Geosci. Model Dev., 13, 139–154, 2020
https://doi.org/10.5194/gmd-13-139-2020

Geoscientific Model Development
Open Access
EGU

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)

Received: 28 March 2019 – Discussion started: 16 May 2019
Revised: 29 November 2019 – Accepted: 4 December 2019 – Published: 16 January 2020

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 the 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 (Coupled Model Intercomparison Project Phase 6) 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 sufficiently long duration of simulation is used.

In regions where El Niño–Southern Oscillation (ENSO) teleconnection patterns were detected, we found large sea surface temperature and sea ice concentration differences on centennial timescales. This indicates that a 100-year constant-forcing climate 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 for many MIP (Model Intercomparison Project) experiments.

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 (Model Intercomparison Project) 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 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 model 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 have 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. (2015a, b) 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 timescales (from decadal to centennial) in order to identify an averaging period and/or 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 an experiment because, provided that the simulation is long enough, it will return a picture of the natural climate variability.

The remainder of the paper is organized as follows. In Sect. 2, mechanisms by which the computing environment can influence the numerical solution of chaotic dynamical systems are reviewed and discussed. In Sect. 3, the numerical simulations are presented, and the methodology used for the data analysis is described. In Sect. 4, the simulation results are presented and discussed. In Sect. 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 3-D Lorenz model (Lorenz, 1963), which is a simplified model for convection in deterministic flows. The Lorenz model consists of the following three differential equations:

$$\frac{\mathrm{d}x}{\mathrm{d}t} = \alpha(y - x), \tag{1}$$

$$\frac{\mathrm{d}y}{\mathrm{d}t} = \gamma x - y - zx, \tag{2}$$

$$\frac{\mathrm{d}z}{\mathrm{d}t} = xy - \beta z, \tag{3}$$

where the parameters $\alpha = 10$, $\gamma = 28$, and $\beta = 8/3$ were chosen to allow for 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 first demonstrate the implications of switching between different computing environments, the Lorenz model was run on the ARCHER platform using the following:

- two different FORTRAN compilers (cce8.5.8 and intel17.0; see Fig. 1a and b);
- same FORTRAN compiler (cce8.5.8) but different degrees of floating-point optimization (`-hfp0` and `-hfp3`; see Fig. 1c and d); and
- the same FORTRAN (cce8.5.8) compiler and compiling options, but the x component in Eqs. (1)–(3) 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 Fig. 1e and f.

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) (Fig. 1g and h).

The divergence of the solutions in Fig. 1a and b can likely be explained by the different "computation order" of the two compilers (i.e. the order in which the same arithmetic expression is computed). In Fig. 1c and d, solutions differ because of the round-off error introduced by the different precision of floating-point computation. In Fig. 1e and f, 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 parameterizations for initialization and perturbation purposes (e.g. clouds, radiation, and turbulence parameterizations) as well as in stochastic parameterizations. The processes by which initial seeds are selected within the model code are thus crucial in order to ensure numerical reproducibility. Furthermore, different compilers may have different default seeds.

As for Fig. 1g and h, 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 Fig. 1g and h the two solutions diverge in time similarly to Fig. 1a–d; however, identifying reasons for the observed differences is not straightforward. While we speculate that reasons may be down to the machine architecture and/or chip set, further investigations on the subject were not pursued as this would be beyond the scope of this study.

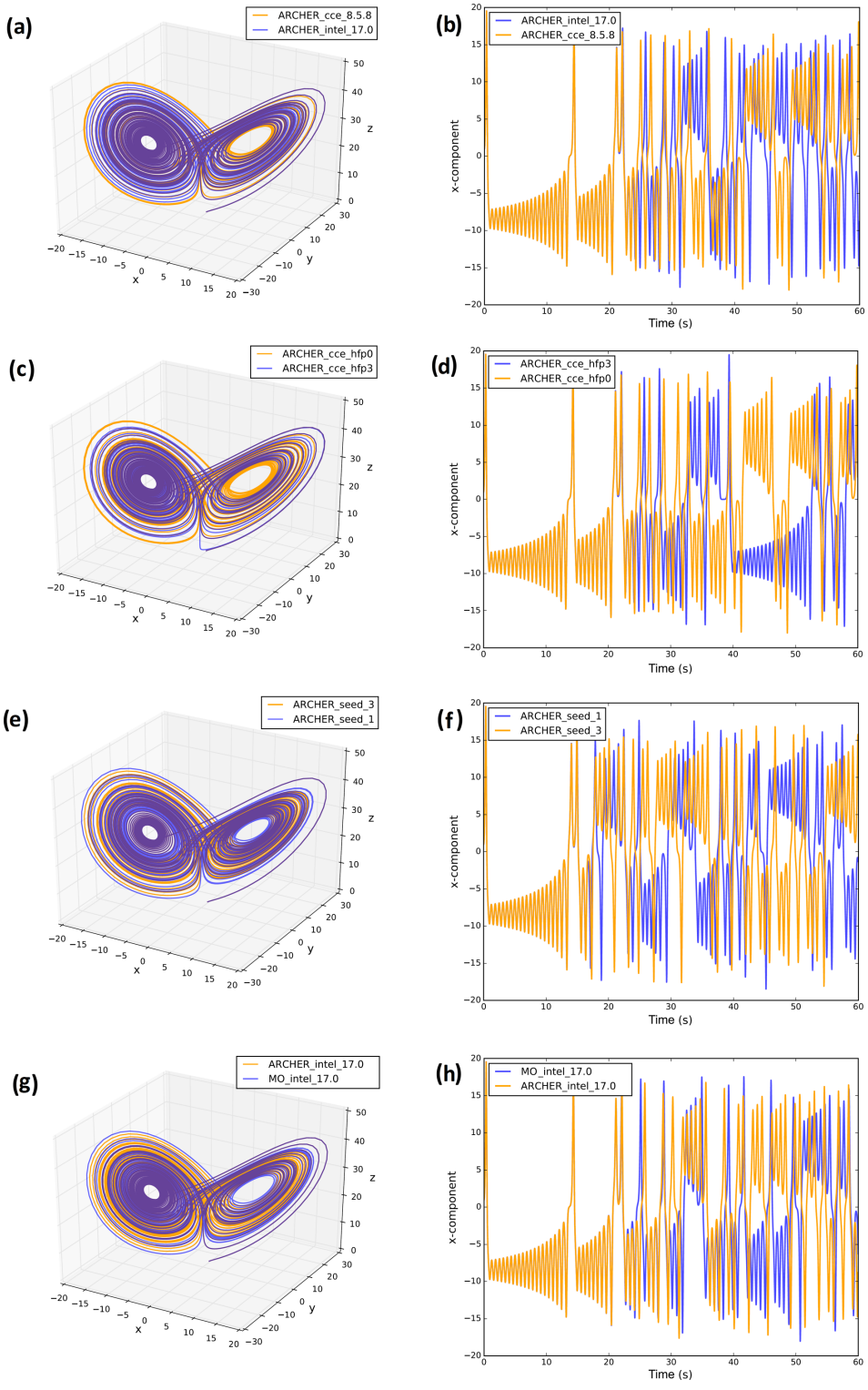

Figure 1. Attractor (left-hand side) and time series of the x component (right-hand side) of the 3-D Lorenz model for simulations run on ARCHER using the cce8.5.8 and intel17.0 compilers **(a, b)**, the same compiler (cce8.5.8) but a different level of floating-point optimization (hfp0, hfp3) **(c, d)**, and the same compiler (cce8.5.8) and compiling options but a different seed for the random number generator (seed 1, 3) **(e, f)**. Panels **(g)** and **(h)** are the Lorenz attractor and the x component time series for the Lorenz model run on MO and ARCHER using the same compiler (intel17.0) and compiling options.

The three mechanisms discussed above were selected because they are 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 the number of processors and processor decomposition, communications software (i.e. MPI libraries), and threading (i.e. OpenMP libraries).

