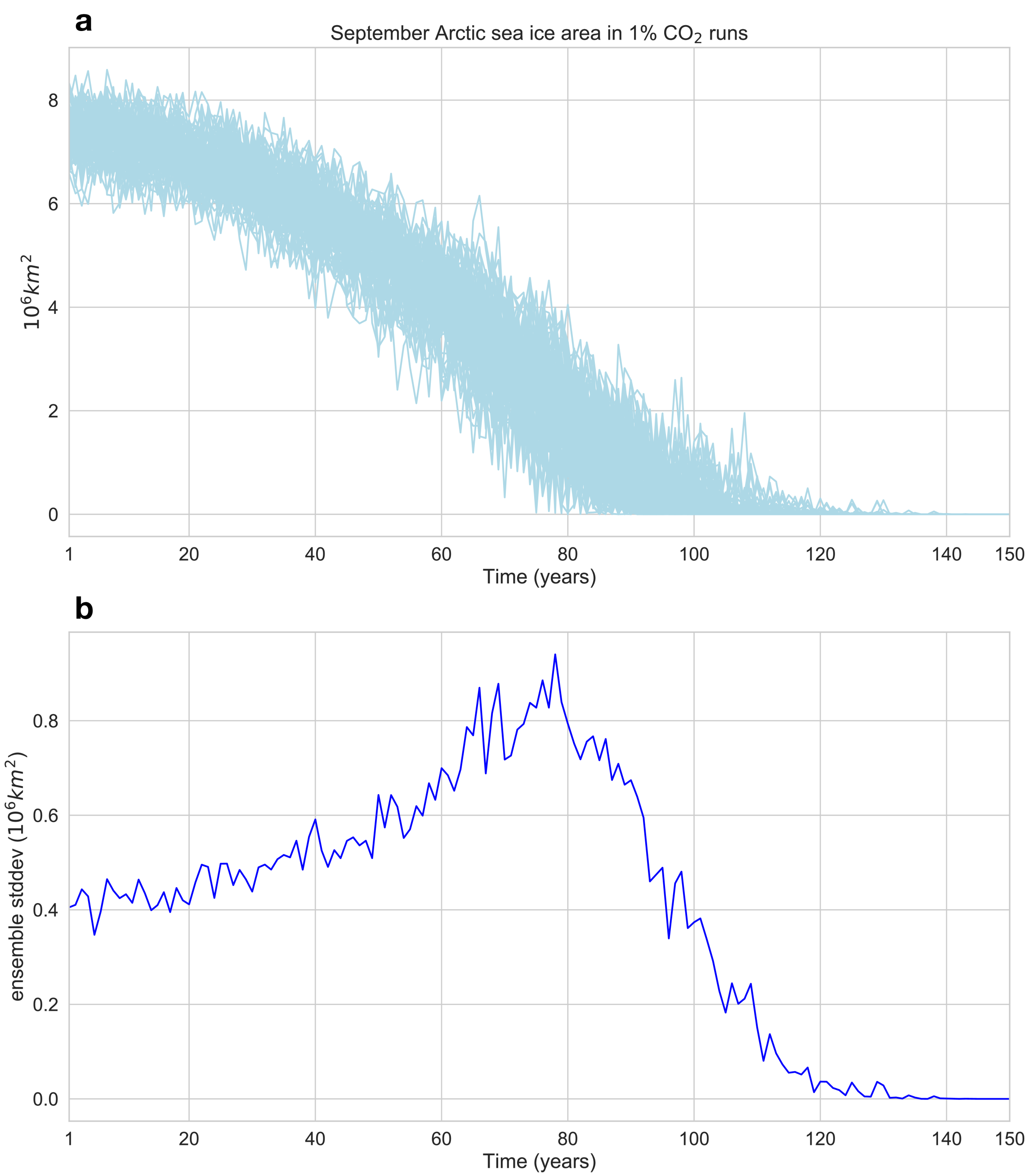

Figure A1. a) September Arctic sea ice area in the 100 realisations for the 1% CO_2 experiment. b) ensemble standard deviation for the 100 realisations.