Please find below our responses to reviewer n1 and n2. We have done our best to address the reviewer comments and clarify any outstanding issue.

Responses to Referee n2

This paper is much improved and is now acceptable. Two minor comments below.

We are pleased that the referee is satisfied with our improved version and would like to thank the referee for the helpful comments improving our manuscript.

- ### Specific Comments

1) "p8. line 27-28. Which MIPs might be affected by this?"

We have added examples of MIPs recommending 100 years or less as minimum run length for their experiments.

For instance, the CMIP6 minimum run-length requirement for a few of the Model Intercomparison Projects (MIPs), excluding DECK and Historical simulations, is 100 years or less and ensembles are not always requested (e.g., some of the Tier 1/2/3 experiments in PMIP (Otto-Bleisner et al., 2017), nonlinMIP (Good et al., 2016), GeoMIP (Kravitz et al., 2015), High-ResMIP (Haarsma et al., 2016), FAFMIP (Gregory et al., 2016)).

2) "p9. line 14. Wittenberg et al (2009. 10.1029/2009GL038710) suggests even longer would be necessary."

The reference was added at page 8, line 14:

As ENSO provides a medium-frequency modulation of the climate system, it is not surprising that it takes longer than 100 years for its variability to be fully represented (see e.g., Wittenberg et al., 2009).

Responses to Referee n1

The article by Maria-Vittoria Guarino, Louise C. Sime, David Schroeder, Grenville M. S. Lister and Rosalyn Hatcher is about estimating the potential changes in the climate simulated by a climate model when run on different computers, and the influence of internal variability in this type of study.

Overall, the paper has largely improved compared with the first version. However, I still find the same tendency along the manuscript for over-interpretation of the results, too many suggestions and not enough clear use and interpretation of the actual results, not enough precision and details on the methodology, and even worse, a wrong initial statement as the starting point of section 4.2 (minimum length for a piControl DECK simulation in CMIP6 of is 100 years when it is actually 500 years), that is the base of one of the main conclusions of the paper. Unfortunately, following those elements, I still not recommend the manuscript for publication in GMD.

We thank the referee for acknowledging the improvements and for his/her censorious attitude towards our manuscript, which is really helpful in eliminating some remaining misunderstandings. However, the referee is mistaken in stating that our initial statement about minimum length is wrong. Please find below our response to the specific comments. We added some details to methodology as suggested and made a few minor changes to the manuscript to avoid misunderstandings.

- **Specific Comments**

1) "Page 4, lines 9-10: 'all' the CMIP6 experiments are not analysed against the piControl. Replace 'all' with 'many'. "

Changed as requested.

2) Page 5, lines 11-13: the sentence "However this method cannot detect code bugs, which may cause a model to behave differently on different machines" is incorrect. First, the method you present is suited to check the resolution of the equations of the model on a period shorter than one day. It could thus detect what you call 'bugs'. And second: what are you actually calling bugs? In your study you are trying to detect whether the resolution of the equations of the model on two different machines could end up with two different climates. But which one is the right one? In case you find that the simulations differ in some way, is there a way to say that one is correct and the other is not? I suggest to replace this sentence with: "However this method is restricted to time scales shorter than one day. The centennial simulations presented in this paper will help understanding whether or not differences can arise on longer time scales in the HadGEM3-GC3.1 model."

The bugs we intended are compiler bugs, i.e. different compilers can interpret a same line of code differently. The method described in the paper, used to test the porting of the HadGEM3 model, is targeted to identify errors such as round-off, computation order, IEEE arithmetic and basic library function errors (and etc.) that would cause the solution to diverge in the two cases immediately. However, errors resulting from, for example, divisions by zeros (which might or might not occur in the first 24 hours of simulations, and could be interpreted differently depending on the compiler), or random seed initialization in parameterization schemes might occur later on in the simulation.
We accept the referee's suggestion and have replaced the sentence at page 5, line 11-13, as we do not think that additional clarifications on 'compiler bugs' would benefit the paper's scientific discussion.

However this method is restricted to time scales shorter than one day. The centennial simulations presented in this paper will help understanding whether or not differences can arise on longer time scales in the HadGEM3-GC3.1 model.

3) Page 6, line 8: I still don't get why you divide your SNR by sqrt(2). I'm not saying that it's not relevant, I just don't find any explicit reason, supported by a reference, or a demonstration, to explain this choice in your manuscript. I ask for clarification in the text (lines 14-15 are not enough, need a reference), especially because you say that MO-AR differences are outside the internal variability range for values greater than one (and not 0.9, or 1.1).

The assumption used in the paper is based on one of the basic properties of variance (i.e. being a not linear operator), the mathematical demonstration is given at the end of this document for readability (please see below). We now mention this at page 6, line 8. The sentence has also been modified for improved clarity and a reference, as requested, has been added:

The signal is represented by the mean of the differences between PI_MO and PI_AR (μ_{MO-AR}) and the noise is represented by the standard deviation of PI_MO (σ_{MO}), our "reference" simulation. Because of the basic properties of variance, for which $Var_{X-Y} = Var_X + Var_Y - 2Cov(X, Y)$ (Loeve, 1977), we can more conveniently express the noise as $\sigma_{MO} = \sigma_{MO-AR}/\sqrt{2}$, under the assumptions that PI_MO and PI_AR are uncorrelated ($Cov(MO, AR) = 0$), and have same variance ($Var_{MO} = Var_{AR}$). This allowed us to compute SNR on one same grid, and avoid divisions by (nearly) zero when the sea ice field between PI_MO and PI_AR evolved differently, resulting in unrealistically high SNR values along the sea ice edges.
Finally, SNR is defined as:

$$SNR = \frac{\mu_{MO-AR}}{\sigma_{MO}} = \frac{\mu_{MO-AR}}{\sigma_{MO-AR}/\sqrt{2}}$$

A signal-to-noise ratio larger than 1 is commonly associated to the existence of a physical process. This is because such condition certainly implies that the signal (the mean) is larger than the noise (the standard deviation). However, we agree that using a net cut-off value of 1 can be, in itself, unphysical. In fact, values very close to 1 might still indicate an 'emerging' signal.
A short discussion of why also values close to 0.9 might be important was present in the first submitted version of the manuscript. However, at referee's recommendation, we modified the discussion so to focus only on values larger than 1. Please see also our previous "Responses to referees" document, responses to Referee n1, specific comments 6 and 7, about this.

4) Page 6, lines 9-12: please reconsider this paragraph with this proposition: "When SNR<1, MO-AR differences can be interpreted as fluctuations within the estimated range of internal variability. When SNR>1, the MO-AR differences in mean are outside the expected range of internal variability. It means that we either evidenced a true difference in mean, or that the estimated range of variability is underestimated".

The sentence at page 6, line 9-12, was modified following the reviewer's suggestion:

When SNR < 1, (PI_MO - PI_AR) differences can be interpreted as fluctuations within the estimated range of internal variability. When SNR > 1, (PI_MO - PI_AR) differences in the mean are outside the expected range of internal variability. This eventuality indicates either a true difference in the mean, or that the expected range of variability is underestimated.

5) Page 7, lines 13-19: the simple and direct way to explain the behaviour of your results is:
- on decadal time scale, the period is too short to adequately sample the longer time scales of the interannual variability; therefore the estimated mean is not stable, and the estimated standard deviation of the simulation is likely underestimated compared with the true standard deviation of the internal variability of the model; it is thus not surprising to have values higher than one when analysing decadal periods

- on longer time scales, the estimate of the mean and standard deviation converge toward their ‘true’ values. Accordingly, we see that the differences between and MO and AR become smaller.
- For the 200-year long period, we find no value greater than one. Following this diagnostic, and for the variables we assessed, the results show that there is no significant difference in mean simulated with HadGEM3-GC3.1 on MO and AR.
And this is valid only for the mean, and for the variables considered. You can thus reconsider your last sentence (line 18-19) by saying that “Our results show that there is no difference in mean when considering a 200-year long period between AR and MO”. Your suggestion is that “the overall physical behaviour of the model has not been affected by the porting” is premature regarding your analyses.

We accept the suggestions and paragraphs at page 7 have been modified as follows:

On decadal timescales, the averaging period is too short to adequately sample the model interannual variability; therefore the estimated mean is not stable, and the estimated standard deviation is likely to be underestimated compared with the true standard deviation of the model internal variability. Large differences in the mean and a SNR >>1 are, thus, not surprising when analysing decadal periods.
On longer timescales, the estimate of the mean and standard deviation converge toward their `true' values. Accordingly, we see that the differences in the mean between PI_MO and PI_AR become smaller and approach zero (Figure 4d to 9d).
When considering the 200-year long-term mean, we find no SNR value greater than one (Figure 2 and 3). Following this diagnostic, and for the variables we assessed, our results show that there is no significant difference in the simulated mean between the two PI_MO and PI_AR HadGEM3-GC3.1 simulations when considering a 200-year long period.

6) Page 8, lines 16-17: I reject your conclusion that “simulations using the HadGEM3-GC3.1 model are reproducible […] long-enough simulation in length is used”. You can only conclude from your analyses that the mean of the variables assessed is not different between MO and AR for your piControl simulations. Actually showing that the model is reproducible would require that your diagnostics provide an exhaustive description of the model physical behaviour, not only mean, but also variability, teleconnection patterns, trends... You can say that your analyses do not show that the model is not reproducible, and that’s already a valuable information, that has not been provided by all the modelling centres (and you should receive credit for this).

We take the opportunity to stress that in the paper we present mean and standard deviation (Table 2) for the considered variables, and that teleconnection patterns are indeed discussed, when results point towards a probable change in their mean characteristics (see section 4.2 about the 100-year timescale). However, as more could be done to fully characterize the physical behaviour of the model, we reformulated our sentence at page 8 as follows:

We thus conclude that the mean climate properties simulated by the HadGEM3-GC3.1 model are reproducible on different HPC platforms, provided that a long-enough simulation length is used.

7) Page 8, line 22: I propose “The large differences observed on time-scales shorter than 200-years are a direct consequence of the (potentially underestimated) internal variability of the model, triggered (at least initially) by the machine-dependent processes (compiler […] 3.1 for details).”

The sentence at page 8, line 22, was modified to take into account the reviewer's suggestion:

> The large differences observed on time-scales shorter than 200-years are a direct consequence of the (potentially underestimated) internal variability of the model, and triggered (at least initially) by machine-dependent processes (compiler, machine architecture etc., see section 2 and 3.1 for details).

. 8) Page 8, line 27: this is a major point: because the analyses presented in your manuscript concern piControl runs performed with a fully-coupled GCM (typically one of the DECK experiments, see https://www.geosci-model-dev.net/9/1937/2016/gmd-9-1937-2016.pdf), I assume that your are talking about the minimum length of the piControl run in the DECK (if not, then you really need to add more details to be more specific). The actual minimum length for a piControl run in the CMIP6 DECK is 500 years. Therefore your section starts with a wrong statement, which is a pretty serious mistake. Indeed, your results suggest that 100 years may not be enough to fully sample the internal variability of HadGEM3-GC3.1. The good news regarding this statement is that the CMIP6 protocol asked for 500 years. The bad news is that several CMIP6 models have much more internal variability than their previous CMIP5 versions, and that 500 years might not be enough. But this is another story. Also, your conclusions and suggestions have a priori no reason to be applicable to other experiments/MIPs, such as AMIP, historical runs, scenarios, etc... if you want to make on any other given experiment/MIP, be more specific.