We conclude this section by stressing that the four case studies presented in Fig. 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 or error into the system (no matter by which means), 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 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 and the NEMO Team, 2015), and the CICE sea ice model (Ridley et al., 2018b). The UM vertical grid contains 85 pressure levels (terrain-following hybrid height coordinates), while the NEMO vertical grid 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 orthogonal 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 greenhouse gas (GHG), 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 a 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), and 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 8.5.8	Ivy Bridge

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

Of the possible mechanisms discussed in Sect. 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 h 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, longwave radiation, 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 (David Case, National Centre for Atmospheric Science, University of Reading, Reading, UK, personal communications, 2019). However, this method is restricted to timescales shorter than 1 d. The centennial simulations presented in this paper will help us understand whether or not differences can arise on longer timescales 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 and sea ice concentration (SIA, SIC), 1.5 m air temperature (SAT), the outgoing longwave and shortwave 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, ($\mathrm{PI_{MO}} - \mathrm{PI_{AR}}$) 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 Sect. 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 extratropics, and southern extratropics according to the underlying physical processes.

Note that, when calculating ($\mathrm{PI_{MO}} - \mathrm{PI_{AR}}$) differences, $\mathrm{PI_{MO}}$ and $\mathrm{PI_{AR}}$ segments are subtracted in chronological order. Thus, for example, the first 10 years of $\mathrm{PI_{AR}}$ are subtracted from the first 10 years of $\mathrm{PI_{MO}}$ 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-year segment is equally representative of the pre-industrial decadal variability. We acknowledge that an equally valid alternative approach would be to subtract the $\mathrm{PI_{AR}}$ and $\mathrm{PI_{MO}}$ segments without a prescribed order.

Discrepancies in the results between the two runs were 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 $\mathrm{PI_{MO}}$ and $\mathrm{PI_{AR}}$ ($\mu_{\mathrm{MO-AR}}$), and the noise is represented by the standard deviation of $\mathrm{PI_{MO}}$ (σ_{MO}), our "reference" simulation. Because of the basic properties of variance, for which $\mathrm{Var}_{X-Y} = \mathrm{Var}_X + \mathrm{Var}_Y - 2\mathrm{Cov}(X, Y)$ (Loeve, 1977), we can more conveniently express the noise as $\sigma_{\mathrm{MO}} = \frac{\sigma_{\mathrm{MO-AR}}}{\sqrt{2}}$ under the assumption that $\mathrm{PI_{MO}}$ and $\mathrm{PI_{AR}}$ are uncorrelated ($\mathrm{Cov}(\mathrm{MO}, \mathrm{AR}) = 0$) and have the same variance ($\mathrm{Var_{MO}} = \mathrm{Var_{AR}}$). This allowed us to compute SNR on one grid and avoid divisions by (nearly) zero when the sea ice field between $\mathrm{PI_{MO}}$ and $\mathrm{PI_{AR}}$ evolved differently, resulting in unrealistically high SNR values along the sea ice edges. Finally, SNR is defined as

$$\mathrm{SNR} = \frac{|\mu_{\mathrm{MO-AR}}|}{\sigma_{\mathrm{MO}}} = \frac{|\mu_{\mathrm{MO-AR}}|}{\frac{\sigma_{\mathrm{MO-AR}}}{\sqrt{2}}}. \quad (4)$$

When $\mathrm{SNR} < 1$, ($\mathrm{PI_{MO}} - \mathrm{PI_{AR}}$) differences can be interpreted as fluctuations within the estimated range of internal variability. When $\mathrm{SNR} > 1$, ($\mathrm{PI_{MO}} - \mathrm{PI_{AR}}$) 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 Oscillation (ENSO) signal was computed for the ARCHER and MO simulations. We used the Niño 3.4 index, with a 3-month running mean, defined as follows:

$$\begin{aligned} \mathrm{NINO3.4} &= \mathrm{SST_{mnth}} - \overline{\mathrm{SST_{30\,yr}}} \\ &\text{if } 5\,^{\circ}\mathrm{N} \leq \text{latitude} \leq 5\,^{\circ}\mathrm{S} \text{ and} \\ &120\,^{\circ}\mathrm{W} \leq \text{longitude} \leq 170\,^{\circ}\mathrm{W}, \end{aligned} \quad (5)$$

where $\mathrm{SST_{mnth}}$ is the monthly sea surface temperature and $\overline{\mathrm{SST_{30\,yr}}}$ 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 Figs. 2 and 3. All the variables exhibit $\mathrm{SNR} < 1$, 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 Figs. 2 and 3 are computed using 10-, 30-, 50-, and 100-year averaging periods (not shown), the magnitude of the anomalies increases and ($\mathrm{PI_{MO}} - \mathrm{PI_{AR}}$) differences become significant ($\mathrm{SNR} \gg 1$). 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 ($\mathrm{PI_{MO}} - \mathrm{PI_{AR}}$) differences as a function of the averaging timescale for each variable (see Sect. 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, $\mathrm{PI_{MO}}$ and $\mathrm{PI_{AR}}$ start diverging quickly after the first few time steps once the system has lost memory of the initial conditions (Figs. 4 to 9, panels a, b, c). See Sect. 2 (Fig. 1) for a 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 Figs. 4d, 6d, 7d, 8d), while SIA anomalies are particularly high in the Southern Hemisphere (red crosses in Fig. 5d) and P anomalies in the tropics (green circles in Fig. 9d). Overall, discrepancies are the largest at decadal timescales at which the spread between the two simulations can reach $|0.2|\,^{\circ}\mathrm{C}$ in global mean air temperature (Fig. 6d), $|1.2|$ million km^2 in Southern Hemisphere

