We thank the editor and referees for the time they have taken to provide valuable comments on our manuscript, the comments have enabled us to make substantial clarifications and other improvements to the manuscript.

We note that the major concern from both referees was on the role of internal variability on our results. In this, the referees raise a valid point. The nature and the purpose of our work was misunderstood by both referees because it was previously insufficiently clear. We have taken the time to clarify the purpose of the paper by rewriting the introduction and have rewritten several other sections within the manuscript, as was required. It is not our aim to fully separate between machine dependence uncertainty and internal variability. Instead we aim to firstly, explore the influence of machine dependence for our UK CMIP6 model HadGEM3-GC3.1 simulations and secondly, give the wider modelling community general practical guidance when simulations have to be run on different machines – as has frequently been the case for the paleo-modelling MIPs – and is the case for our HadGEM3-GC3.1 simulations.

To make our responses complete, the revised manuscript is provided at the end of the document.

Responses to Referee n1

- ### General Comments

1) "The authors assume that computing the difference between two preindustrial simulations run from the same initial conditions and with the same boundary conditions is a measure of machine dependence. This is fundamentally wrong: it is primarily a measure of differences due to internal variability."

Ref.n1 is right in that we call "machine dependence" those differences that we can observe by simply running a same climate simulation (same model, set-up, forcing etc.) on two different machines.

The context in which the term "machine dependence" is used in the paper is explained in section 2, where we explain possible reasons that can cause the solution in the two cases to evolve differently. We conclude section 2, p4 line 1, clearly stating that differences are triggered by machine-dependent processes (compiler, optimization and etc.) *but eventually exist because of the chaotic nature of the system* (internal variability). Thus, the differences among the two runs analysed throughout the paper are not a measure of the machine dependence over the internal variability (the type of machine dependence Ref.1 refers to) but a measure of how the internal variability responds to the different computing environment in the two cases (see also response to Ref.n2 point 1).

The purpose of our work is to find out for how long we should run a simulation (or analyse its results) on the ARCHER HPC platform to capture/sample the same climate variability exhibited by the reference simulation on the Met Office supercomputer (as we know the two runs do not bit-compare).

To clarify, we reformulated section 1, and adopted a change of terminology, requiring a change of the title of the manuscript alongside a revision of the results and conclusions sections.

2) "I invite you to read the publication by Milroy et al (2018) (https://www.geosci-model-dev.net/11/697/2018/) to see an interesting approach to compare model results on different computers."

The porting of the HadGEM3-GC3.1 model from the Met Office computing platform to the ARCHER platform was tested as a part of routine tests performed by the UK Met Office and NCAS-CMS teams. As the purpose of our work is not to assess the porting of the specific PI simulation, the results of the porting and of its testing are not included in the manuscript. However, we added a paragraph (revised manuscript p5, line 5), explaining model porting and testing.

For completeness, here we provide additional details on the procedure used by the UK team: consistency across simulations on different computers are routinely tested by running 50 ensemble members (each 24 hours long) on both platforms. Each ensemble member is created by adding a random bit-level perturbation to a set of variables of the model initial conditions. These variables are: x- and y- components of the wind, air potential temperature, specific humidity, mass fraction of cloud, air pressure, long-wave and sort-wave radiation.

Variables from each set of ensembles are 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 (please note that these tests were not performed directly by the authors, thus we cannot share the results). Note also that this method cannot detect code bugs, which may cause a same model to behave differently on different machines. In this respect, when we conclude at page 8 line 1-3 that the long-term statistics of the two runs are similar, we provide a strong indication that the HadGEM3-GC3.1 model does not suffer from code bugs giving different outcomes depending on the computing environment. We have now added a sentence highlighting this at page 8, line 18-19:

"Our results also provide a strong indication that HadGEM3-GC3.1 does not suffer from code/compiler bugs that would make the model behave differently on different machines."

3) The whole section on the physical implications is out of subject. You are speculating on differences due to internal variability.

Please see response to Ref.n2 point 4.

- **Specific Comments**

4) "Therefore, I agree that the machine dependence uncertainty could be estimated someway and taken into account, but I think it is wrong to say that it could be removed by running all the models on the same machine. If we can't select a model today on climate-based criteria and comparisons with observations, I don't see any chance to select one computer. Running all the models on the same computer is not reducing or removing machine dependence: it's ignoring it (what we do today), which is fundamentally different.

The Introduction has been substantially changed and the paragraph the reviewer refers to has been removed. Please refer to revised manuscript.

5) "Page 6, lines 13-20: I got really confused by this paragraph. You say "using a chronological order in the strictest sense is meaningless because every 10 years segment is equally representative of the pre-industrial climate variability". And the last two sentences of the paragraph (lines 19-20) contradict this statement.

We agree with the reviewer that this was written in a confusing way, to address this, Lines 19-20 are removed.

6) "Section 3.2: the SNR measure you use to estimate if the mean of PIMO is different from the mean of PIAR should come with a test to determine more objectively when the difference becomes significant. For instance, if you compute your SNR on n years, you could sample 100 couples of n random years (bootstrap) in the same simulation to check the distribution of the SNR when computed between two (random) periods of the same simulation. This would give you an estimate of the influence of internal variability on your SNR (but would also need a longer simulation for this)… An alternative would be the use of the Student t-test on the difference of mean, and the Fisher F-test on the difference of variance."

Alongside computing SNR, PI_MO – PI_AR differences were tested using a 2-tailed Welch's T-test (where our H0 is mean_MO = mean_AR).

In the paper, only results based on SNR are shown. That is because, while most of the time the T-test and the SNR-based analysis gave us the same answer, in a few instances we had discordant results (see Figure 1). Overall, the behaviour detected in comparing the two methods was: t-test indicating significant differences despite SNR < 1. Bearing in mind that what a t-test really signifies is: IF H0 were true the probability of obtaining, in repeated experiments, the observed mean_MO – mean_AR value is less than 0.05, we made the decision of keeping on using SNR as, in our view, this quantity has a greater physical weight and can more readily be interpreted. Additionally, by doing so we chose the most conservative method of the two (i.e. the one that less frequently points to significant differences).