In this paragraph we refer to individual Model Intercomparison projects (MIPs), and not to DECK (which the PI control run belongs to) or Historical simulations. This terminology is confirmed by the same reference the reviewer provides in their comment, where the structure of CMIP6 (DECK + Historical + MIPs) is presented and explained.

The minimum simulation length required by many MIPs is 100 years or less. At page 8, line 29-32, we have now added a list of individual MIPs (with references) that recommend 100 years or less, and added a sentence to clarify that we do not refer to DECK or Historical simulations (see response to referee n2, point 1). In relation to the above, please see also the HighMIP documentation paper (https://www.geosci-model-dev.net/9/4185/2016/gmd-9-4185-2016.pdf) where they say: *"The future end-date is based on a compromise between what is computationally affordable by a sufficient number of centres (~100 years of integration) and what is scientifically relevant."*

The relevance of our results to MIPs is explained in Introduction (p1, lines 17-20) and in the Conclusions (p10, lines 26-31). As an example, two of the authors of the present manuscript are part of the PMIP4 community, the analysis this study is based on was of crucial importance to the UK PMIP community in order to decide how long Tier 1 and 2 experiments should be.

In the manuscript, it is not stated that PI control simulations should be run for longer than 100 years, but is shown that, for comparison purposes (with the PI control run), other MIPs experiments should be run for at least 200 years when possible. This will assure that the HadGEM3 internal variability is sampled correctly, and that differences in means due to a wrong sampling will not be confused with system responses to a different climate forcing. The paragraph at page 10, lines 26-31, is reformulated to help clarify:

This result has immediate implications for those members of the UK CMIP6 community who will run individual MIP experiments on the ARCHER HPC platform, and will compare results against the reference PI simulation run on the MO platform by the UK Met Office. The magnitude of (PI_MO - PI_AR) differences presented in this paper should be regarded as threshold values below which differences between ARCHER and MO simulations must be interpreted with caution (as they might be the consequence of a wrong sampling of the model internal variability rather than the climate response to a different forcing).

. 9) Page 8, lines 4-14: I don't understand where you want to go with the 2/3 power law, although the result is surprisingly consistent among the variables. And the "plateau" you describe is supported by three consecutive points on your plots on figure 10, the last one being slightly higher than expected by the line. I would agree that there is a plateau if it was described by more than one single point being higher than expected. And once more, you conclude that this results "suggests" something. I would recommend using your results to "show" things, and stick to what they actually show.

We agree that longer simulations, resulting in more data points in Figure 10 and Figure 4d to 9d, would allow a better visualization of the plateau. However, we show that, as the time-scale increases, data points in Figure 10 vary following a power law relationship. (PI_MO – PI_AR) differences are very small and close to zero at the 200-year timescale (Figure 4d to 9d). Since they are expected to become even smaller (closer and closer to zero) with time, the trend exhibited by the data is to eventually plateau at zero for timescales ≥ 200 years.

These results are not described as "the data plateaus at the 200-year timescale", rather as "approaches a plateau near the 200-year time-scale" (page 8, line 9). This reflects the reviewer's comment regarding just one point being above the line in Figure 10.

Figure 10 provides an alternative/additional way to quantify the HadGEM3 model behaviour on the two HPC platforms: this analysis tells us that, not only the considered variables converge to their true value at the 200-year timescale (Figure 4d – 9d), but that the rate at which this happens is the same for all variables at the global scale (which is not immediate when you look at Figure 4d - 9d). As the reviewer recognized, this behaviour is remarkably consistent among all the considered variables and is thus worth mentioning.

See below changes at page 8, lines 7-11, and page 10, lines 12-14, to take into account the reviewer's comment:

Thus, the straight-lines that best fit the global mean data in Figure 10 have a slope of ~ 2/3. The existence of a ~ 2/3 power law, which does not depend on the single quantity, shows a consistent scaling of PI_MO - PI_AR) differences with the timescale that approaches a plateau near the 200-year timescale (note that an actual plateau can only be reached for longer simulations, as differences computed over all timescales longer than 200 years would be ~ 0).

There exist a consistent time-dependent scaling of (PI_MO - PI_AR) differences across the whole climate simulation, so that variables converge toward their true values at the same rate, independently on the physical processes that they represent.

Mathematical demonstration of $\boldsymbol{STDEV_X = STDEV_{X-Y} \ / \ \sqrt{2}}$

We use the basic property of variance for which:

$$Var_{X-Y} = Var_X + Var_Y - 2Cov(X,Y)$$

If variables are uncorrelated (i.e. independent), so that $ov(X,Y) = 0$, and have the same variance ($Var_X = Var_Y$) :

$$Var_{X-Y} = 2Var_X$$

As the square root of the variance is the standard deviation:

$$STDEV_{X-Y} = \sqrt{2}\,STDEV_X \ \rightarrow \ STDEV_X = STDEV_{X-Y}/\sqrt{2}$$

Machine dependence and reproducibility for coupled climate simulations: The HadGEM3-GC3.1 CMIP Preindustrial simulation

Maria-Vittoria Guarino[1], Louise C. Sime[1], David Schroeder[2], Grenville M. S. Lister[3], and Rosalyn Hatcher[3]

[1]British Antarctic Survey, Cambridge, UK
[2]Department of Meteorology, University of Reading, Reading, UK
[3]National Centre for Atmospheric Science, University of Reading, Reading, UK

Correspondence: Maria-Vittoria Guarino (m.v.guarino@bas.ac.uk)

Abstract.

When the same weather or climate simulation is run on different High Performance Computing (HPC) platforms, model outputs may not be identical for a given initial condition. While the role of HPC platforms in delivering better climate projections is to some extent discussed in literature, attention is mainly focused on scalability and performance rather than on the impact of machine-dependent processes on the numerical solution.

Here we investigate the behaviour of the Preindustrial (PI) simulation prepared by the UK Met Office for the forthcoming CMIP6 under different computing environments.

Discrepancies between the means of key climate variables were analysed at different timescales, from decadal to centennial. We found that for the two simulations to be statistically indistinguishable, a 200-year averaging period must be used for the analysis of the results. Thus, constant-forcing climate simulations using the HadGEM3-GC3.1 model are reproducible on different HPC platforms provided that a long-enough duration of simulation is used.

In regions where ENSO teleconnection patterns were detected, we found large sea surface temperature and sea ice concentration differences on centennial time-scales. This indicates that a 100-year constant-forcing simulation may not be long enough to adequately capture the internal variability of the HadGEM3-GC3.1 model, despite this being the minimum simulation length recommended by CMIP6 protocols.

On the basis of our findings, we recommend a minimum simulation length of 200 years whenever possible.

1 Introduction

The UK CMIP6 (Coupled Model Intercomparison Project Phase 6) community runs individual MIP experiments on differing computing platforms, but will generally compare results against the reference simulations run on the UK Met Office platform. For this reason, within the UK CMIP community, the possible influence of machine dependence on simulation results is often informally discussed among scientists, but yet surprisingly an analysis to quantify its impact has not been attempted.

The issue of being able to reproduce identical simulation results across different supercomputers, or following a system upgrade on the same supercomputer, has long been known by numerical modellers and computer scientists. However, the impact that a different computing environment can have on otherwise identical numerical simulations appears to be little known by climate models users and model data analysts. In fact, the subject is rarely ever addressed in a way that helps the community understand the magnitude of the problem, or to develop practical guidelines that take account of the issue.

To the extent of our knowledge, only a few authors discussed the existence of machine dependence uncertainty and highlighted the importance of bit-for-bit numerical reproducibility in the context of climate model simulations. Song et al. (2012) and Hong et al. (2013) investigated the uncertainty due to the round-off error in climate simulations. Liu et al. (2015b) and Liu et al. (2015a) discussed the importance of bitwise identical reproducibility in climate models.

In this paper, we investigate the behaviour of the UK CMIP6 Preindustrial (PI) control simulation with the HadGEM3-GC3.1 model on two different High Performance Computing (HPC) platforms. We first study whether the two versions of the PI simulation show significant differences in their long-term statistics. This answers our first question of whether the HadGEM3-GC3.1 model gives different results on different HPC platforms.

Machine-dependent processes can influence the model internal variability by causing it to be sampled differently on the two platforms (i.e. similarly to what happens to ensemble members initiated from different initial conditions). Therefore, our second objective is to quantify discrepancies between the two simulations at different time-scales (from decadal to centennial) in order to identify an averaging period/simulation length for which the two simulations return the same internal variability.

Note that the PI control simulation is a constant-forcing simulation. Therefore, no ensemble members are required for such experiment because, provided that the simulation is long enough, it will return a picture of the natural variability.

The remainder of the paper is organized as follows. In section 2, mechanisms by which the computing environment can influence the numerical solution of chaotic dynamical systems are reviewed and discussed. In section 3, the numerical simulations are presented and the methodology used for the data analysis is described. In section 4, the simulation results are presented and discussed. In section 5, the main conclusions of the present study are summarized.

2 The impact of machine dependence on the numerical solution

In this section, possible known ways in which machine-dependent processes can influence the numerical solution of chaotic dynamical systems are reviewed and discussed.

Different compiling options, degrees of code optimization and basic library functions all have the potential to affect the reproducibility of model results across different HPC platforms, and on the same platform under different computing environments. Here we provide a few examples of machine-dependent numerical solutions using the 3D Lorenz model (Lorenz, 1963), which is a simplified model for convection in deterministic flows. The Lorenz model consists of the following three differential equations:

$$\begin{aligned} \frac{dx}{dt} &= \alpha(y-x) \\ \frac{dy}{dt} &= \gamma x - y - zx \\ \frac{dz}{dt} &= xy - \beta z \end{aligned} \tag{1}$$

where the parameters $\alpha = 10$, $\gamma = 28$ and $\beta = 8/3$ were chosen to allow the generation of flow instabilities and obtain chaotic solutions (Lorenz, 1963). The model was initialized with $(x_0, y_0, z_0) \equiv (1, 1, 1)$ and numerically integrated with a 4th-order Runge-Kutta scheme using a time step of 0.01. The Lorenz model was run on two HPC platforms, namely: the UK Met Office Supercomputer (hereinafter simply "MO") and ARCHER.

To demonstrate first the implications of switching between different computing environments, the Lorenz model was run on the ARCHER platform using:

- two different FORTRAN compilers (cce8.5.8 and intel17.0), see Figure 1a and 1b;
- same FORTRAN compiler (cce8.5.8) but different degrees of floating-point optimization (`-hfp0` and `-hfp3`), see Figure 1c and 1d;
- same FORTRAN compiler and compiling options but the x-component in (1) was perturbed by adding a noise term obtained using the `random_number` and `random_seed` intrinsic FORTRAN functions. In particular, the seed of the random number generator was set to 1 and 3 in two separate experiments, see Figure 1e and 1f.