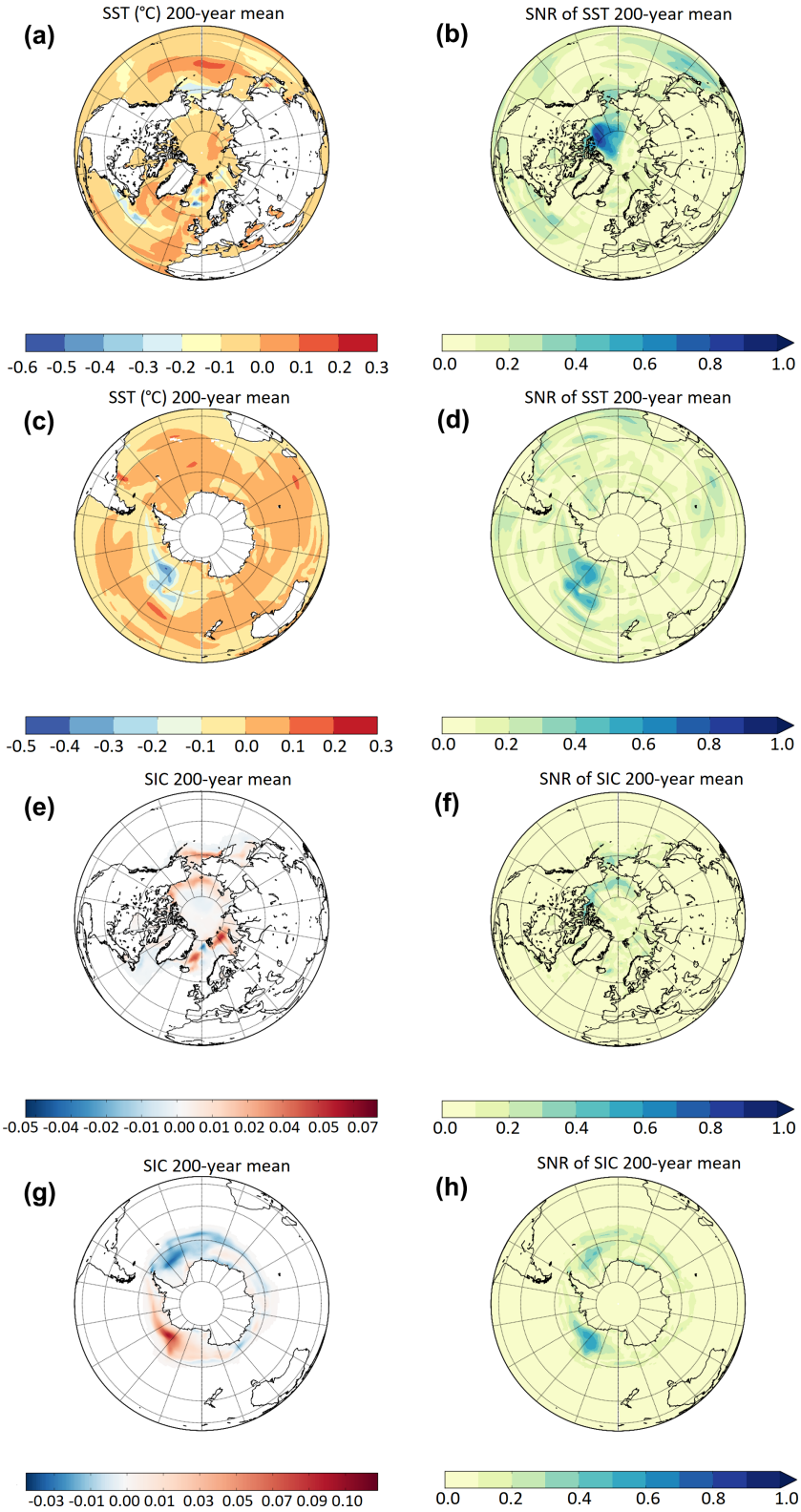

Figure 2. The 200-year means and corresponding SNR of ($PI_{MO} - PI_{AR}$) differences for NH SST **(a, b)**, SH SST **(c, d)**, NH SIC **(e, f)**, and SH SIC **(g, h)**.

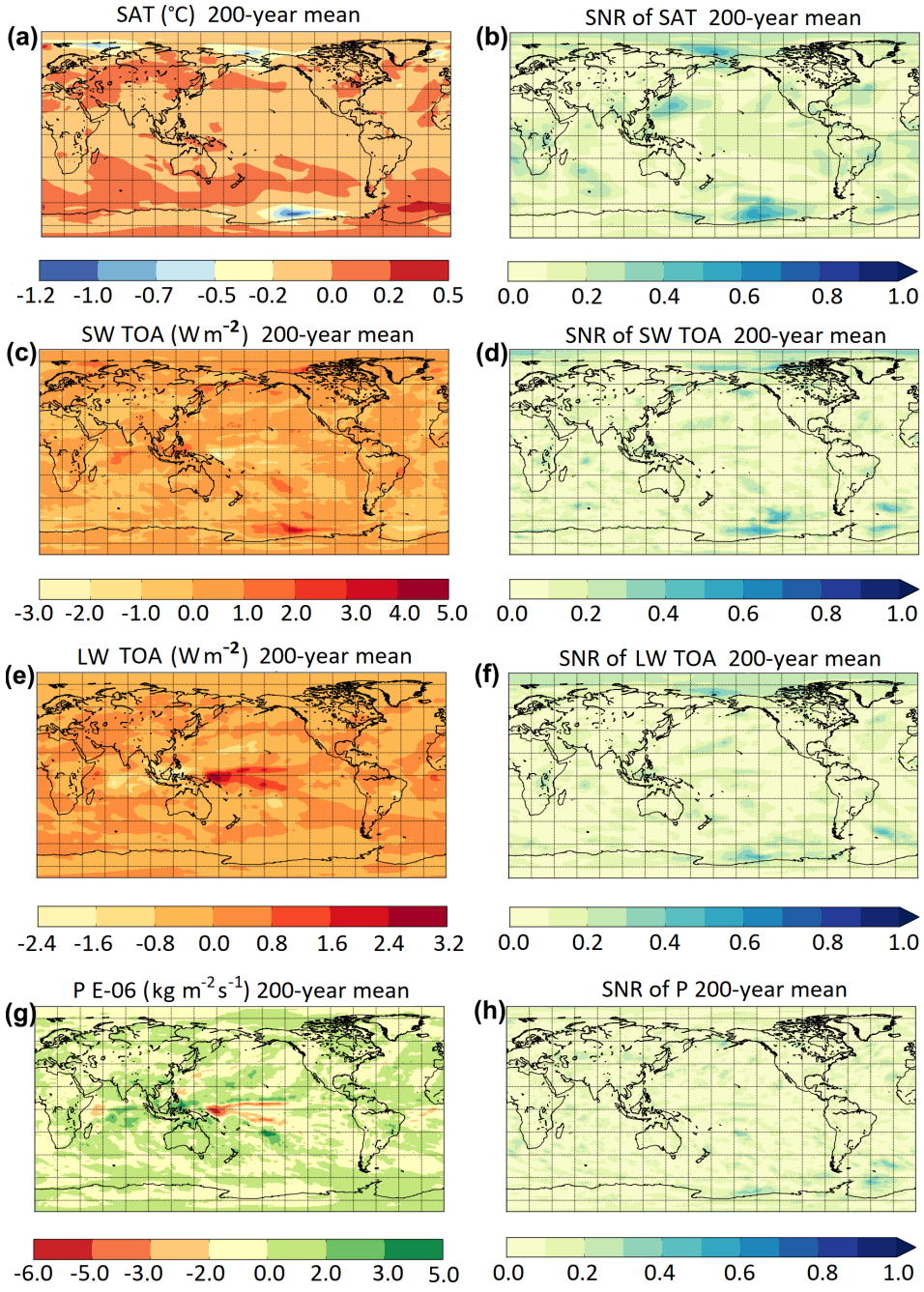

Figure 3. The 200-year means and corresponding SNR of ($PI_{MO} - PI_{AR}$) differences for SAT **(a, b)**, SW TOA **(c, d)**, LW TOA **(e, f)**, and P **(g, h)**.

sea ice area (Fig. 5d), or $|1|\,W\,m^{-2}$ in global TOA outgoing LW flux (Fig. 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 SNR $\gg 1$ are thus not surprising when analysing decadal periods.

On longer timescales, the estimates 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 as the timescale increases (Figs. 4d to 9d). When we consider the 200-year timescale, we find no SNR value greater than 1 (Figs. 2 and