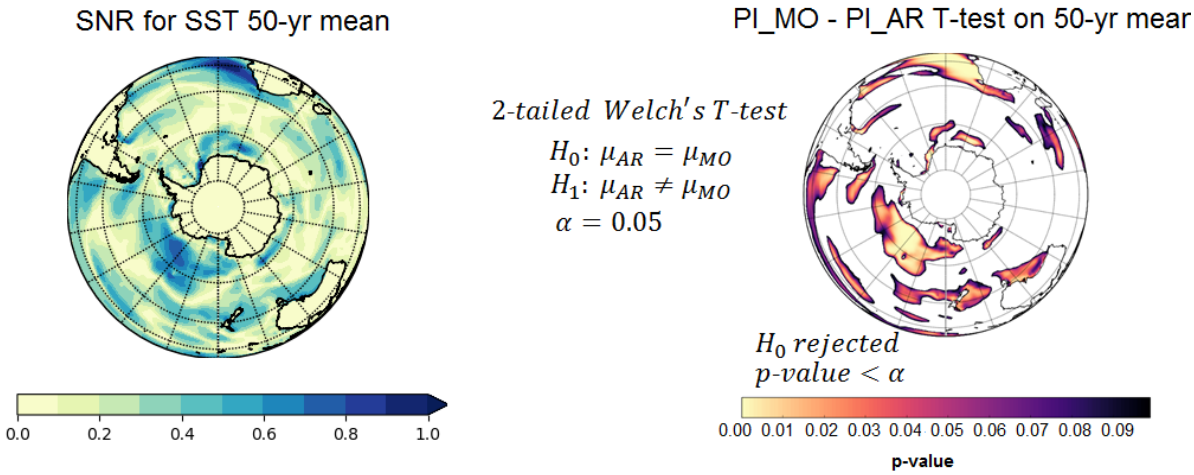

Figure 1 a) SNR computed as in manuscript eq (2) for PI_MO – PI_AR SST differences (on a 50-years period): maximum value for SNR is 0.9, indicating that there are differences but SNR still below the threshold value of 1. b) 2-tailed Welch's T-test on same data: where SNR ~ 0.7 - 0.9, the p-value is less than alpha. According to this, differences should be considered significant.

7) "… as long as those differences are lower than one we should not care about them because they do not show any significant difference (according to your definition of the SNR). And then you talk about SNR values of 0.9, 0.7, that are supposed to be close to 1. How close from one are those results?"

We address this comment by modifying the manuscript so that we analyse results in regions only where SNR ≥ 1. See for example p6 line 24, p9 line6, p9 line 29.

Responses to Referee n2

- ## General Comments

1) "Part of the problem I think is conceptual… Subsequent time-steps will simply repeat the exercise (i.e. perturbing the initial conditions for the next time step by machine precision). I do not see how this will produce anything fundamentally different from a standard initial condition ensemble."

A rather detailed explanation of how the machine may affect the numerical solution is given in section 2. However, it is clear that the previous version did not fully successfully communicate which questions our analysis and manuscript address (see below).

Internal variability is usually assessed in climate simulations via two methods: ensemble members for varying-forcing simulations and long centennial runs for constant-forcing simulations. In the first case, the question is "how many" ensemble members are needed to sample correctly the climate variability, in the second case the question is "how long". While the obvious answers are the more the better and the longest the better, one may ask what is the minimum number of ensembles, or the minimum simulation length required, that would guarantee an acceptable result.

Our work finds its reason in this context, i.e. we want to know for how long we should run a simulation (or analyse its results) on the ARCHER HPC platform to capture the same magnitude of climate variability exhibited by the reference simulation on the Met Office supercomputer (see also responses to Ref.n1 point 1), as we know that the two runs do not bit-compare.

We have now made this clear in the manuscript. Page 2, line 8-17:

"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."

and page 8 line 19-24:

"The large differences observed on time-scales shorter than 200 years are a direct consequence of machine-dependent processes (compiler, machine architecture etc., see section 2 and 3.1 for details),

but eventually exist because of the chaotic nature of the system. 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)."

Please see also responses to Ref.n1 point 2 for details about the model porting and tests performed on IC ensembles, and responses to Ref.n1 point 6 for details on differences on single machine.

2) "Therefore the question to be asked of the two simulations discussed here is whether the simulations are distinguishable from an IC ensemble on a single machine, not whether they diverge at all".
   We added a section to the manuscript; see page 9, starting at line 15 (and Figure 11e and 11f). In this section, we now show the signal-to-noise ratio computed by taking the differences between two 100-year periods for each simulation (note that each 100-year period can be considered as an ensemble member run on the same machine). This shows that differences between the two machines are comparable to differences among ensemble members run on a single machine. This confirms that the differences we observe between PI_MO and PI_AR, although triggered by the different computing environment, are largely dominated by the internal variability. This also highlights that 100 years is a too-short length for constant-forcing simulations on the same, or on a different, machine.

3) "However, the results presented here demonstrate that the climatology of the two simulations is the same - given a long enough averaging period the simulations are indistinguishable. This is a good result, however, it is not the conclusion that the authors come to."

The first result presented in the paper (in the revised manuscript page 7, line 5-7) shows that on multi-centennial time-scales the differences are not significant and the long-term statistics of the two runs are similar. We have now given more strength to this result. Page 8, line 15-19:

"In summary, although large differences can be observed at smaller time-scales (see next section for further discussion), the climate of PIMO and PIAR is indistinguishable on the 200-year time-scale. We thus conclude that simulations using the HadGEM3-GC3.1 model are reproducible on different HPC platforms, provided that a long-enough simulation length is used. Our results also provide a strong indication that HadGEM3-GC3.1 does not suffer from code/compiler bugs that would make the model behave differently on different machines "

4) "Given this, the analysis in sections 3 and 4 are of little interest."

Section 4 answers our question of how long we should run a copy–simulation on the ARCHER platform. In this section, we show that only when using a 200-year averaging period we capture the same internal variability in both simulations. This is the main result of our paper, which implies not only that running a constant-forcing simulation for less than 100 years may potentially lead to different outcomes but also that running it for longer may not be necessary (this applies however only to those variables considered in the paper).

The additional analysis done on the ENSO signal is meant to be a demonstration of what physical process can cause such big differences. However, we agree with the reviewers that it is not surprising

that a low-frequency process like ENSO is the one still showing differences on a 100-year timescale. This remark was added at page 9, line 13-14.

As both reviewers agree that this section is of lesser interest, we have shortened section 4.2 in the revised manuscript.