Finally, to illustrate the role of using different HPC platforms, the Lorenz model was run on the ARCHER and MO platforms using the same compiler (intel17.0) and identical compiling options (i.e. level of code optimization, floating-point precision, vectorization) (Figure 1g and 1h).

The divergence of the solutions in Figure 1a and 1b can likely be explained by the different 'computation order' of the two compilers (i.e. the order in which a same arithmetic expression is computed). In Figure 1c and 1d, solutions differ because of the round-off error introduced by the different precision of floating-point computation. In Figure 1e and 1f, the different seed used to generate random numbers caused the system to be perturbed differently in the two cases. While this conclusion is straightforward, it is worth mentioning that the use of random numbers is widespread in weather and climate modelling. Random number generators are largely used in physics parametrizations for initialization and perturbation purposes (e.g. clouds, radiation and turbulence parametrizations) and, as obvious, in stochastic parametrizations. The processes by which initial seeds are selected within the model code are thus crucial in order to assure numerical reproducibility. Furthermore, different compilers may have different default seeds.

As for Figure 1g and 1h, this is probably the most relevant result for the present paper. It highlights the influence of the HPC platform (and of its hardware specifications) on the final numerical solution. In Figure 1g and 1h the two solutions diverge in time similarly to Figure 1a - 1d, however identifying reasons for the observed differences is not straightforward. While we speculate that reasons may be down to machine architecture and/or chip-set, further investigations on the subject were not pursued as this would be beyond the scope of this study.

The three mechanisms discussed above were selected because illustrative of the problem and easily testable via a simple model such as the Lorenz model. However, there are a number of additional software and hardware specifications that can influence numerical reproducibility, and that only emerge when more complex codes, like weather and climate models, are run. These are: number of processors and processor decomposition, communications software (i.e. MPI libraries), threading (i.e. OpenMP libraries).

We conclude this section stressing that the four case studies presented in Figure 1 (and the additional mechanisms discussed in this section) are all essentially a consequence of the chaotic nature of the system. When machine-dependent processes introduce a small perturbation/error into the system (no matter by which mean), they cause it to evolve differently after a few time-steps.

3 Methodology

3.1 Numerical simulations

In this study, we consider two versions of the Preindustrial PI control simulation prepared by the UK Met Office for the sixth coupled model intercomparison project CMIP6 (Eyring et al., 2016). This PI control experiment is used to study the (natural) unforced variability of the climate system and it is one of the reference simulations against which many of the other CMIP6 experiments will be analysed.

The PI simulation considered in this paper uses the N96 resolution version of the HadGEM3-GC3.1 climate model (N96ORCA1). The model set-up, initialization, performance and physical basis are documented in Menary et al. (2018) and Williams et al. (2018), to which publications the reader is referred for a detailed description. In summary, HadGEM3-GC3.1 is a global coupled atmosphere-land-ocean-ice model that comprises the Unified Model (UM) atmosphere model (Walters et al., 2017), the JULES land surface model (Walters et al., 2017), the NEMO ocean model (Madec et al., 2015) and the CICE sea ice model (Ridley et al., 2018). The UM vertical grid contains 85 pressure levels (terrain-following hybrid height coordinates) while the NEMO vertical gird contains 75 depth levels (rescaled-height coordinates). In the N96 resolution version, the atmospheric model utilizes a horizontal grid-spacing of approximately 135 km on a regular latitude-longitude grid. The grid spacing of the ocean model, which employs an ortoghonal curvilinear grid, is 1° everywhere but decreases down to 0.33° between 15° N and 15° S of the equator, as described by Kuhlbrodt et al. (2018).

Following the CMIP6 guidelines, the model was initialized using constant 1850 GHGs, ozone, solar, tropospheric aerosol, stratospheric volcanic aerosol and land use forcings. The UK CMIP6 PI control simulation (hereinafter referred to as PI_{MO}) was originally run on the MO HPC platform on 2500 cores. The model was at first run for 700 model-years to allow the atmospheric and oceanic masses to attain a steady state (model spin-up), and then run for further 500 model-years (actual run length) (see Menary et al. (2018) for details). A copy of the PI control simulation was ported to the ARCHER HPC platform (hereinafter referred to as PI_{AR}), initialized using the atmospheric and oceanic fields from the end of the spin-up and run for 200 model-years using 1500 cores. The source codes of the atmosphere and ocean models were compiled on the two platforms using the same levels of code optimization (`-O` option), vectorization (`-Ovector` option), floating-point precision (`-hfp` option) and, for numerical reproducibility purposes, selecting the least tolerant behaviour in terms of code optimization when the number of ranks or threads varies (`-hflex_mp` option). For the atmosphere component the following options were used: `-O2 -Ovector1 -hfp0 -hflex_mp=strict`. For the ocean component the following options were used: `-O3 -Ovector1 -hfp0 -hflex_mp=strict` .

Table 1. Hardware and software specifications of the ARCHER and MO HPC platforms as used to run the HadGEM3-GC3.1 model.

HPC Platform	Machine	Compiler	Processor
MO	Cray XC40	cce 8.3.4	Broadwell
ARCHER	Cray XC30	cce 8.5.5	Ivy Bridge

Table 1 provides an overview of the hardware and software specifications of the two HPC platforms where the model was run.

Of the possible mechanisms discussed in section 2, the ARCHER and MO simulations were likely affected by differences in compiler, processor type, number of processors and processor decomposition (alongside the different machine).

Note that the porting of the HadGEM3-GC3.1 model from the Met Office computing platform to the ARCHER platform was tested by running 50 ensemble members (each 24 hours long) on both platforms (this was done by the UK Met Office and NCAS-CMS teams). Each ensemble member was created by adding a random bit-level perturbation to a set of selected variables (x- and y- components of the wind, air potential temperature, specific humidity, long-wave radiation and etc.). Variables from each set of ensembles were then tested for significance using a Kolmogorov-Smirnov test to determine whether they can be assumed to be drawn from the same distribution. These tests did not reveal any significant problem with the porting of the HadGEM3-GC3.1 model (Personal Communications). However this method is restricted to time scales shorter than one day. The centennial simulations presented in this paper will help understanding whether or not differences can arise on longer time scales in the HadGEM3-GC3.1 model.

3.2 Data post-processing and analysis

During the analysis of the results, the following climate variables were considered: sea surface temperature (SST), sea ice area/concentration (SIA/SIC), 1.5m air temperature (SAT), the outgoing long-wave and short-wave radiation fluxes at top of the atmosphere (LW TOA and SW TOA), and the precipitation flux (P). These variables were selected as representative of the ocean and atmosphere domains and because they are commonly used to evaluate the status of the climate system.

Discrepancies between the means of the selected variables were analysed at different timescales, from decadal to centennial. To compute 10-, 30-, 50- and 100-year means, (PI$_{MO}$ - PI$_{AR}$) 200-year time-series were divided into 20, 6, 4 and 2 segments respectively. Spatial maps were simply created by averaging each segment over time. Additionally, to create the scatter plots presented in section 4.1, the time average was combined with an area-weighted spatial average. Except for SIC, all the variables were averaged globally. Additionally, SIC, SST and SAT were regionally-averaged over the Northern and Southern Hemisphere, while SW TOA, LW TOA and P were regionally-averaged over the tropics, Northern extra-tropics and Southern extra-tropics according to the underlying physical processes.

Note that, when calculating (PI$_{MO}$ - PI$_{AR}$) differences, PI$_{MO}$ and PI$_{AR}$ segments are subtracted in chronological order. Thus, for example, the first 10 years of PI$_{AR}$ are subtracted from the first 10 years of PI$_{MO}$ and so on. In fact, because the PI control simulation is run with a constant climate forcing, using a 'chronological order' in the strictest sense is meaningless, as every 10 years segment is equally representative of the pre-industrial decadal variability. We acknowledge that an alternative approach, equally valid, would be to subtract PI$_{AR}$ and PI$_{MO}$ segments without a prescribed order.

Discrepancies in the results between the two runs was quantified by computing the Signal-to-Noise Ratio (SNR) for each considered variable at each timescale. The signal is represented by the mean of the differences between PI$_{MO}$ and PI$_{AR}$ (μ_{MO-AR}) and the noise is represented by the standard deviation of PI$_{MO}$ (σ_{MO}), our 'reference' simulation. Because of the basic properties of variance, for which $Var_{X-Y} = Var_X + Var_Y - 2Cov(X,Y)$ (Loeve, 1977), we can more conveniently express the noise as $\sigma_{MO} = \frac{\sigma_{MO-AR}}{\sqrt{2}}$, under the assumptions that PI$_{MO}$ and PI$_{AR}$ are uncorrelated ($Cov(MO,AR) = 0$), and have same variance ($Var_{MO} = Var_{AR}$). This allowed us to compute SNR on one same grid, and avoid divisions by (nearly) zero when the sea ice field between PI$_{MO}$ and PI$_{AR}$ evolved differently, resulting in unrealistically high SNR values along the sea ice edges. Finally, SNR is defined as:

$$SNR = \frac{|\mu_{MO-AR}|}{\sigma_{MO}} = \frac{|\mu_{MO-AR}|}{\frac{\sigma_{MO-AR}}{\sqrt{2}}} \tag{2}$$

When SNR < 1, (PI$_{MO}$ - PI$_{AR}$) differences can be interpreted as fluctuations within the estimated range of internal variability. When SNR > 1, (PI$_{MO}$ - PI$_{AR}$) differences in the mean are outside the expected range of internal variability. This eventuality indicates either a true difference in the mean, or that the expected range of variability is underestimated.

For the final step of the analysis, the El Niño Southern Hemisphere Oscillation (ENSO) signal was computed for the ARCHER and MO simulations. We used the NINO3.4 index, with a 3-month running mean, defined as follows:

$$NINO3.4 = SST_{mnth} - \overline{SST_{30yr}} \quad \text{if} \quad 5^\circ\ N \leq \text{latitude} \leq 5^\circ\ S \quad \text{and} \quad 120^\circ\ W \leq \text{longitude} \leq 170^\circ\ W \tag{3}$$

where SST_{mnth} is the monthly sea surface temperature and $\overline{SST_{30yr}}$ is the climatological mean of the first 30 years of simulation used to compute the anomalies.

4 Results and discussion

4.1 Multiple Timescales

The long-term means of the selected variables, and the associated SNR, are shown in Figures 2 and 3. All the variables exhibit a SNR < 1, indicating that on multi-centennial timescales the differences observed between the two simulations fall into the expected range of variability of the PI control run.

When maps like the ones in Figure 2 and 3 are computed using 10-, 30-, 50- and 100-year averaging periods (not shown), the magnitude of the anomalies increase and (PI$_{MO}$ - PI$_{AR}$) differences become significant (SNR » 1). This behaviour is discussed below.

Figures 4 to 9 show annual-mean time-series of spatially averaged SST, SIA, SAT, SW TOA, LW TOA and P, respectively. Figures 4d to 9d show (PI$_{MO}$ - PI$_{AR}$) differences as a function of the averaging timescale for each variable (see section 3.2

Table 2. 200-year global mean and standard deviation for SST, SIA, SAT, SW TOA, LW TOA and P.