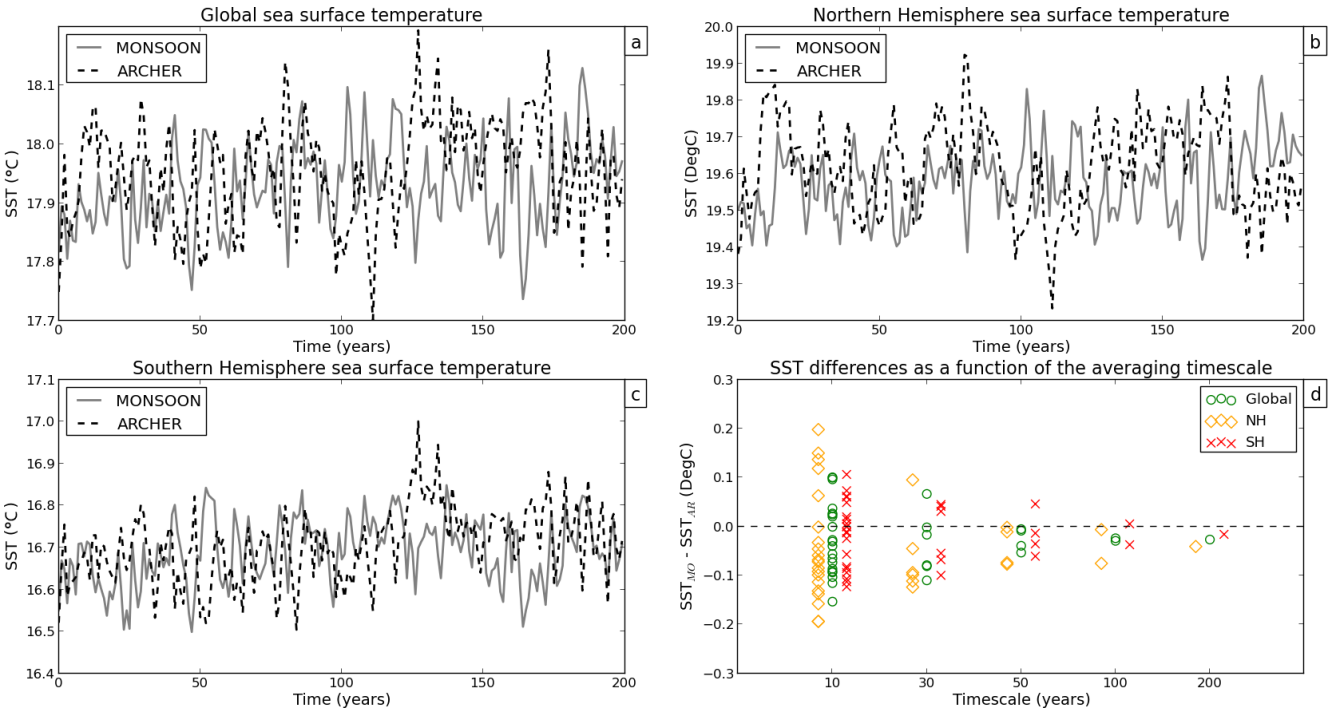

Figure 4. Annual mean time series of global SST **(a)**, Northern Hemisphere SST **(b)**, and Southern Hemisphere SST **(c)** for PI_{MO} (grey line) and PI_{AR} (dashed line). Panel **(d)** shows how SST differences vary as a function of the timescale.

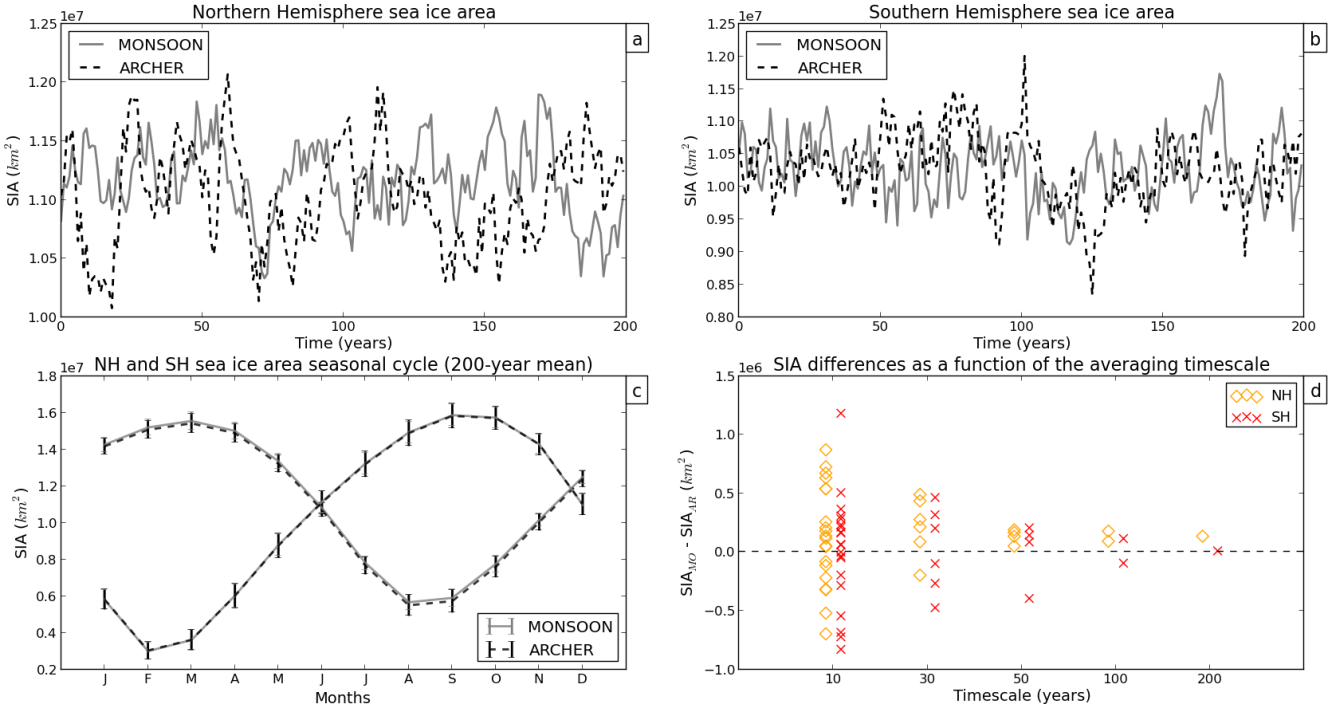

Figure 5. Annual mean time series of Northern Hemisphere SIA **(a)** and Southern Hemisphere SIA **(b)** for PI_{MO} (grey line) and PI_{AR} (dashed line). The 200-year mean of the NH and SH SIA seasonal cycle is shown in **(c)**. Panel **(d)** shows how SIA differences vary as a function of the timescale.

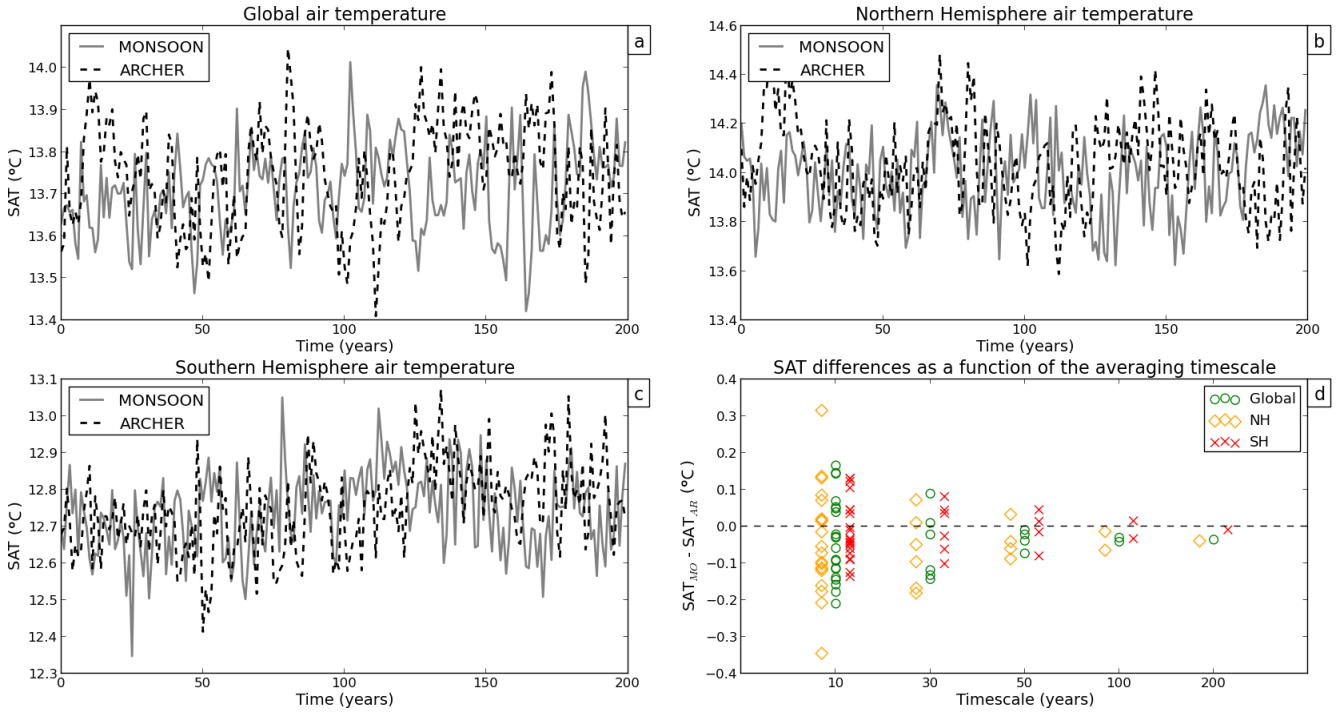

Figure 6. As in Fig. 4 but for SAT.