Finally, we believe that the reference to the CMIP project is appropriate. CMIP is much more than Historical and Scenario simulations (for which ensembles are requested). Many individual MIPs run constant-forcing simulations with varying lengths depending on time availability, computational resources and length requirements. The CMIP6 minimum run length requirement for many of the Model Intercomparison Projects (MIPs) is 100 years. Our results suggest that 100 years may not be long to capture correctly the internal variability of the HadGEM3-GC3.1 model.

- **Specific Comments**

Please refer to the revised manuscript, as Introduction and Conclusions have substantially changed since the previous version.

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 simulation 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 all 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 cannot detect code bugs, which may cause a same model to behave differently on different machines. The centennial simulations presented in this paper will help understanding whether or not such code bugs exist for 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} - PI_{AR}) (σ_{MO-AR}) divided by $\sqrt{2}$ (see below for details). Thus, SNR is defined as:

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

when SNR < 1, (PI_{MO} - PI_{AR}) differences can be interpreted as fluctuations of the system not necessarily linked to the different computing environment (i.e. PI_{MO} and PI_{AR} do not differ more than PI_{MO} (or PI_{AR}) evaluated at two different points in time). When SNR > 1, the observed differences are outside of the expected range of variability and PI_{MO} and PI_{AR} are considered to be different.

Note that (2) makes use of two mathematical assumptions: PI_{MO} and PI_{AR} are uncorrelated (i.e. their covariance is zero), and have the same variance. Under these assumptions, the noise can be represented as the standard deviation of the differences between PI_{MO} and PI_{AR} divided by $\sqrt{2}$.

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.

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

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 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).

As the timescale increases, (PI_{MO} - PI_{AR}) differences get smaller and approach zero when a 200-year timescale is considered. This happens because 200 years is a long enough averaging-period for the positive and negative extremes in the time-series of Figure 4 - 9 to average out. On shorter time-intervals, strong increasing/decreasing trends in one simulation may not be compensated by trends of opposite sign in the the other simulation and may result in SNR » 1. See for example the first 10 years of the NH SIA time-series in Figure 5a. Additionally, the 200-year mean of the SIA seasonal cycle shown in Figure 5c is almost identical for ARCHER and MO, confirming that on a 200-year timescale the two runs are comparable. This suggests that the overall physical behaviour of the model has not been affected by the porting.

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.

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, all 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, suggests a consistent scaling of (PI_{MO} - PI_{AR}) differences with the timescale that approaches a plateau near the 200-year time-scale.

SIA (not shown) was the only variables 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 simulations using the HadGEM3-GC3.1 model are reproducible on different HPC platforms, provided that a long-enough simulation length is used.

Our results also provide a strong indication that HadGEM3-GC3.1 does not suffer from code/compiler bugs that would make the model behave differently on different machines.

4.2 The 100-year timescale

The large differences observed on time-scales shorter than 200 years are a direct consequence of machine-dependent processes (compiler, machine architecture etc., see section 2 and 3.1 for details), but eventually exist because of the chaotic nature of the system. 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 many of the Model Intercomparison Projects (MIPs) is 100 years or less, with no ensembles required. 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).

[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.

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.

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, 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. This suggests a consistent time-dependent scaling of (PI_{MO} - PI_{AR}) differences across the whole climate simulation.

Larger inconsistencies between the two runs were found for shorter timescales (where SNR $\geq$ 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 $\geq$ 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 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.

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).

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.

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.

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.

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.

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.

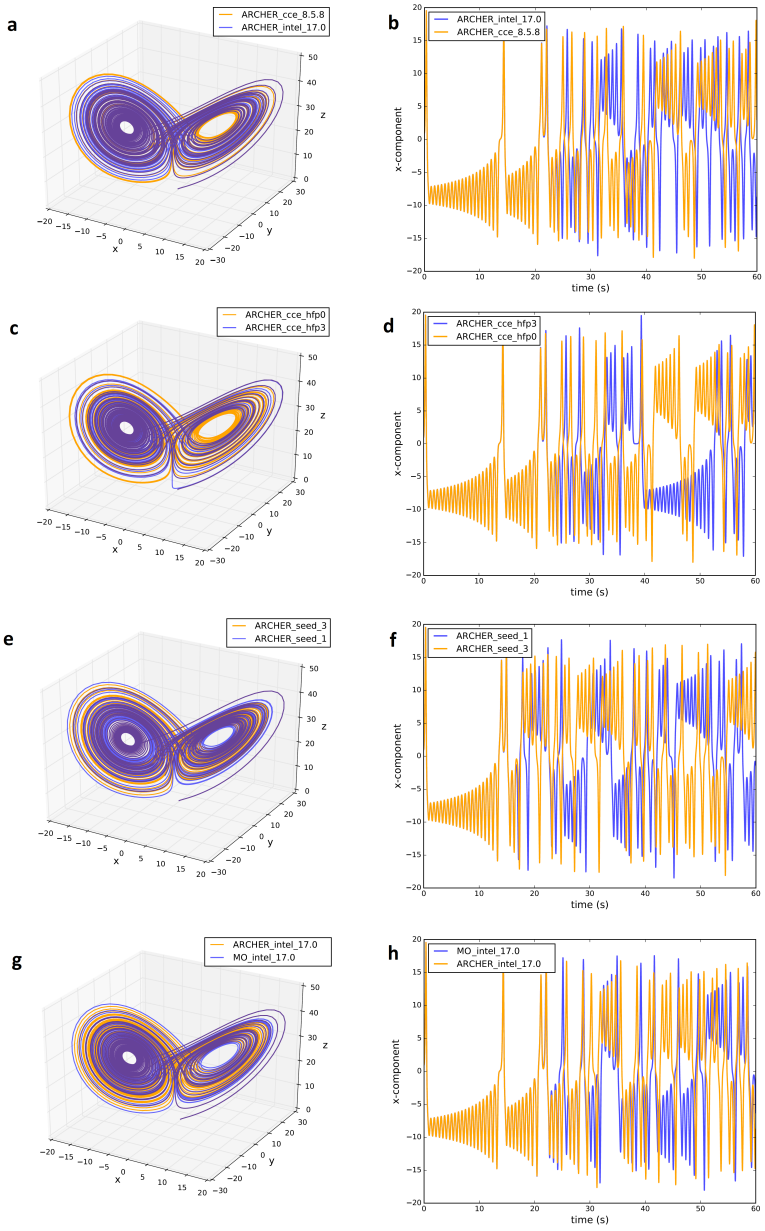

Figure 1. Attractor (left-hand side) and time-series of the x-component (right-hand side) of the 3D Lorenz model for simulations run on ARCHER using: the cce8.3.4 and intel17.0 compilers (a, b), same compiler but different level of floating-point optimization (c, d), same compiler and compiling options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