	MO	ARCHER
	Mean , StDev	Mean , StDev
SST (°C)	17.93 , 0.07	17.95 , 0.08
SIA (10^6 km^2)	21.44 , 0.65	21.30 , 0.68
SAT (°C)	13.71 , 0.10	13.75 , 0.12
SW TOA (W /m^2)	98.83 , 0.24	98.76 , 0.27
LW TOA (W /m^2)	241.29 , 0.27	241.36 , 0.33
P (10^{-6} kg /m^2 /s)	36.22 , 0.12	36.25 , 0.14

for details on the computation of the means). The 200-year global-mean and standard deviation of each variable are shown in Table 2.

For all the considered variables, PI_{MO} and PI_{AR} start diverging quickly after the first few time-steps, once the system has lost memory of the initial conditions. See section 2 (Figure 1) for further discussion on how machine-dependent processes can influence the temporal evolution of the system.

SST, SAT, SW TOA and LW TOA differ the most in the Northern Hemisphere (and particularly on decadal timescales) (yellow diamonds in Figures 4d,6d,7d,8d), while SIA anomalies are particularly high in the Southern Hemisphere (red crosses in Figure 5d) and P anomalies in the tropics (green circles in Figure 9d). Overall, discrepancies are the largest at decadal timescales where the spread between the two simulations can reach $|0.2|$ °C in global mean air temperature (Figure 6d), $|1.2|$ million km^2 in Southern Hemisphere sea ice area (Figure 5d), or $|1|$ W /m^2 in global TOA outgoing LW flux (Figure 8d).

On decadal timescales, the averaging period is too short to adequately sample the model interannual variability; therefore the estimated mean is not stable, and the estimated standard deviation is likely to be underestimated compared with the true standard deviation of the model internal variability. Large differences in the mean and a SNR »1 are, thus, not surprising when analysing decadal periods. On longer timescales, the estimate of the mean and standard deviation converge toward their 'true' values. Accordingly, we see that the differences in the mean between PI_{MO} and PI_{AR} become smaller and approach zero (Figure 4d to 9d). When considering the 200-year long-term mean, we find no SNR value greater than one (Figure 2 and 3). Following this diagnostic, and for the variables we assessed, our results show that there is no significant difference in the simulated mean between the two PI_{MO} and PI_{AR} HadGEM3-GC3.1 simulations when considering a 200-year long period.

In Figures 4d to 9d, the variation of (PI_{MO} - PI_{AR}) with the timescale suggests the existence of power law relationship[1]. To investigate this behaviour, a base-10 logarithmic transformation was applied to the x- and y-axes of Figure 4d to 9d and linear regression was used to find the straight-lines that best fit the data.

[1] Note that, for readability, the ticks of the x-axes of Figures 4d to 9d were equally spaced. This partially masks the power law behaviour discussed in the paper, which can be better detected when the natural x-axes are used.

Figure 10 shows log-log plots for SST, SAT, SW TOA, LW TOA and P for the maximum (PI$_{MO}$ - PI$_{AR}$) values at each timescale. To ease the comparison, all the variables were averaged globally and over the SH and NH Hemispheres. Global, NH and SH mean data all align along a straight line, supporting the existence of a power law. However, the most interesting result emerges at the global scale where (PI$_{MO}$ - PI$_{AR}$) differences vary following a same power law relationship, regardless the physical quantity considered. More precisely, the actual slope values for SST, SAT, SW TOA, LW TOA and P are: -0.65, -0.65, -0.64, -0.66, -0.67 respectively. Thus, the straight-lines that best fit the global mean data in Figure 10 have a slope of $\approx$ 2/3. The existence of a $\approx$ 2/3 power law, which does not depend on the single quantity, shows a consistent scaling of (PI$_{MO}$ - PI$_{AR}$) differences with the timescale that approaches a plateau near the 200-year timescale (note that an actual plateau can only be reached for longer simulations, as differences computed over all timescales longer than 200 years would be $\approx$ 0).

SIA (not shown) was the only variable that did not show a $\approx$ 2/3 power law relationship. This however should not invalidate the analysis presented above. The sea ice area is an integral computed on a limited area, and not a mean computed on a globally uniform surface (like all the other variables considered here), and thus represents a signal of a different nature.

In summary, although large differences can be observed at smaller time-scales (see next section for further discussion), the climate of PI$_{MO}$ and PI$_{AR}$ is indistinguishable on the 200-year time-scale. We thus conclude that the mean climate properties simulated by the HadGEM3-GC3.1 model are reproducible on different HPC platforms, provided that a long-enough simulation length is used.

Our results also show that HadGEM3-GC3.1 does not suffer from compiler bugs that would make the model behave differently on different machines for integration times longer than 24 hours (for which the model was previously tested, see section 3.1).

4.2 The 100-year timescale

The large differences observed on time-scales shorter than 200-years are a direct consequence of the (potentially underestimated) internal variability of the model, and triggered (at least initially) by machine-dependent processes (compiler, machine architecture etc., see section 2 and 3.1 for details). The two simulations behave similarly to ensemble members initiated from different initial conditions. Therefore, they exhibit different phases of the same internal variability but over longer time-scales differences converge to zero (Figure 4 - 9).

While in section 4.1 we showed that PI$_{MO}$ and PI$_{AR}$ necessitate 200 years to become statistically indistinguishable, an interesting case to look at is the 100-year time-scale.

For instance, the CMIP6 minimum run-length requirement for a few of the Model Intercomparison Projects (MIPs), excluding DECK and Historical simulations, is 100 years or less, and not always ensembles are requested (e.g., some of the Tier 1/2/3 experiments in PMIP (Otto-Bleisner et al., 2017), nonlinMIP (Good et al., 2016), GeoMIP (Kravitz et al., 2015), HighResMIP (Haarsma et al., 2016), FAFMIP (Gregory et al., 2016)). This is likely because longer fully-coupled climate simulations are not always possible. They demand significant computational resources or impractically long running-times (for instance, simulating 200 years with the HadGEM3-GC3.1 model on ARCHER in its CMIP6 configuration takes about 4 months).

Our results suggest that 100 years may not be enough to allow HadGEM3-GC3.1 to sample the same climate variability on different HPC platforms. This is particularity evident when we look at the spatial patterns of (PI$_{MO}$ - PI$_{AR}$) differences and the associated SNR.

In Figure 11, (PI$_{MO}$ - PI$_{AR}$) differences materialize into spatial patterns that are signatures of physical processes. SST (Figure 11a,b) and SIC (Figure 11c,d) anomalies are the largest in West Antarctica where ENSO teleconnection patterns are expected, they correspond to regions where SNR becomes equal to/larger than one. This suggests that (PI$_{MO}$ - PI$_{AR}$) differences are driven by two different ENSO regimes (the connection between SIC (and SST) anomalies in the Southern Hemisphere and ENSO has been widely documented in literature, e.g. Kwok and Comiso (2002), Liu et al. (2002), Turner (2004), Welhouse et al. (2016), Pope et al. (2017)).

This hypothesis is confirmed by the ENSO signal in Figure 12. A few times, to a strong El Niño (/La Niña) event in PI$_{MO}$ corresponds a strong La Niña (/El Niño) event in PI$_{AR}$. This opposite behaviour enlarges SIC (and SST) differences between the two runs and strengthens the μ_{MO-AR} signal, resulting in a strong SNR.

As ENSO provides a medium-frequency modulation of the climate system, it is not surprising that it takes longer than 100 years for its variability to be fully represented (see e.g., Wittenberg (2009)).

Finally, we want to know whether the two ENSO regimes in PI$_{MO}$ and PI$_{AR}$ are a reflection of the different computing environment or solely the result of natural variability (i.e. if a similar behaviour can be detected for simulations run on a same machine). This can be done by splitting the 200-year simulations in two segments and assuming that each 100-year period of PI$_{MO}$ and PI$_{AR}$ is a member of an ensemble of size two. Therefore, the ARCHER ensemble is made of PI$_{AR}$1st and PI$_{AR}$2nd, and the MO ensemble comprises PI$_{MO}$1st and PI$_{MO}$2nd.

Figure 11e and 11f show the signal-to-noise ratio corresponding to SST differences between PI$_{AR}$1st and PI$_{AR}$2nd and PI$_{MO}$1st and PI$_{MO}$2nd. In Figure 11e, the SNR pattern exhibited by the ARCHER ensemble members resemble the one shown by (PI$_{MO}$ - PI$_{AR}$) differences in Figure 11b. Thus, we conclude that differences between ARCHER and MO are comparable to differences between ensemble members run on a single machine.

As for PI$_{MO}$, in Figure 11f, large differences (and SNR > 1) between the two ensemble members are found in East Antarctica. While this suggests that in this case a climate process other than ENSO is in action, the large SNR confirms that 100 years is a too short length for constant-forcing HadGEM3-GC3.1 simulations even on the same machine.

In summary, the analysis above confirms that (PI$_{MO}$ - PI$_{AR}$) differences, while triggered by the computing environment, are largely dominated by the internal variability as they persist among ensemble members on the same machine (in Figure 11 SNR > 1 always).

5 Discussion and Conclusions

In this paper, the effects of different computing environments on the reproducibility of coupled climate model simulations are discussed. Two versions of the UK CMIP6 PI control simulation, one run on the UK Met Office supercomputer (MO) (PI$_{MO}$)

and the other run on the ARCHER (PI_{AR}) HPC platform, were used to investigate the impact of machine-dependent processes of the N96ORCA1 HadGEM3-GC3.1 model.

Discrepancies between the means of key climate variables (SST, SIA/SIC, SAT, SW TOA, LW TOA and P) were analysed at different timescales, from decadal to centennial (see section 3.2 for details on methodology).

Although the two versions of the same PI control simulation do not bit-compare, we found that the long-term statistics of the two runs are similar and that, on multi-centennial timescales, the considered variables show a signal-to-noise ratio (SNR) less than one. We conclude that in order for PI_{MO} and PI_{AR} to be statistically indistinguishable a 200-year averaging period must be used for the analysis of the results. This indicates that simulations using the HadGEM3-GC3.1 model are reproducible on different HPC platforms (in their mean climate properties), provided that a long-enough simulation length is used.

Additionally, the relationship between global mean differences and timescale exhibits a 2/3 power law behaviour, regardless the physical quantity considered, that approaches a plateau near the 200-year time-scale. Thus, there exist a consistent time-dependent scaling of (PI_{MO} - PI_{AR}) differences across the whole climate simulation, so that variables converge toward their true values at the same rate, independently on the physical processes that they represent .

Larger inconsistencies between the two runs were found for shorter timescales (where SNR ≥ 1), being the largest at decadal timescales. For example, when a 10-year averaging period is used, discrepancies between the runs can be equal to up to $|0.2|$ °C global mean air temperature anomalies, or $|1.2|$ million km^2 Southern Hemisphere sea ice area anomalies. The observed differences are a direct consequence of the different sampling of the internal variability when the same climate simulation is run on different machines. They become approximately zero when a 200-year averaging period is used, confirming that the overall physical behaviour of the model was not affected by the different computing environment.