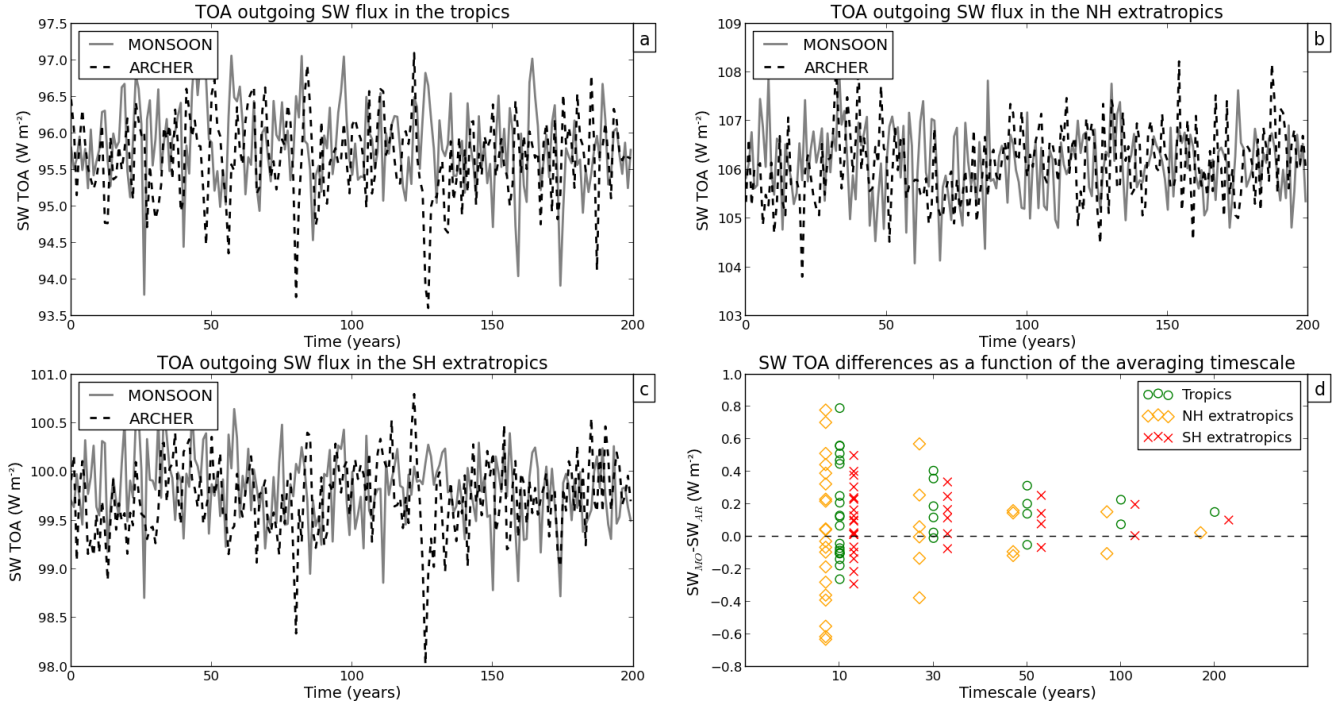

Figure 7. Annual mean time series of SW TOA in the tropics **(a)**, SW TOA in the northern extratropics **(b)**, and SW TOA in the southern extratropics **(c)** for PI_{MO} (grey line) and PI_{AR} (dashed line). Panel **(d)** shows how SW TOA differences vary as a function of the timescale.

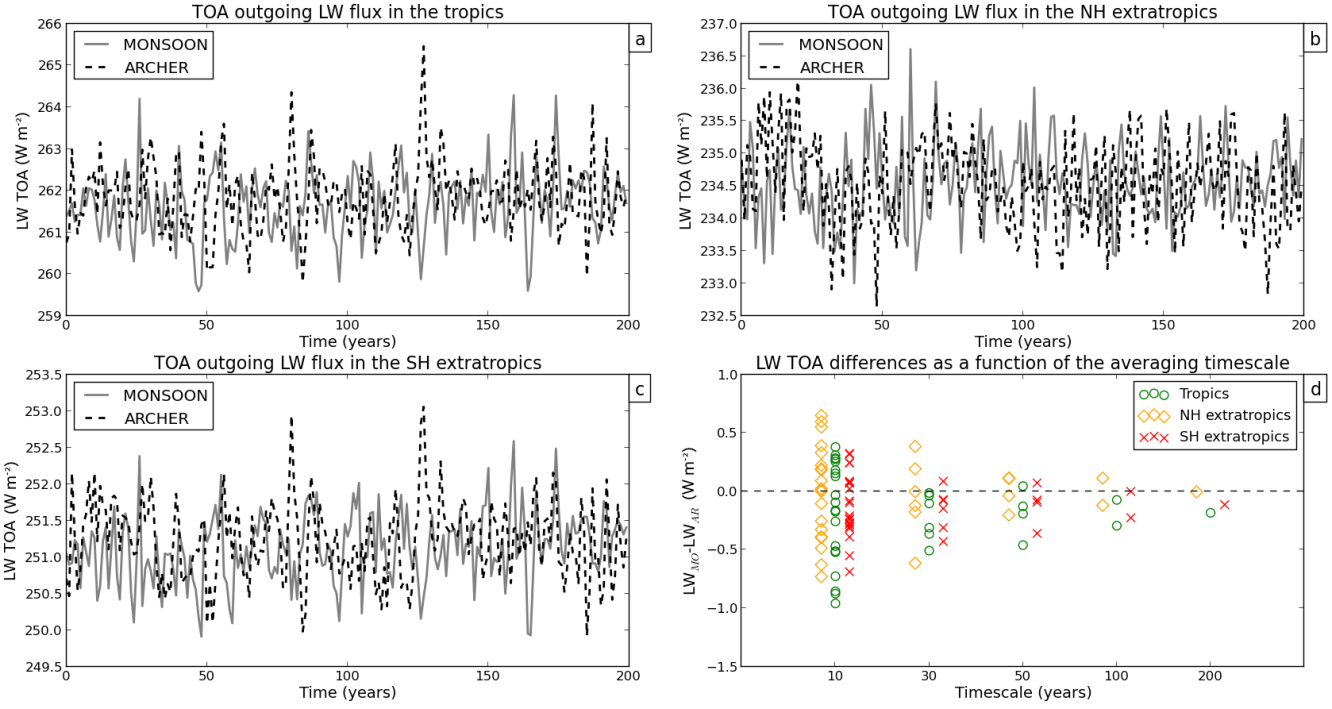

Figure 8. As in Fig. 4 but for LW TOA.