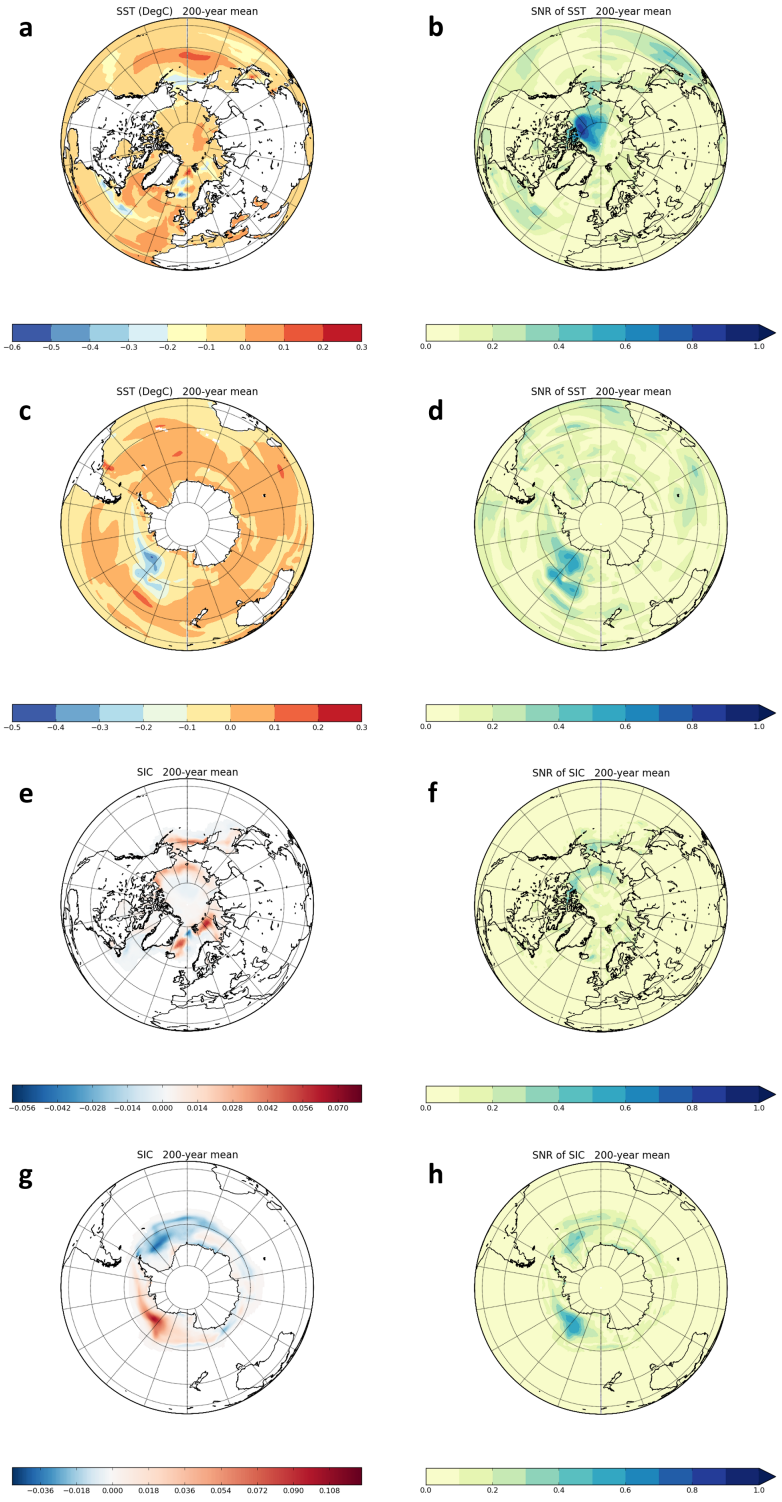

Figure 2. 200-year means and corresponding SNR of (PI_{MO} - PI_{AR}) differences for NH SST (a, b), SH SST (c, d), NH SIC (e, f) and SH SIC (g, h).