On a 100-year timescale, large SST and SIC differences (with SNR ≥ 1) where found where ENSO teleconnection patterns are expected. Medium-frequency climate processes like ENSO need longer than 100 years to be fully represented. Thus, a 100-year constant-forcing simulation may not be long enough to correctly capture the internal variability of the HadGEM3-GC3.1 model (on the same, or on a different, machine). While this result is not per se unexpected, it is relevant to CMIP6 experiments as CMIP6 protocols recommend a minimum simulation length of 100 years (or less) for many of the MIP experiments.

This result has immediate implications for those members of the UK CMIP6 community who will run individual MIP experiments on the ARCHER HPC platform, and will compare results against the reference PI simulation run on the MO platform by the UK Met Office. The magnitude of (PI_{MO} - PI_{AR}) differences presented in this paper should be regarded as threshold values below which differences between ARCHER and MO simulations must be interpreted with caution (as they might be the consequence of a wrong sampling of the model internal variability rather than the climate response to a different forcing).

In the light of our results, our recommendation to the UK MIPs studying the climate response to different forcings is to run HadGEM3-GC3.1 for at least 200 years, even when CMIP6 minimum requirements are of 100 years (see for example PMIP protocols (Otto-Bleisner et al., 2017)).

Finally, although the quantitative analysis presented in this paper applies strictly to HadGEM3-GC3.1 constant-forcing climate simulations only, this study has the broader purpose of increasing the awareness of the climate modelling community on the subject of machine dependence of climate simulations.

Code availability. Access to the model code used in the manuscript has been granted to the editor. The source code of the UM model is available under licence. To apply for a licence go to http://www.metoffice.gov.uk/research/modelling-systems/unified-model. JULES is available under licence free of charge, see https://jules-lsm.github.io/. The NEMO model code is available from http://www.nemo-ocean.eu. The model code for CICE can be downloaded from https://code.metoffice.gov.uk/trac/cice/browser.

Data availability. Access to the data used in the manuscript has been granted to the editor. The CMIP6 PI simulation run by the UK Met Office will be made available on the Earth System Grid Federation (ESGF)

(https://cera-www.dkrz.de/WDCC/ui/cerasearch/cmip6?input=CMIP6.CMIP.MOHC.UKESM1-0-LL), the data repository for all CMIP6 output. CMIP6 outputs are expected to be public by 2020. Dataset used for the analysis of the PI simulation ported to ARCHER can be shared, under request, via the CEDA platform (https://help.ceda.ac.uk). Please contact the authors.

Author contributions. M.V.G ran the ARCHER simulation, processed the data and carried out the scientific analysis with the contribution of L.C.S and D.S. M.V.G carried out tests with simple model described in section 2. G.L and R.H ported the PI simulation to the ARCHER supercomputer, provided technical support and advised on the nature of machine-dependent processes. All authors revised the manuscript.

Acknowledgements. M.V.G. and L.S. acknowledge the financial support of the NERC research grants NE/P013279/1 and NE/P009271/1. This work used the ARCHER UK National Supercomputing Service (http://www.archer.ac.uk). Authors acknowledge use of the UK Met Office supercomputing facility in providing data for model comparisons.

References

Eyring, V., Bony, S., Meehl, G. A., Senior, C. A., Stevens, B., Stouffer, R. J., and Taylor, K. E.: Overview of the Coupled Model Intercomparison Project Phase 6 (CMIP6) experimental design and organization, Geoscientific Model Development (Online), 9, 2016.

Good, P., Andrews, T., Chadwick, R., Dufresne, J.-L., Gregory, J. M., Lowe, J. A., Schaller, N., and Shiogama, H.: nonlinMIP contribution to CMIP6: model intercomparison project for non-linear mechanisms: physical basis, experimental design and analysis principles (v1. 0), Geoscientific Model Development, 9, 4019–4028, 2016.

Gregory, J. M., Bouttes, N., Griffies, S. M., Haak, H., Hurlin, W. J., Jungclaus, J., Kelley, M., Lee, W. G., Marshall, J., Romanou, A., et al.: The Flux-Anomaly-Forced Model Intercomparison Project (FAFMIP) contribution to CMIP6: investigation of sea-level and ocean climate change in response to CO forcing, Geoscientific Model Development, 9, 3993–4017, 2016.

Haarsma, R. J., Roberts, M. J., Vidale, P. L., Senior, C. A., Bellucci, A., Bao, Q., Chang, P., Corti, S., Fučkar, N. S., Guemas, V., et al.: High resolution model intercomparison project (HighResMIP v1. 0) for CMIP6, Geoscientific Model Development, 9, 4185–4208, 2016.

Hong, S.-Y., Koo, M.-S., Jang, J., Esther Kim, J.-E., Park, H., Joh, M.-S., Kang, J.-H., and Oh, T.-J.: An evaluation of the software system dependency of a global atmospheric model, Monthly Weather Review, 141, 4165–4172, 2013.

Kravitz, B., Robock, A., Tilmes, S., Boucher, O., English, J. M., Irvine, P. J., Jones, A., Lawrence, M. G., MacCracken, M., Muri, H., et al.: The geoengineering model intercomparison project phase 6 (GeoMIP6): Simulation design and preliminary results, Geoscientific Model Development, 8, 3379–3392, 2015.

Kuhlbrodt, T., Jones, C. G., Sellar, A., Storkey, D., Blockley, E., Stringer, M., Hill, R., Graham, T., Ridley, J., Blaker, A., et al.: The Low-Resolution Version of HadGEM3 GC3. 1: Development and Evaluation for Global Climate, Journal of Advances in Modeling Earth Systems, 10, 2865–2888, 2018.

Kwok, R. and Comiso, J. C.: Spatial patterns of variability in Antarctic surface temperature: Connections to the Southern Hemisphere Annular Mode and the Southern Oscillation, Geophysical Research Letters, 29, 50–1, 2002.

Liu, J., Yuan, X., Rind, D., and Martinson, D. G.: Mechanism study of the ENSO and southern high latitude climate teleconnections, Geophysical Research Letters, 29, 24–1, 2002.

Liu, L., Li, R., Zhang, C., Yang, G., Wang, B., and Dong, L.: Enhancement for bitwise identical reproducibility of Earth system modeling on the C-Coupler platform, Geoscientific Model Development Discussions, 8, 2403–2435, 2015a.

Liu, L., Peng, S., Zhang, C., Li, R., Wang, B., Sun, C., Liu, Q., Dong, L., Li, L., Shi, Y., et al.: Importance of bitwise identical reproducibility in earth system modeling and status report, Geosci. Model Dev, 8, 4375–4400, 2015b.

Loeve, M.: Elementary probability theory, pp. 1–52, 1977.

Lorenz, E. N.: Deterministic nonperiodic flow, Journal of the atmospheric sciences, 20, 130–141, 1963.

Madec, G. et al.: NEMO ocean engine, 2015.

Menary, M. B., Kuhlbrodt, T., Ridley, J., Andrews, M. B., Dimdore-Miles, O. B., Deshayes, J., Eade, R., Gray, L., Ineson, S., Mignot, J., et al.: Preindustrial Control Simulations With HadGEM3-GC3. 1 for CMIP6, Journal of Advances in Modeling Earth Systems, 2018.

Otto-Bleisner, B. L., Braconnot, P., Harrison, S. P., Lunt, D. J., Abe-Ouchi, A., Albani, S., Bartlein, P. J., Capron, E., Carlson, A. E., Dutton, A., et al.: The PMIP4 contribution to CMIP6–Part 2: Two interglacials, scientific objective and experimental design for Holocene and Last Interglacial simulations, Geoscientific Model Development, 10, 3979–4003, 2017.

Pope, J. O., Holland, P. R., Orr, A., Marshall, G. J., and Phillips, T.: The impacts of El Niño on the observed sea ice budget of West Antarctica, Geophysical Research Letters, 44, 6200–6208, 2017.

Ridley, J. K., Blockley, E. W., Keen, A. B., Rae, J. G., West, A. E., and Schroeder, D.: The sea ice model component of HadGEM3-GC3. 1, Geoscientific Model Development, 11, 713–723, 2018.

Song, Z., Qiao, F., Lei, X., and Wang, C.: Influence of parallel computational uncertainty on simulations of the Coupled General Climate Model, Geoscientific Model Development, 5, 313–319, 2012.

Turner, J.: The El Niño–Southern Oscillation and Antarctica, International Journal of Climatology: A Journal of the Royal Meteorological Society, 24, 1–31, 2004.

Walters, D., Brooks, M., Boutle, I., Melvin, T., Stratton, R., Vosper, S., Wells, H., Williams, K., Wood, N., Allen, T., et al.: The Met Office unified model global atmosphere 6.0/6.1 and JULES global land 6.0/6.1 configurations, Geoscientific Model Development, 10, 1487–1520, 2017.

Welhouse, L. J., Lazzara, M. A., Keller, L. M., Tripoli, G. J., and Hitchman, M. H.: Composite analysis of the effects of ENSO events on Antarctica, Journal of Climate, 29, 1797–1808, 2016.

Williams, K., Copsey, D., Blockley, E., Bodas-Salcedo, A., Calvert, D., Comer, R., Davis, P., Graham, T., Hewitt, H., Hill, R., et al.: The Met Office global coupled model 3.0 and 3.1 (GC3. 0 and GC3. 1) configurations, Journal of Advances in Modeling Earth Systems, 10, 357–380, 2018.