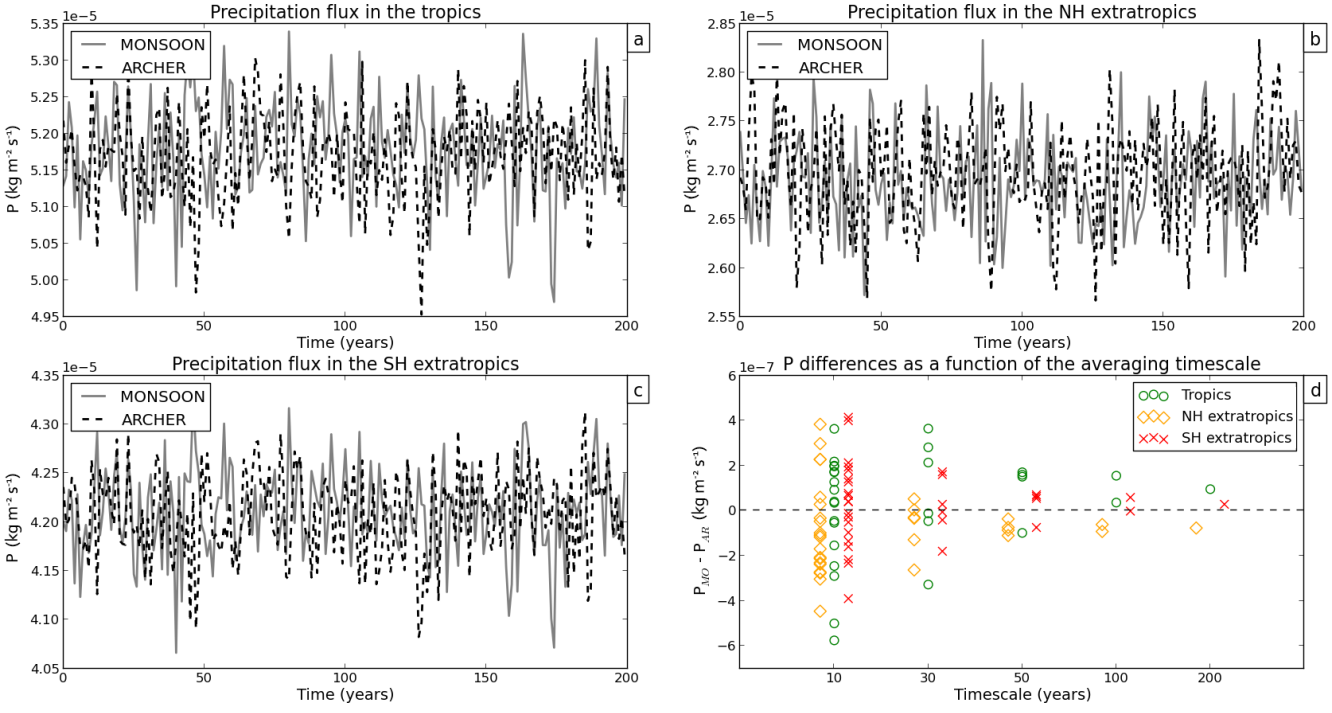

Figure 9. As in Fig. 4 but for P.

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

	MO	ARCHER
	Mean, SD	Mean, SD
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^{-1})	36.22, 0.12	36.25, 0.14

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 a 200-year-long period is considered.

In Figs. 4d to 9d, the variation of ($PI_{MO} - PI_{AR}$) differences with the timescale suggests the existence of a power-law relationship.[1] To investigate this behaviour, a base-10 logarithmic transformation was applied to the x and y axes in Figs. 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 Southern Hemisphere (SH) and Northern Hemisphere (NH). 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 on which ($PI_{MO} - PI_{AR}$) differences vary following the same power-law relationship, regardless of 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, and -0.67, respectively. Thus, the straight lines that best fit the global mean data in Fig. 10 all 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 ≈ 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 over 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 timescales (see the next section for a further discussion), the climate of PI_{MO} and PI_{AR} is indistinguishable on the 200-year timescale. We thus conclude that the mean climate properties simulated by the HadGEM3-GC3.1 model are reproducible on different HPC platforms, provided that a sufficiently long 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 h (for which the model was previously tested; see Sect. 3.1).

4.2 The 100-year timescale

The large differences observed on timescales 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 Sects. 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 timescales differences converge to zero (Figs. 4–9).

While in Sect. 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 timescale.

For instance, the minimum simulation length required by CMIP6 protocols for a few of the MIP experiments (excluding the DECK and Historical simulations) is 100 years or less, and ensembles are not always requested (e.g. some of the Tier 1, 2, and 3 experiments of 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 example, simulating 200 years with the HadGEM3-GC3.1 model on ARCHER in its CMIP6 configuration takes about 4 months).

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

In Fig. 11, ($PI_{MO} - PI_{AR}$) differences materialize into spatial patterns that are signatures of physical processes. SST (Fig. 11a, b) and SIC (Fig. 11c, d) anomalies are the largest in West Antarctica where ENSO teleconnection patterns are expected; they correspond to regions where SNR becomes equal to or larger than 1. This suggests that ($PI_{MO} - PI_{AR}$) differences are driven by two different ENSO regimes (the connection between SIC (and SST) anomalies in the South-

[1] Note that, for readability, the ticks of the x axes in Figs. 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.

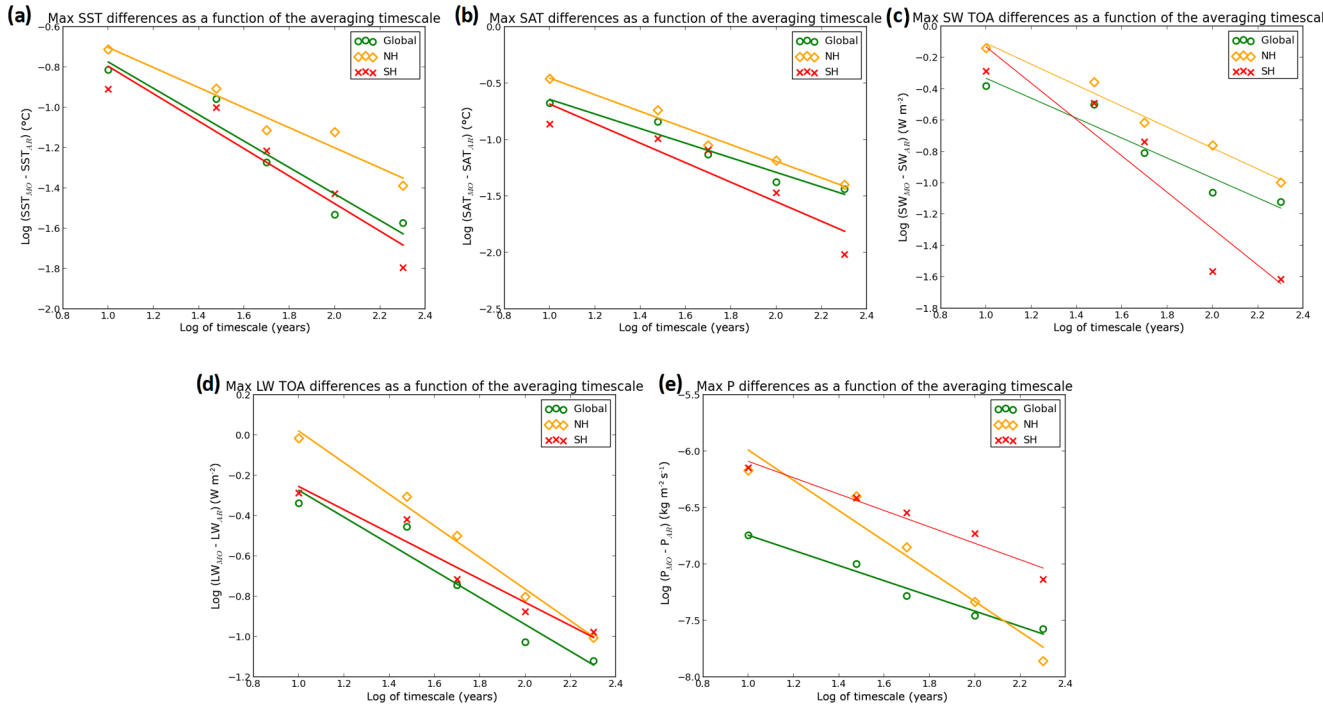

Figure 10. Log–log plots of SST **(a)**, SAT **(b)**, SW TOA **(c)**, LW TOA **(d)**, and P **(e)** representing maximum ($PI_{MO} - PI_{AR}$) differences as a function of the timescale. All the variables were averaged globally (green circles) and over the SH (red crosses) and NH (yellow diamonds). The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

ern Hemisphere and ENSO has been widely documented in the 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 Fig. 12. A few times, a strong El Niño (La Niña) event in PI_{MO} corresponds to 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 the 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 f show the signal-to-noise ratio corresponding to SST differences between $PI_{AR}1st$ and $PI_{AR}2nd$ and between $PI_{MO}1st$ and $PI_{MO}2nd$. In Fig. 11e, the SNR pattern exhibited by the ARCHER ensemble members resembles the one shown by ($PI_{MO} - PI_{AR}$) differences in Fig. 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 Fig. 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 too short a 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 Fig. 11 SNR $>$ 1).

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.