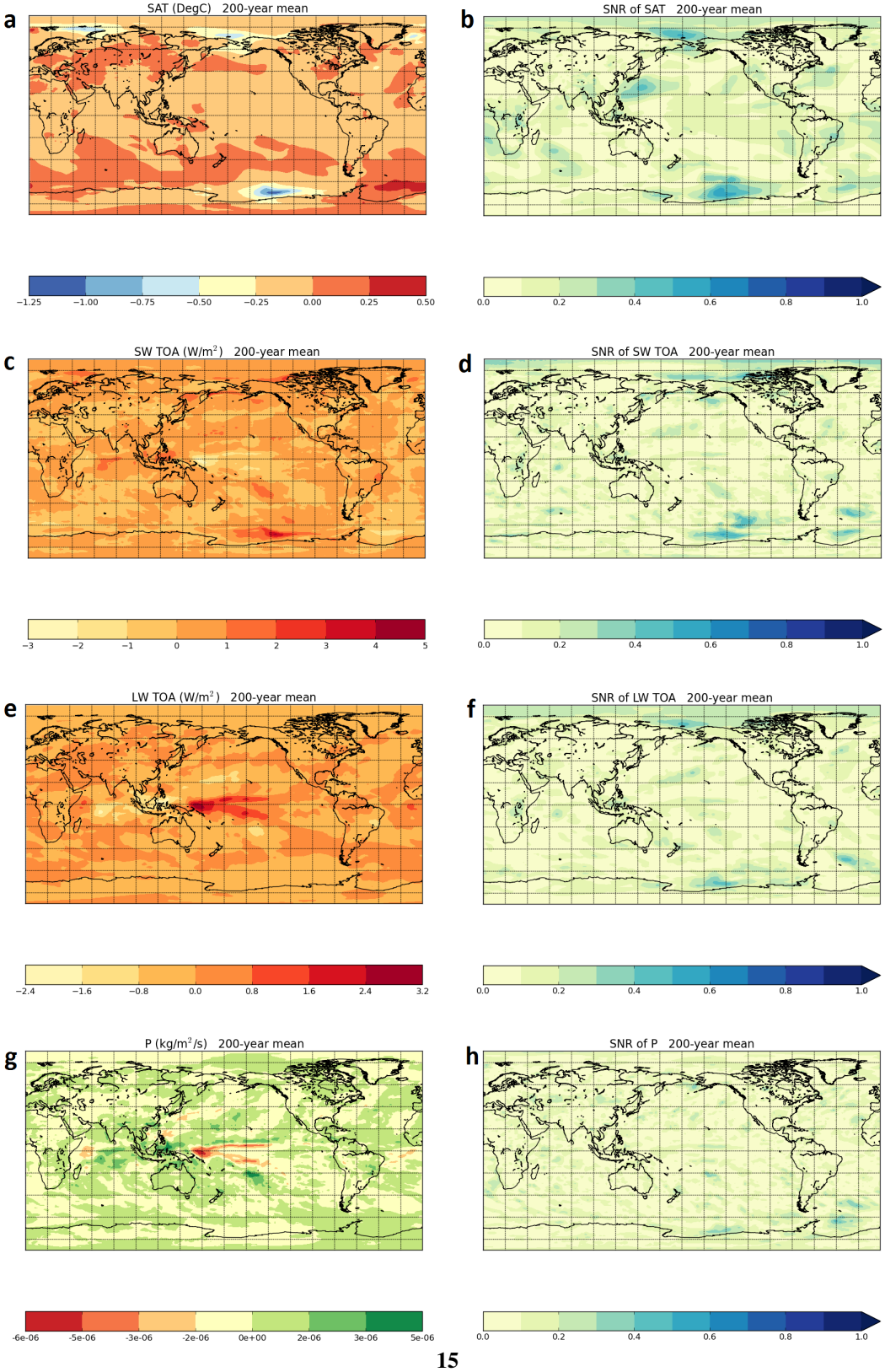

Figure 3. 200-year means and corresponding SNR of (PI_{MO} - PI_{AR}) differences for SAT (a, b), SW TOA (c, d), LW TOA (e, f) and P (g, h).

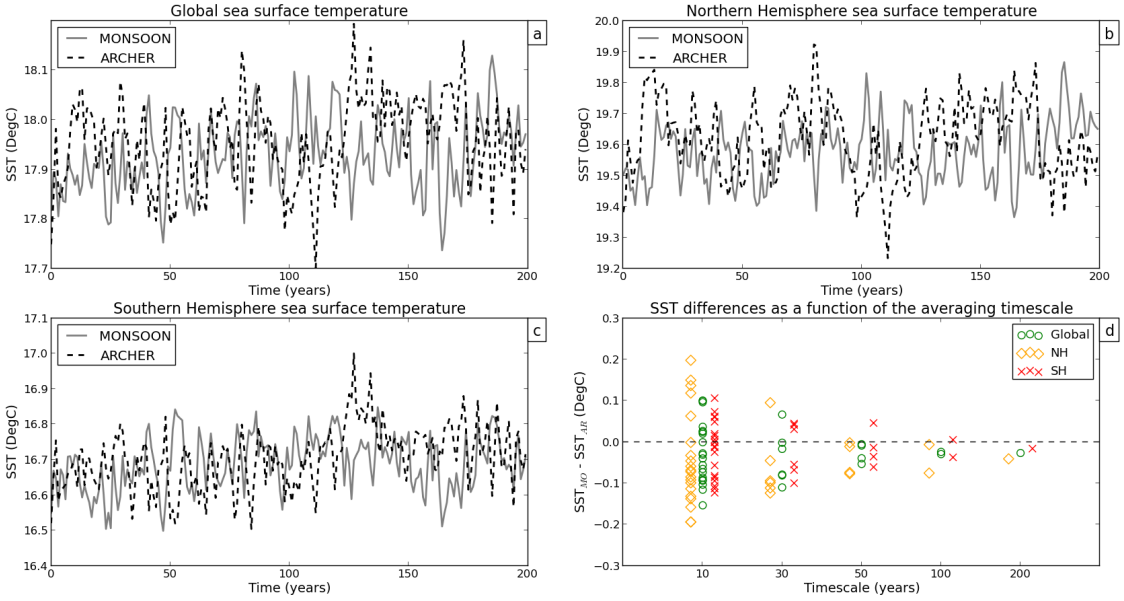

Figure 4. Annual-mean time-series of Global SST (a), Northern Hemisphere SST (b) and Southern Hemisphere SST (c) for PI_{MO} (grey line) and PI_{AR} (dashed line). d shows how SST differences vary as a function of the timescale.

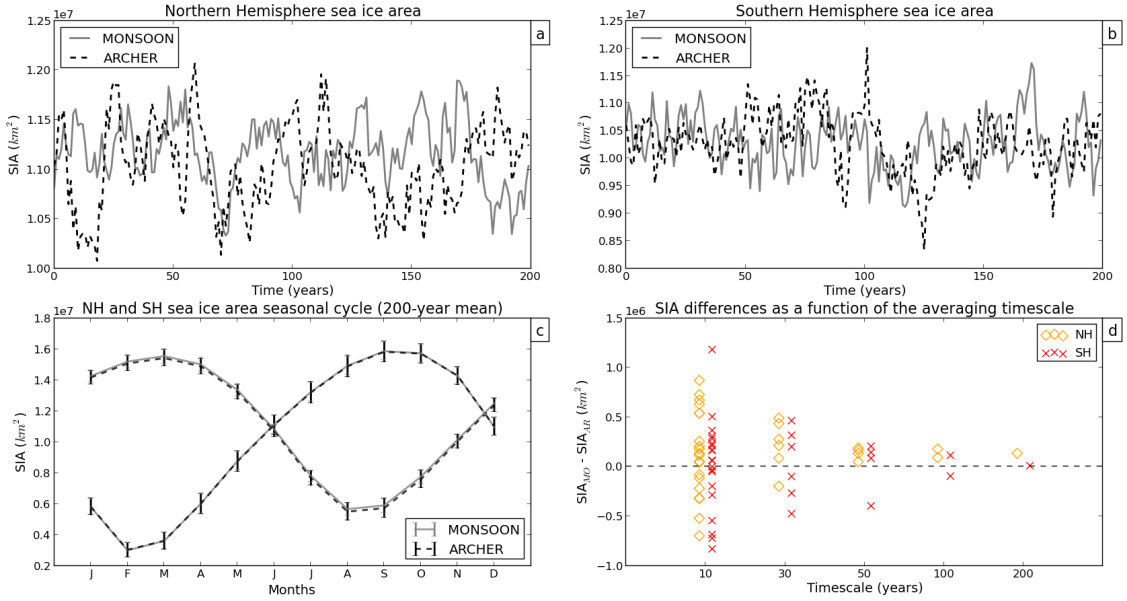

Figure 5. Annual-mean time-series of Northern Hemisphere SIA (a) and Southern Hemisphere SIA (b) for PI_{MO} (grey line) and PI_{AR} (dashed line). The 200-year mean of the NH and SH SIA seasonal cycle is shown in c. d shows how SIA differences vary as a function of the timescale.