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

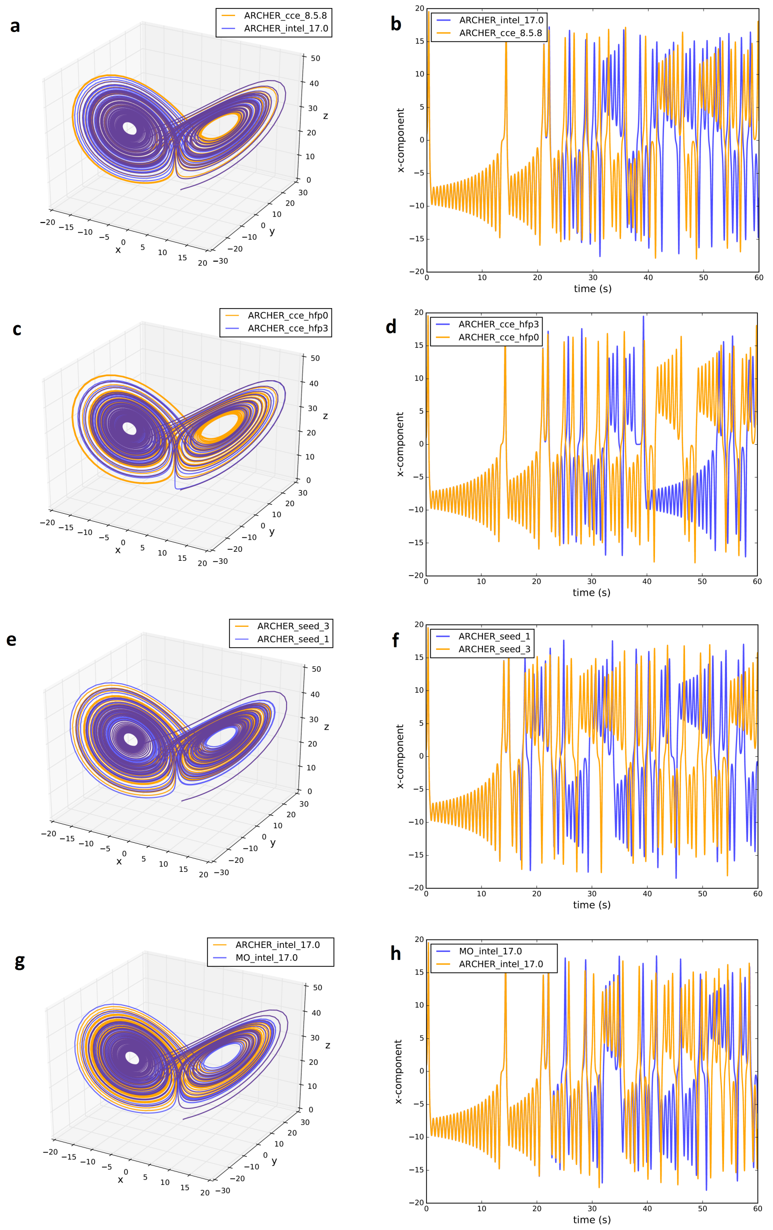

Figure 1. Attractor (left-hand side) and time-series of the x-component (right-hand side) of the 3D Lorenz model for simulations run on ARCHER using: the cce8.3.4 and intel17.0 compilers (a, b), same compiler but different level of floating-point optimization (c, d), same compiler and compiling options but different seed for random number generator (e, f). g and h are the Lorenz attractor and the x-component time-series for the Lorenz model run on MO and ARCHER using same compiler and compiling options.

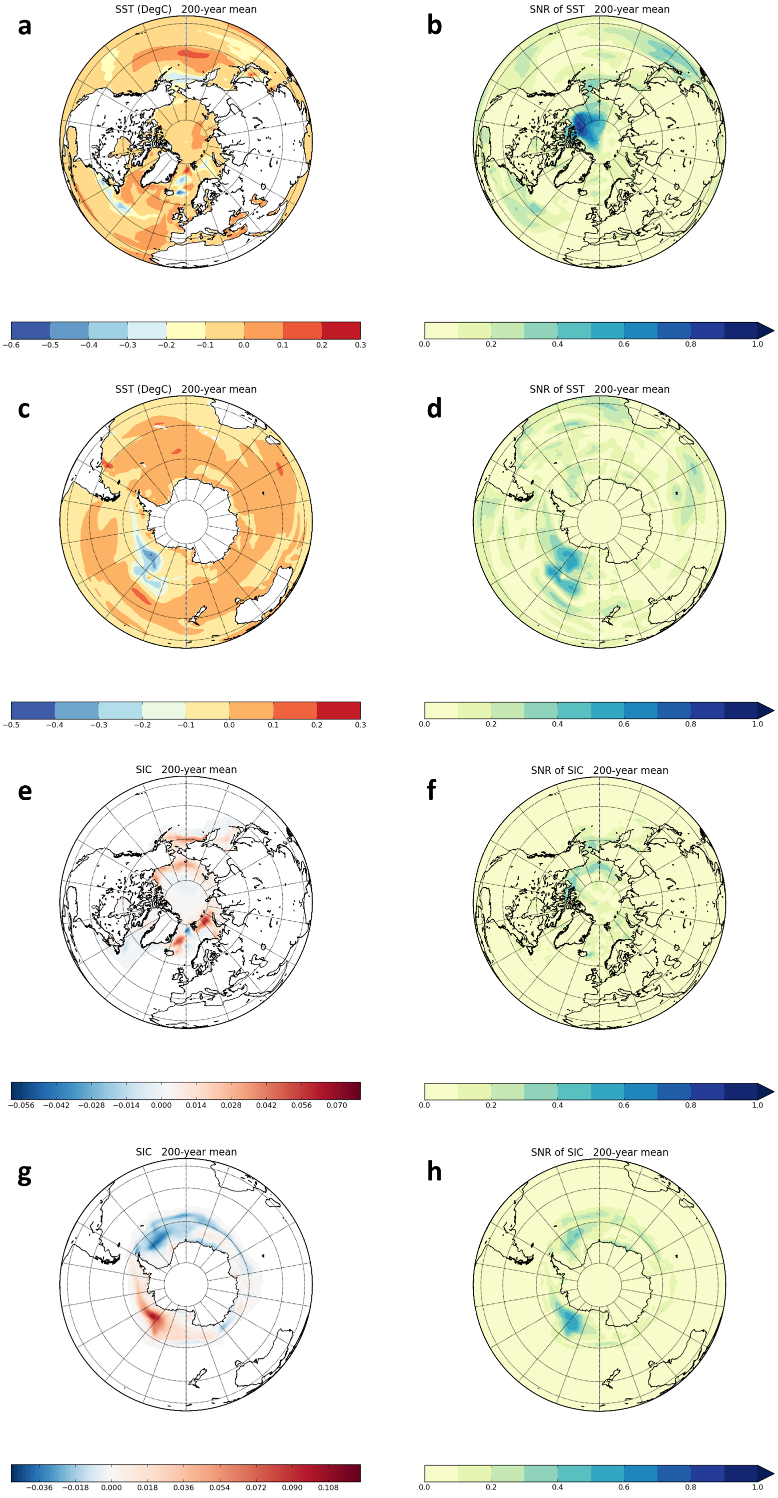

Figure 2. 200-year means and corresponding SNR of (PI$_{MO}$ - PI$_{AR}$) differences for NH SST (a, b), SH SST (c, d), NH SIC (e, f) and SH SIC (g, h).

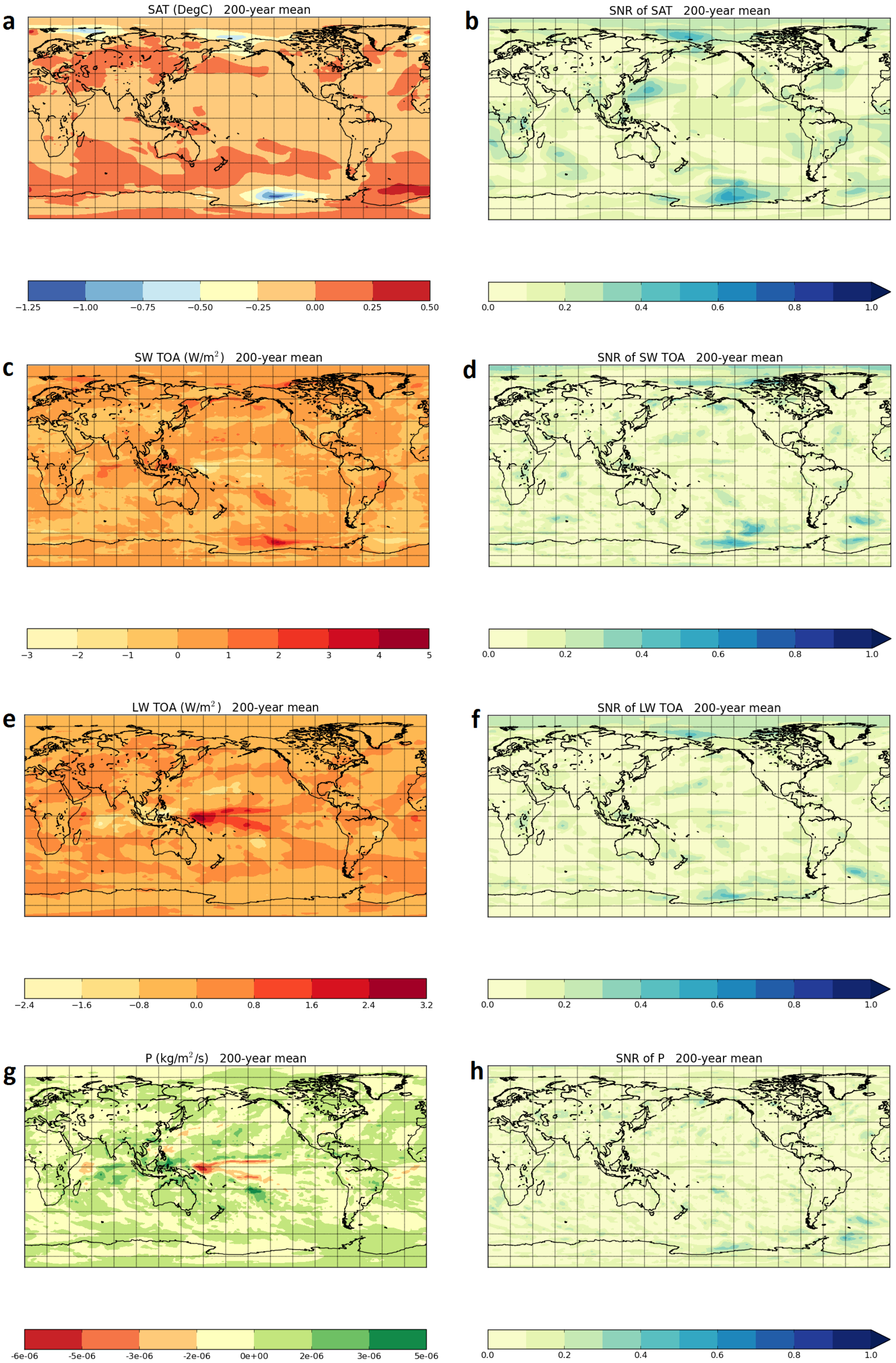

Figure 3. 200-year means and corresponding SNR of (PI$_{MO}$ - PI$_{AR}$) differences for SAT (a, b), SW TOA (c, d), LW TOA (e, f) and P (g, h).

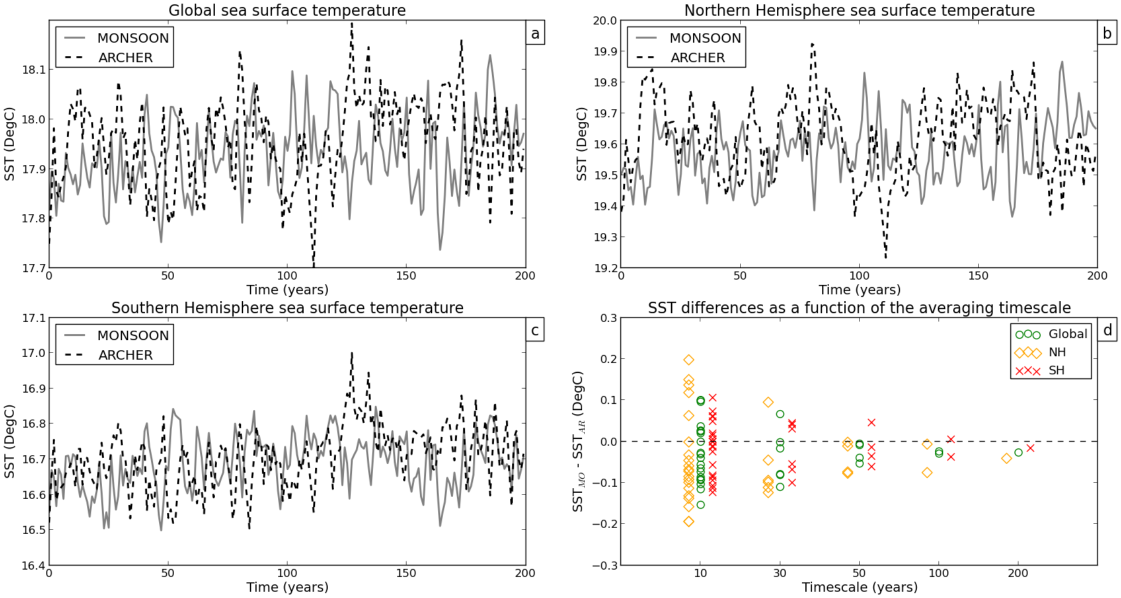

Figure 4. Annual-mean time-series of Global SST (a), Northern Hemisphere SST (b) and Southern Hemisphere SST (c) for PI_{MO} (grey line) and PI_{AR} (dashed line). d shows how SST differences vary as a function of the timescale.