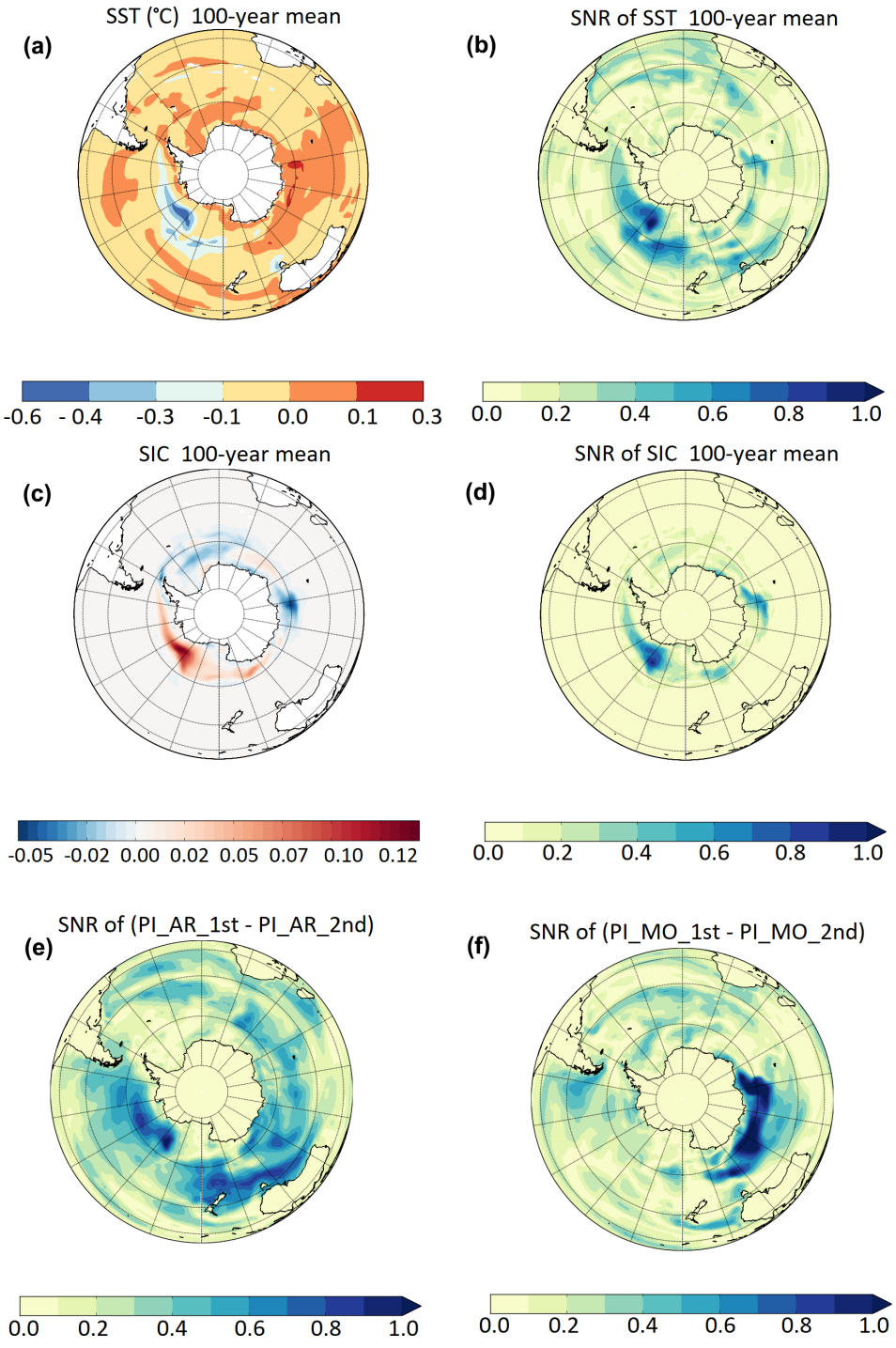

Figure 11. The 100-year means and corresponding SNR of ($PI_{MO} - PI_{AR}$) differences for SH SST **(a, b)** and SH SIC **(c, d)**. Panels **(e)** and **(f)** show SNR of ($PI_{AR}1st - PI_{AR}2nd$) and ($PI_{MO}1st - PI_{MO}2nd$) differences for SH SST, respectively.

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 Sect. 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 1. We conclude that in order for PI_{MO}

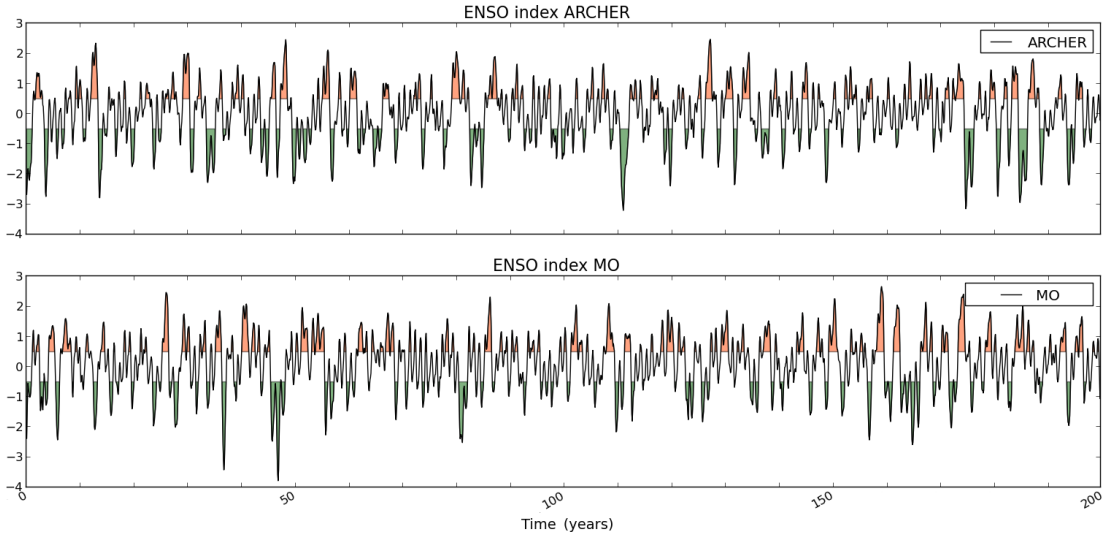

Figure 12. The Niño 3.4 index for PI_{MO} and PI_{AR}. A 3-month running mean was applied to the ENSO signal, and values greater and smaller than or equal to ± 0.5 are shaded in orange and green.

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 sufficiently long simulation length is used.

Additionally, the relationship between global mean differences and timescale exhibits a $\approx 2/3$ power-law behaviour, regardless the physical quantity considered, that approaches a plateau near the 200-year timescale. Thus, there is 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 of the physical processes that they represent.

Larger inconsistencies between the two runs were found for shorter timescales (at which SNR ≥ 1), the largest being at decadal timescales. For example, when a 10-year averaging period is used, discrepancies between the runs can be 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 environments.

On a 100-year timescale, large SST and SIC differences (with SNR ≥ 1) were 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 unexpected per se, 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 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 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 100 years (see, for example, the 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 awareness in the climate modelling community of the subject of the machine dependence of climate simulations.

Code availability. Access to the model code used in the paper has been granted to the editor. The source code of the UM is available under licence. To apply for a licence, go to http://www.metoffice.gov.uk/research/modelling-systems/unified-model (UK Met Office, 2020). JULES is available under licence free of charge; see

https://jules-lsm.github.io/ (Joint UK Land Environment Simulator, 2020). The NEMO model code is available from http://www.nemo-ocean.eu (NEMO Consortium, 2020). The model code for CICE can be downloaded from https://code.metoffice.gov.uk/trac/cice/browser (CICE Consortium, 2020).

Data availability. Access to the data used in the paper 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.HadGEM3-GC31-LL, https://doi.org/10.22033/ESGF/CMIP6.419; Ridley et al., 2018a), the data repository for all CMIP6 output. CMIP6 outputs are expected to be public by 2020. The 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, last access: 13 January 2020). Please contact the authors.

Author contributions. MVG ran the simulation on the ARCHER supercomputer, designed and carried out the tests in Sect. 2, and analysed all simulation results with the contribution of LCS and DS. GL and RH ported the HadGEM3 PI simulation to the ARCHER supercomputer, provided technical support, and advised on the nature of machine-dependent processes. All authors revised the paper.

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

Acknowledgements. We would like to thank the two anonymous referees and the topical editor, Sophie Valcke, for their time and their valuable comments.