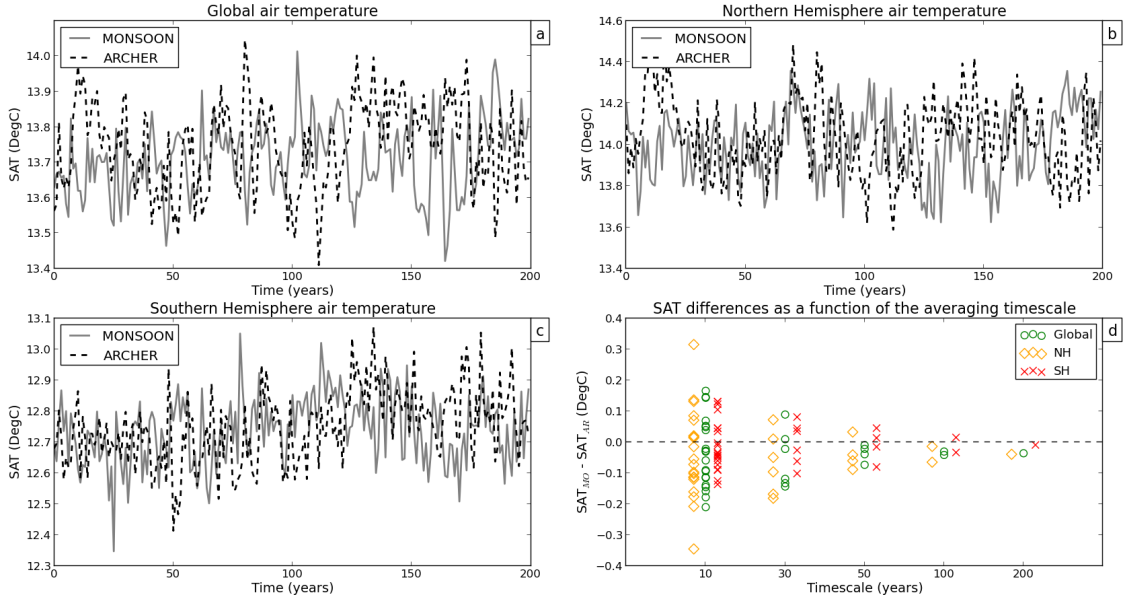

Figure 6. As in 4 but for SAT.

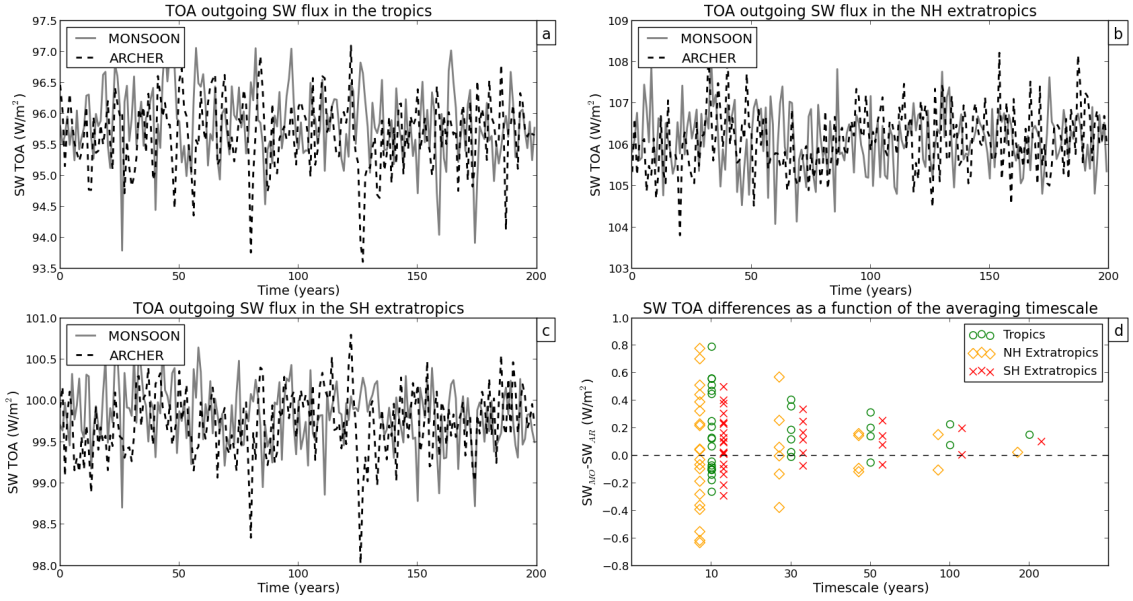

Figure 7. Annual-mean time-series of SW TOA in the tropics (a), SW TOA in the Northern Extratropics (b) and SW TOA in the Southern Extratropics (c) for PI$_{MO}$ (grey line) and PI$_{AR}$ (dashed line). d shows how SW TOA differences vary as a function of the timescale.