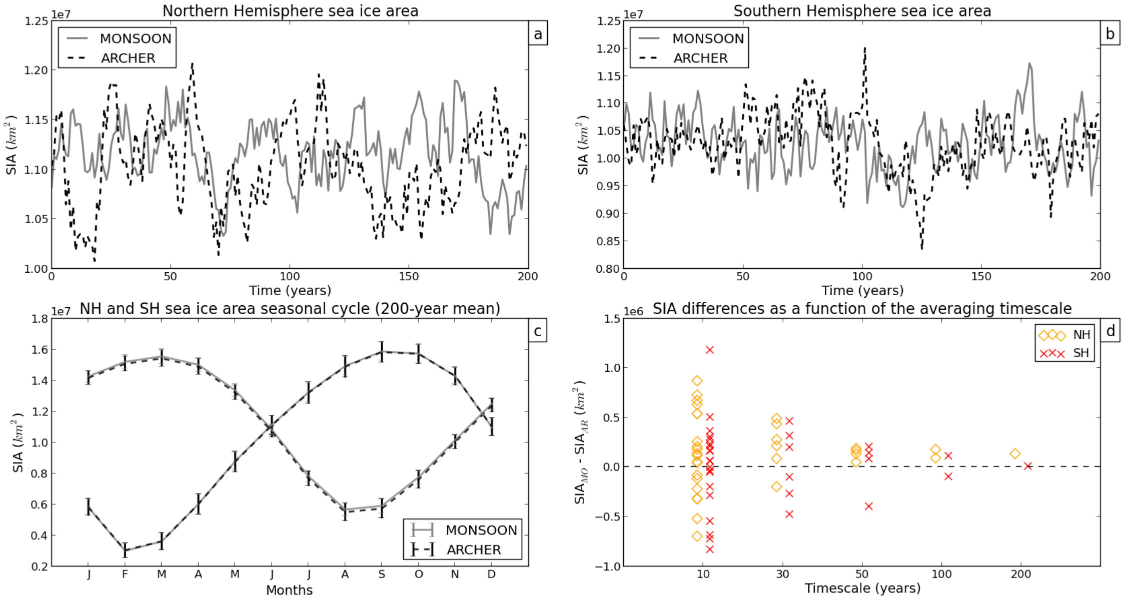

Figure 5. Annual-mean time-series of Northern Hemisphere SIA (a) and Southern Hemisphere SIA (b) for PI$_{MO}$ (grey line) and PI$_{AR}$ (dashed line). The 200-year mean of the NH and SH SIA seasonal cycle is shown in c. d shows how SIA differences vary as a function of the timescale.

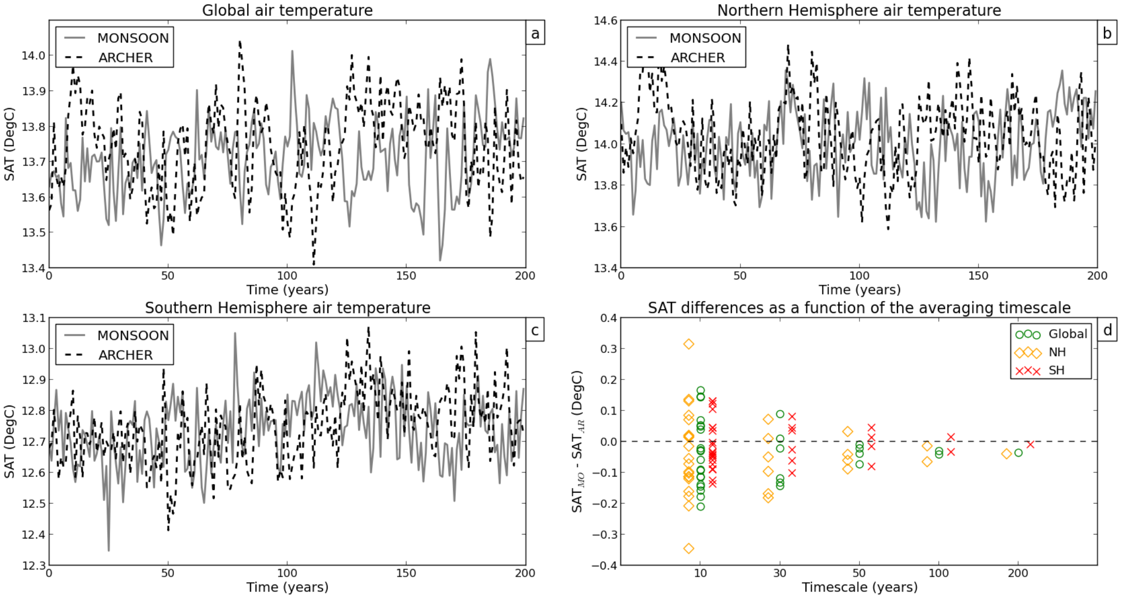

Figure 6. As in 4 but for SAT.

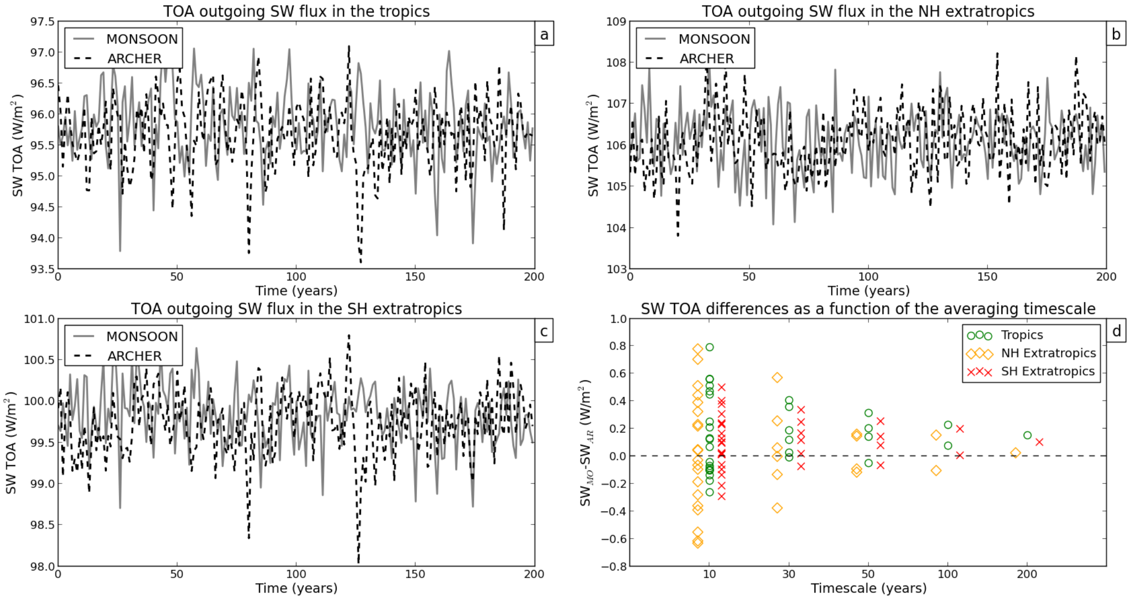

Figure 7. Annual-mean time-series of SW TOA in the tropics (a), SW TOA in the Northern Extratropics (b) and SW TOA in the Southern Extratropics (c) for PI$_{MO}$ (grey line) and PI$_{AR}$ (dashed line). d shows how SW TOA differences vary as a function of the timescale.

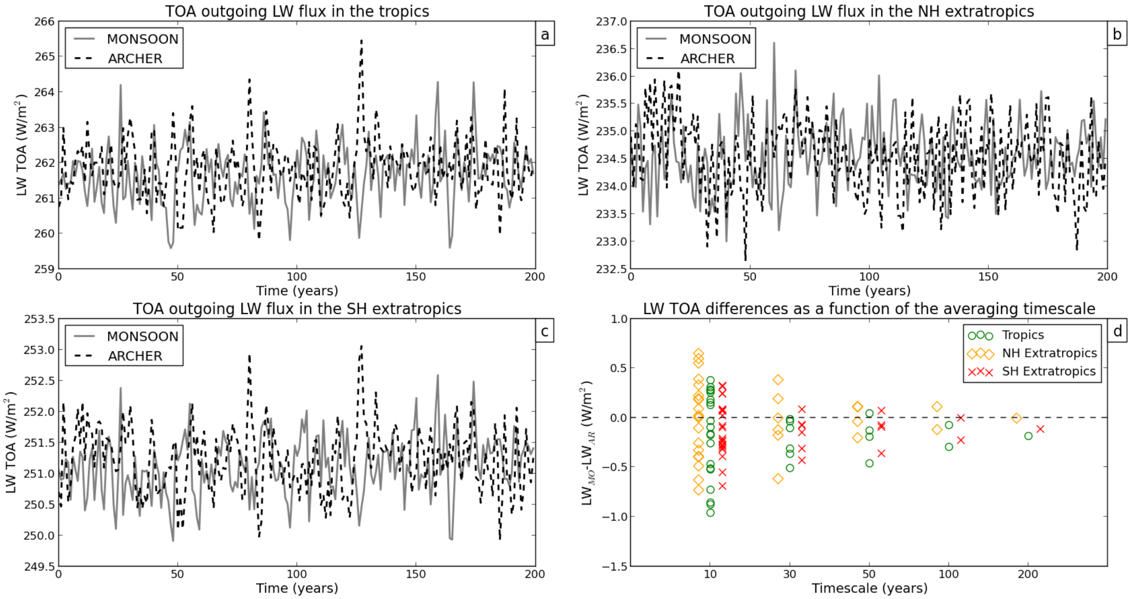

Figure 8. As in 4 but for LW TOA.