Maria-Vittoria Guarino and Louise C. Sime acknowledge the financial support of NERC research grants NE/P013279/1 and NE/P009271/1. This work used the ARCHER UK National Supercomputing Service (http://www.archer.ac.uk, last access: 13 January 2020). The authors acknowledge the use of the UK Met Office supercomputing facility in providing data for model comparisons.

Financial support. This research has been supported by NERC (grant nos. NE/P013279/1 and NE/P009271/1).

Review statement. This paper was edited by Sophie Valcke and reviewed by two anonymous referees.

References

CICE Consortium: CICE, available at: https://github.com/CICE-Consortium last access: 13 January 2020.

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, Geosci. Model Dev., 9, 1937–1958, https://doi.org/10.5194/gmd-9-1937-2016, 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), Geosci. Model Dev., 9, 4019–4028, https://doi.org/10.5194/gmd-9-4019-2016, 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., Saenko, O. A., Stammer, D., and Winton, M.: The Flux-Anomaly-Forced Model Intercomparison Project (FAFMIP) contribution to CMIP6: investigation of sea-level and ocean climate change in response to CO_2 forcing, Geosci. Model Dev., 9, 3993–4017, https://doi.org/10.5194/gmd-9-3993-2016, 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., von Hardenberg, J., Hazeleger, W., Kodama, C., Koenigk, T., Leung, L. R., Lu, J., Luo, J.-J., Mao, J., Mizielinski, M. S., Mizuta, R., Nobre, P., Satoh, M., Scoccimarro, E., Semmler, T., Small, J., and von Storch, J.-S.: High Resolution Model Intercomparison Project (HighResMIP v1.0) for CMIP6, Geosci. Model Dev., 9, 4185–4208, https://doi.org/10.5194/gmd-9-4185-2016, 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, Mon. Weather Rev., 141, 4165–4172, 2013.

Joint UK Land Environment Simulator: JULES, available at: https://jules-lsm.github.io/, last access: 13 January 2020.

Kravitz, B., Robock, A., Tilmes, S., Boucher, O., English, J. M., Irvine, P. J., Jones, A., Lawrence, M. G., MacCracken, M., Muri, H., Moore, J. C., Niemeier, U., Phipps, S. J., Sillmann, J., Storelvmo, T., Wang, H., and Watanabe, S.: The Geoengineering Model Intercomparison Project Phase 6 (GeoMIP6): simulation design and preliminary results, Geosci. Model Dev., 8, 3379–3392, https://doi.org/10.5194/gmd-8-3379-2015, 2015.

Kuhlbrodt, T., Jones, C. G., Sellar, A., Storkey, D., Blockley, E., Stringer, M., Hill, R., Graham, T., Ridley, J., Blaker, A., Calvert, D., Copsey, D., Ellis, R., Hewitt, H., Hyder, P., Ineson, S., Mulcahy, J., Siahaan, A., and Walton, J.: The Low-Resolution Version of HadGEM3 GC3. 1: Development and Evaluation for Global Climate, J. Adv. Model. Earth Sy., 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, Geophys. Res. Lett., 29, https://doi.org/10.1029/2002GL015415, 2002.

Liu, J., Yuan, X., Rind, D., and Martinson, D. G.: Mechanism study of the ENSO and southern high latitude climate teleconnections, Geophys. Res. Lett., 29, https://doi.org/10.1029/2002GL015143, 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, Geosci. Model Dev. Discuss., 8, 2403–2435, https://doi.org/10.5194/gmdd-8-2403-2015, 2015a.

Liu, L., Peng, S., Zhang, C., Li, R., Wang, B., Sun, C., Liu, Q., Dong, L., Li, L., Shi, Y., He, Y., Zhao, W., and Yang, G.: Importance of bitwise identical reproducibility in earth system mod-

eling and status report, Geosci. Model Dev. Discuss., 8, 4375–4400, https://doi.org/10.5194/gmdd-8-4375-2015, 2015b.

Loeve, M.: Elementary probability theory, in: Probability Theory I, Springer, 1–52, 1977.

Lorenz, E. N.: Deterministic nonperiodic flow, J. Atmos. Sci., 20, 130–141, 1963.

Madec, G. and the NEMO Team: NEMO ocean engine, available at: https://epic.awi.de/id/eprint/39698/1/NEMO_book_v6039.pdf (last access: 13 January 2020), 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., Roberts, C. and Robson, J., Wood, R., and Xavier, P.: Preindustrial Control Simulations With HadGEM3-GC3. 1 for CMIP6, J. Adv. Model. Earth Sy., 10, 3049–3075, 2018.

NEMO Consortium: NEMO, available at: https://www.nemo-ocean.eu/, last access: 13 January 2020.

Otto-Bliesner, 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., Fischer, H., Goelzer, H., Govin, A., Haywood, A., Joos, F., LeGrande, A. N., Lipscomb, W. H., Lohmann, G., Mahowald, N., Nehrbass-Ahles, C., Pausata, F. S. R., Peterschmitt, J.-Y., Phipps, S. J., Renssen, H., and Zhang, Q.: The PMIP4 contribution to CMIP6 – Part 2: Two interglacials, scientific objective and experimental design for Holocene and Last Interglacial simulations, Geosci. Model Dev., 10, 3979–4003, https://doi.org/10.5194/gmd-10-3979-2017, 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, Geophys. Res. Lett., 44, 6200–6208, 2017.

Ridley, J., Menary, M., Kuhlbrodt, T., Andrews, M., and Andrews, T.: MOHC HadGEM3-GC31-LL model output prepared for CMIP6 CMIP, Earth System Grid Federation, https://doi.org/10.22033/ESGF/CMIP6.419, 2018a.

Ridley, J. K., Blockley, E. W., Keen, A. B., Rae, J. G. L., West, A. E., and Schroeder, D.: The sea ice model component of HadGEM3-GC3.1, Geosci. Model Dev., 11, 713–723, https://doi.org/10.5194/gmd-11-713-2018, 2018b.

Song, Z., Qiao, F., Lei, X., and Wang, C.: Influence of parallel computational uncertainty on simulations of the Coupled General Climate Model, Geosci. Model Dev., 5, 313–319, https://doi.org/10.5194/gmd-5-313-2012, 2012.

Turner, J.: The El Niño–Southern Oscillation and Antarctica, Int. J. Climatol., 24, 1–31, 2004.

UK Met Office: Unified Model, available at: http://www.metoffice.gov.uk/research/modelling-systems/unified-model, last access: 13 January 2020.

Walters, D., Boutle, I., Brooks, M., Melvin, T., Stratton, R., Vosper, S., Wells, H., Williams, K., Wood, N., Allen, T., Bushell, A., Copsey, D., Earnshaw, P., Edwards, J., Gross, M., Hardiman, S., Harris, C., Heming, J., Klingaman, N., Levine, R., Manners, J., Martin, G., Milton, S., Mittermaier, M., Morcrette, C., Riddick, T., Roberts, M., Sanchez, C., Selwood, P., Stirling, A., Smith, C., Suri, D., Tennant, W., Vidale, P. L., Wilkinson, J., Willett, M., Woolnough, S., and Xavier, P.: The Met Office Unified Model Global Atmosphere 6.0/6.1 and JULES Global Land 6.0/6.1 configurations, Geosci. Model Dev., 10, 1487–1520, https://doi.org/10.5194/gmd-10-1487-2017, 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, J. 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., Hyder, P., Ineson, S., Johns, T. C., Keen, A. B., Lee, R. W., Megann, A., Milton, S. F., Rae, J. G. L., Roberts, M. J., Scaife, A. A., Schiemann, R., Storkey, D., Thorpe, L., Watterson, I. G., Walters, D. N., West, A., Wood, R. A., Woollings, T., and Xavier, P. K.: The Met Office global coupled model 3.0 and 3.1 (GC3.0 and GC3.1) configurations, J. Adv. Model. Earth Sy., 10, 357–380, 2018.

Wittenberg, A. T.: Are historical records sufficient to constrain ENSO simulations?, Geophys. Res Lett., 36, L12702, https://doi.org/10.1029/2009GL038710, 2009.