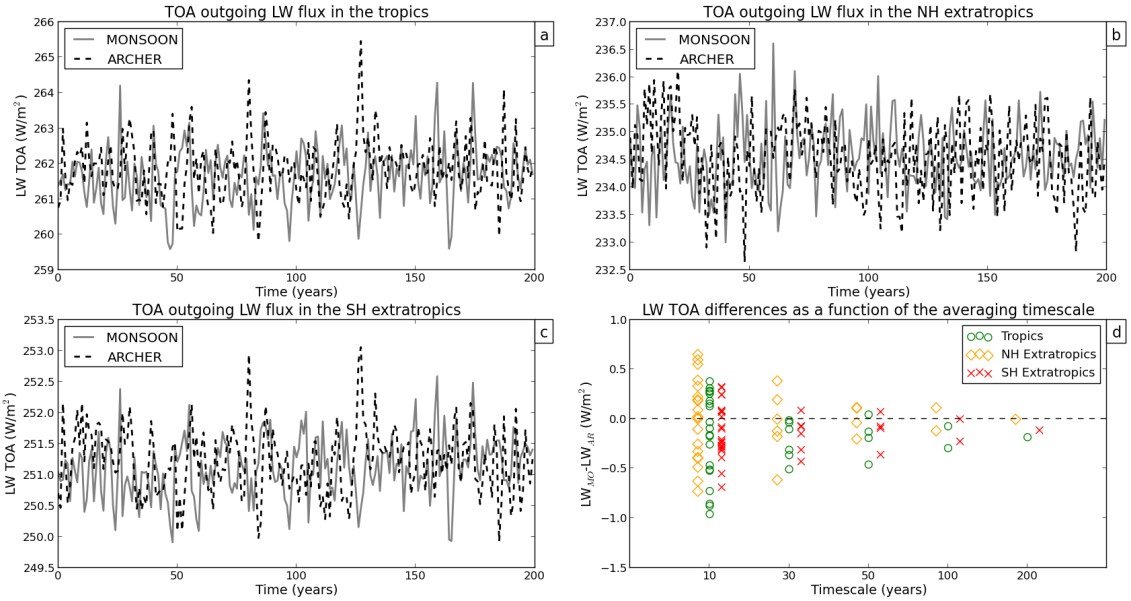

Figure 8. As in 4 but for LW TOA.

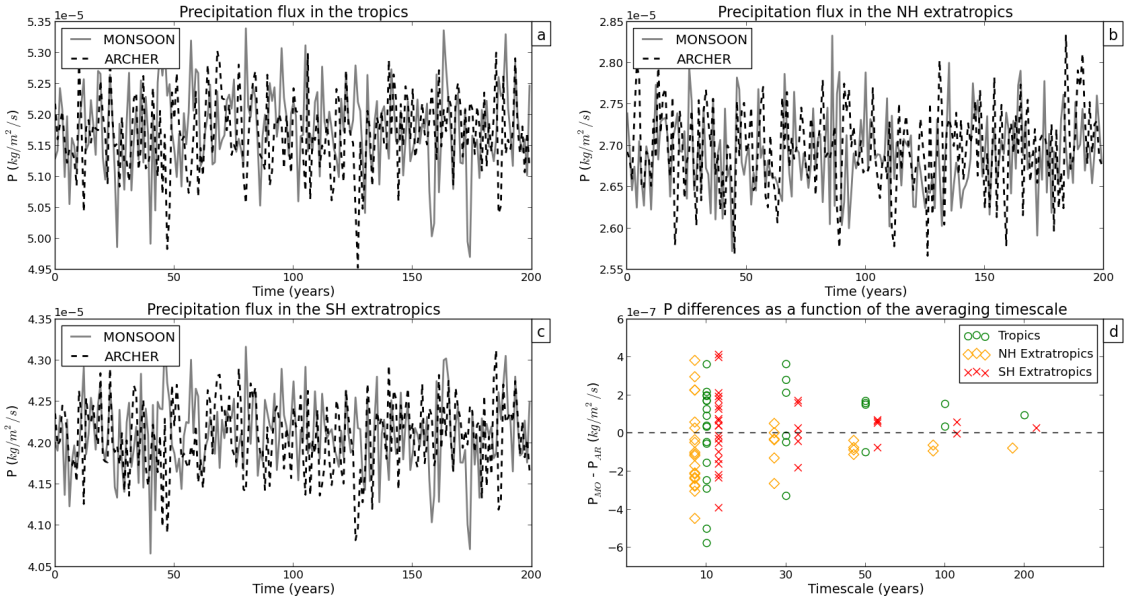

Figure 9. As in 4 but for P.

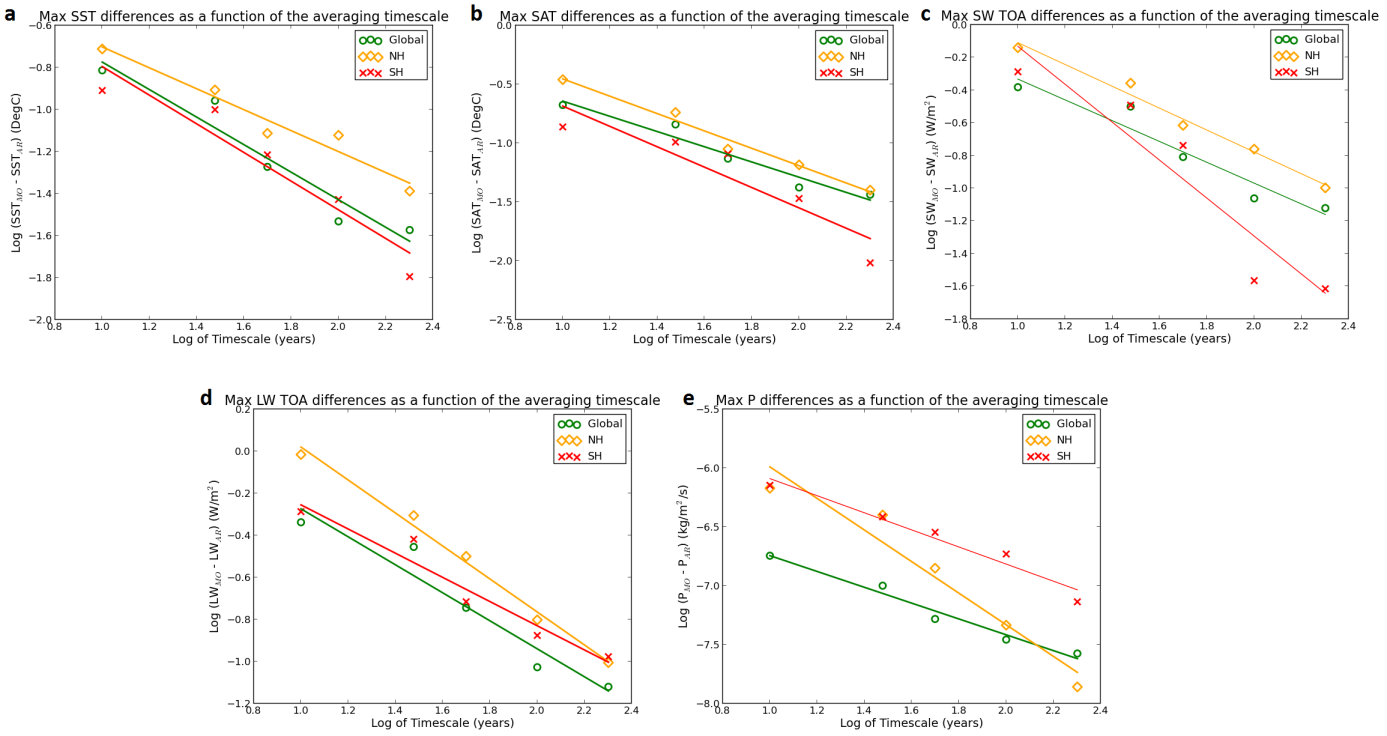

Figure 10. Log-log plots of SST (a), SAT (b), SW TOA (c), LW TOA (d) and P (e) representing maximum (PI_{MO} - PI_{AR}) differences as a function of the timescale. All the variables were averaged globally (green circles) and over the SH (red crosses) and NH (yellow diamonds) Hemispheres. The straight-lines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