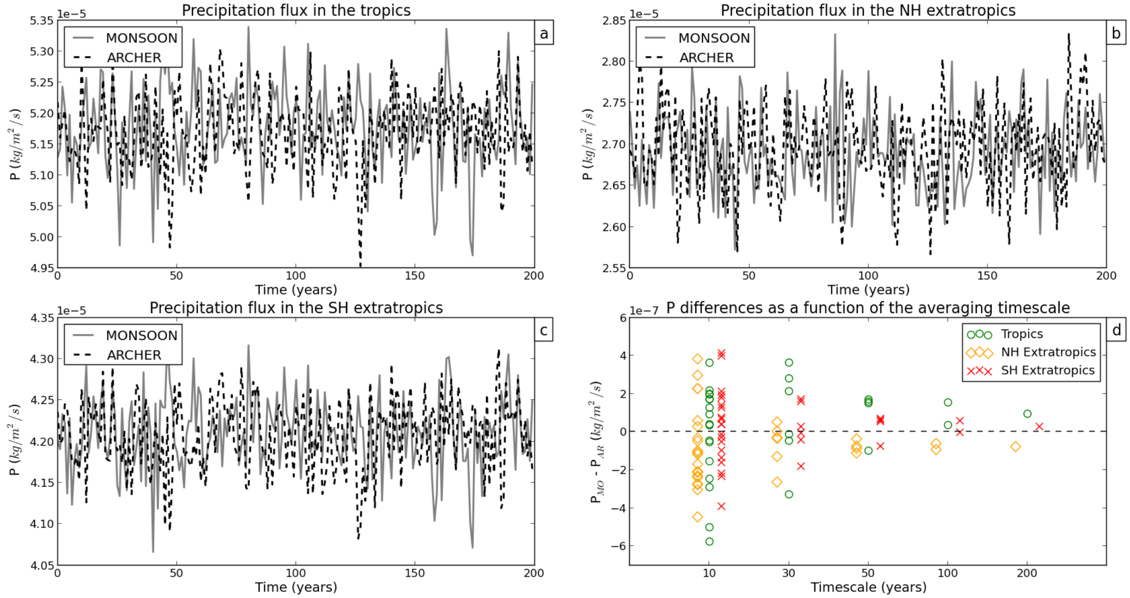

Figure 9. As in 4 but for P.

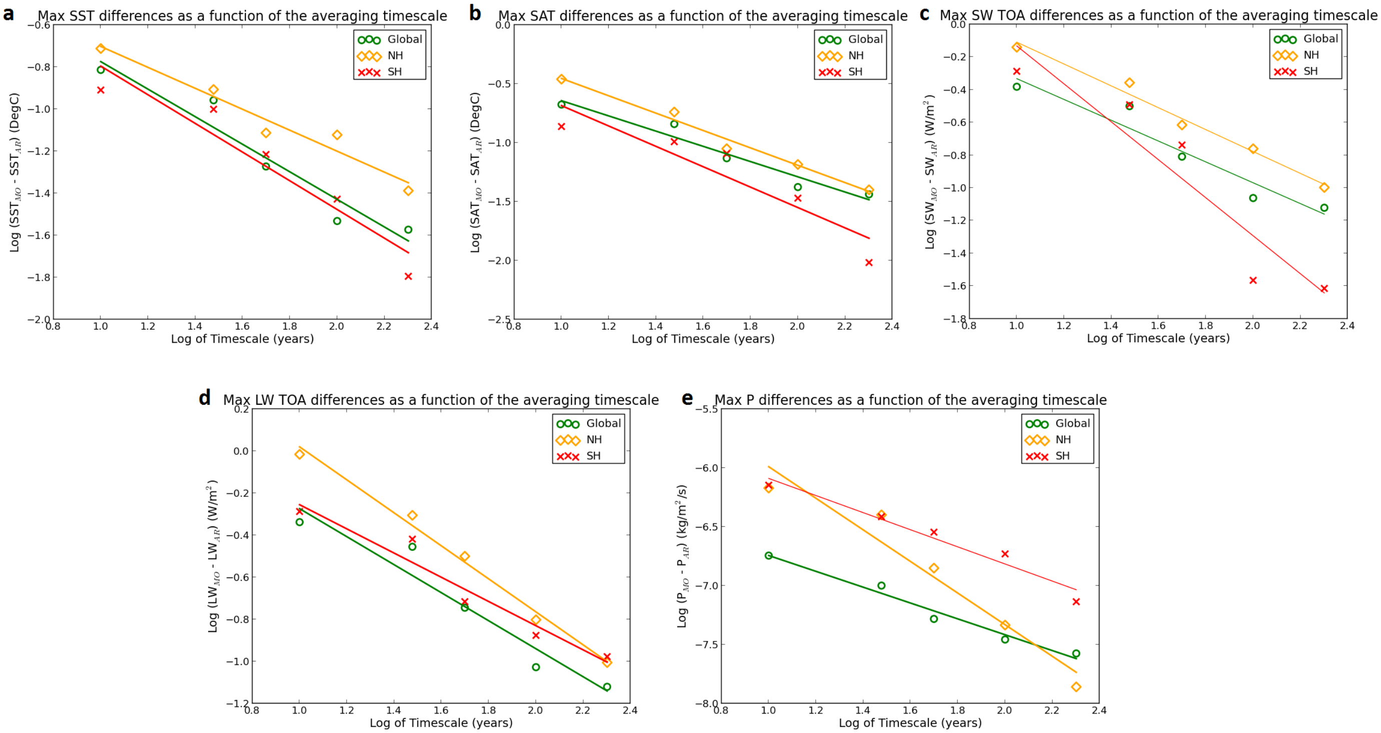

Figure 10. Log-log plots of SST (a), SAT (b), SW TOA (c), LW TOA (d) and P (e) representing maximum (PI$_{MO}$ - PI$_{AR}$) differences as a function of the timescale. All the variables were averaged globally (green circles) and over the SH (red crosses) and NH (yellow diamonds) Hemispheres. The straight-lines represent the best fit lines for the data obtained by linear regression.

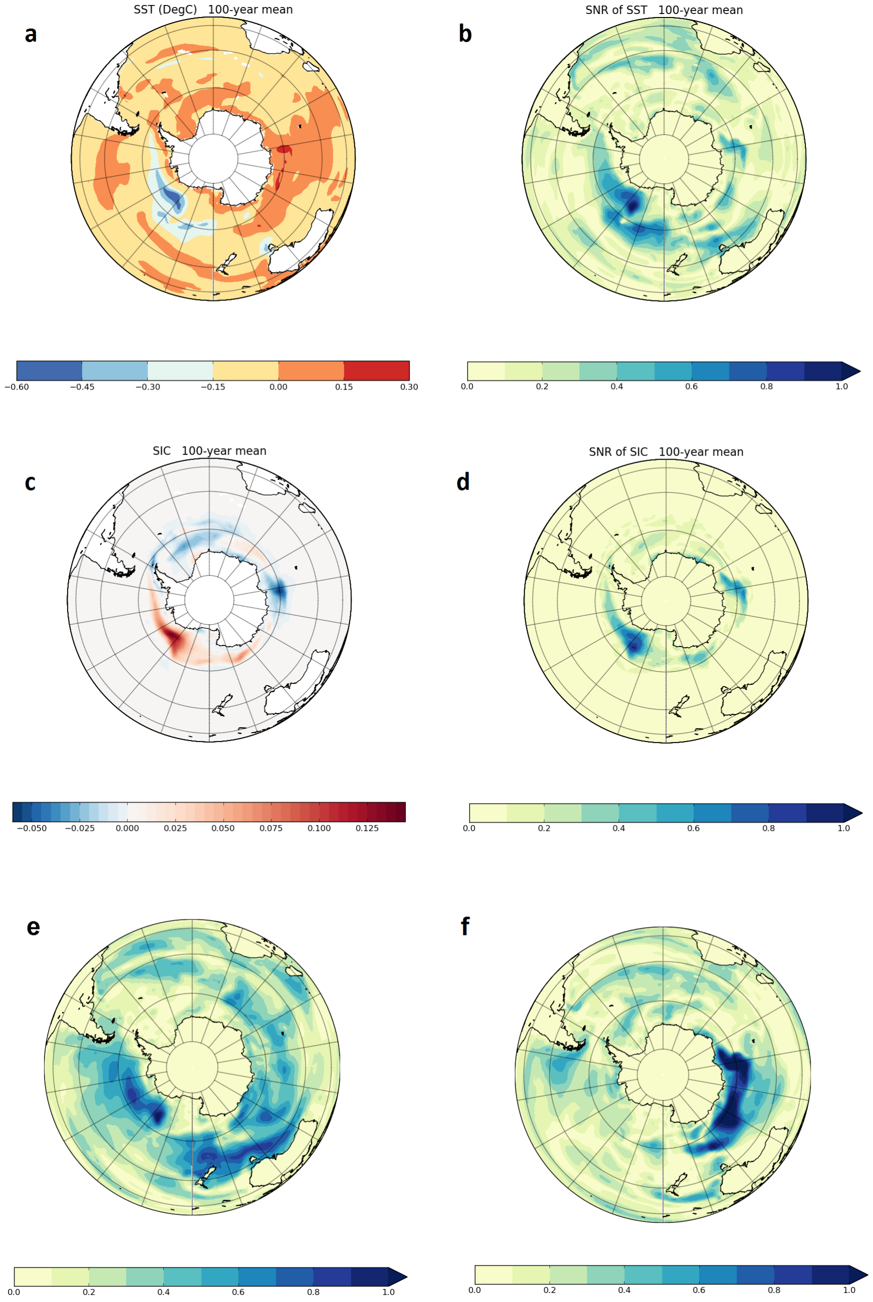

Figure 11. 100-year means and corresponding SNR of (PI_{MO} - PI_{AR}) differences for SH SST (a, b) and SH SIC (c, d). e and f show SNR of (PI_{AR}1st - PI_{AR}2nd) and (PI_{MO}1st - PI_{MO}2nd) differences for SH SST respectively.

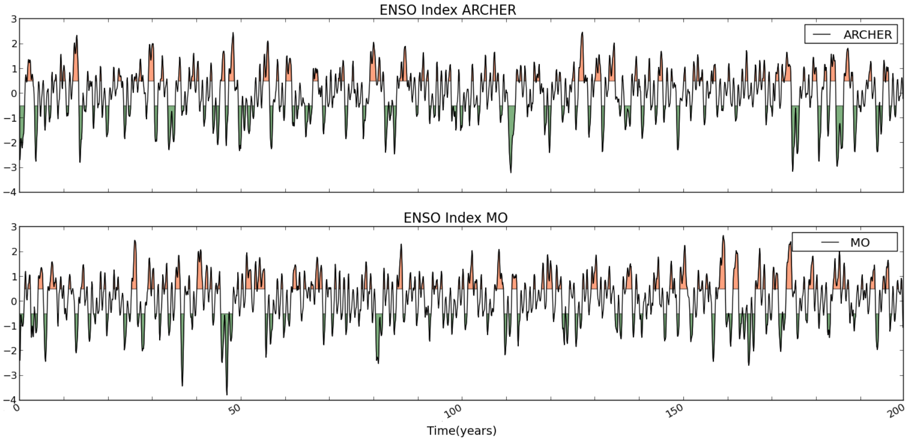

Figure 12. The NINO3.4 index for PI$_{MO}$ and PI$_{AR}$. A 3-month running mean was applied to the ENSO signal and values greater/smaller than or equal to $\pm$ 0.5 are shaded in orange/green.