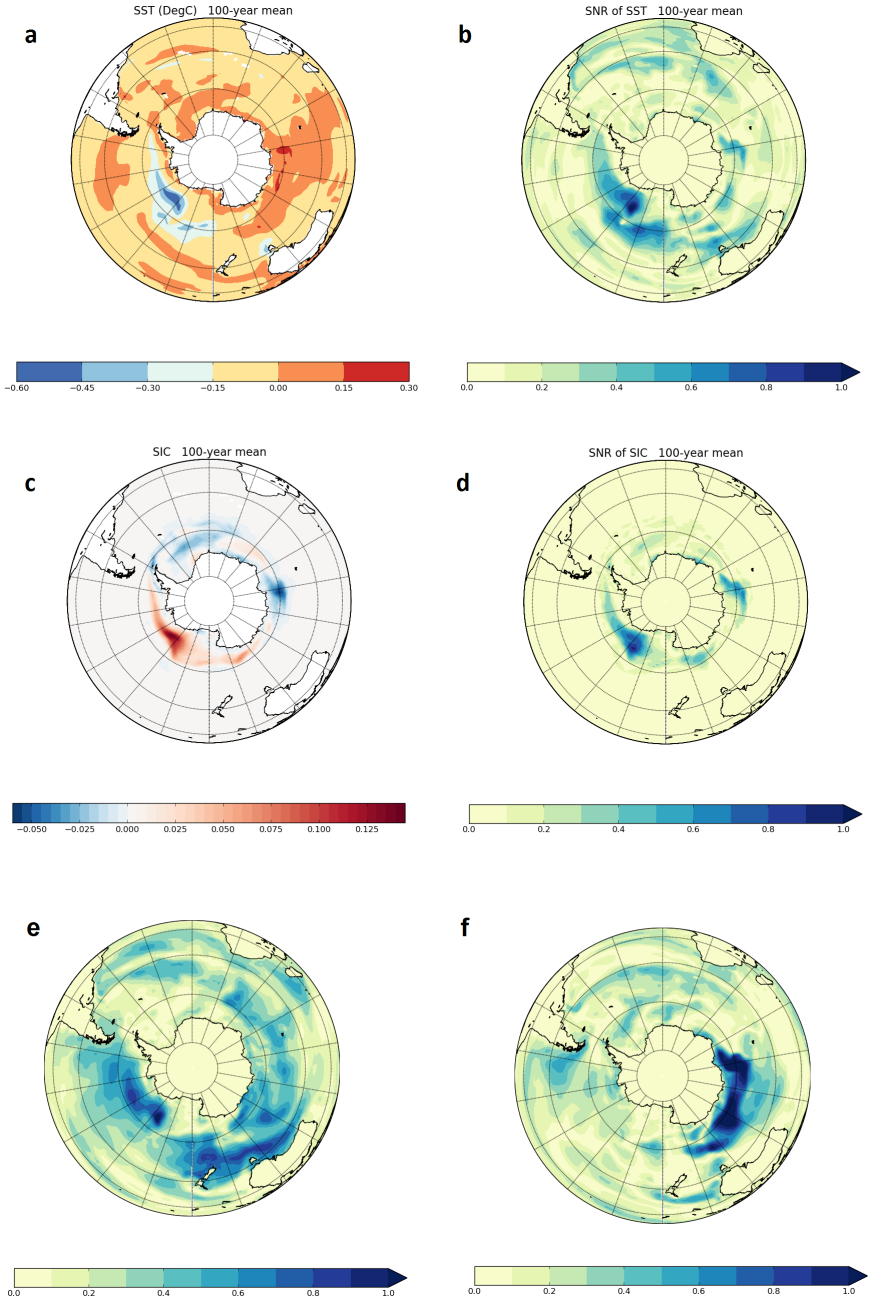

Figure 11. 100-year means and corresponding SNR of (PI_{MO} - PI_{AR}) differences for SH SST (a, b) and SH SIC (c, d). e and f show SNR of (PI_{AR}1st - PI_{AR}2nd) and (PI_{MO}1st - PI_{MO}2nd) differences for SH SST respectively.

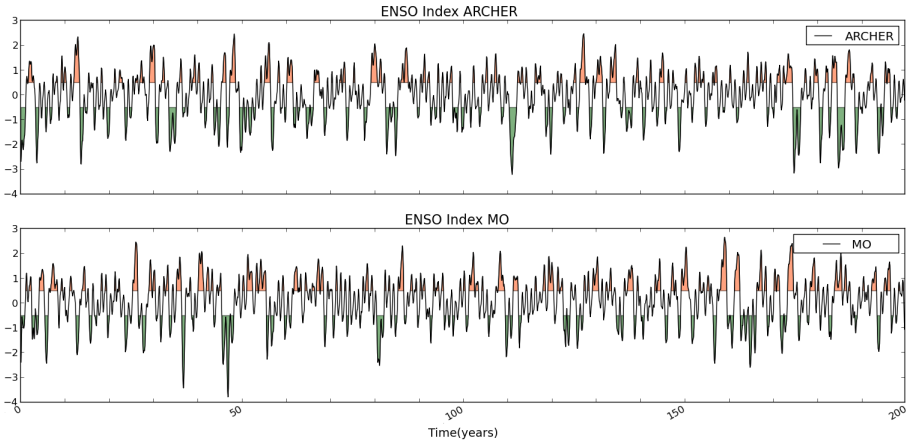

Figure 12. The NINO3.4 index for PI_{MO} and PI_{AR}. A 3-month running mean was applied to the ENSO signal and values greater/smaller than or equal to $\pm$ 0.5 are shaded in orange/green.