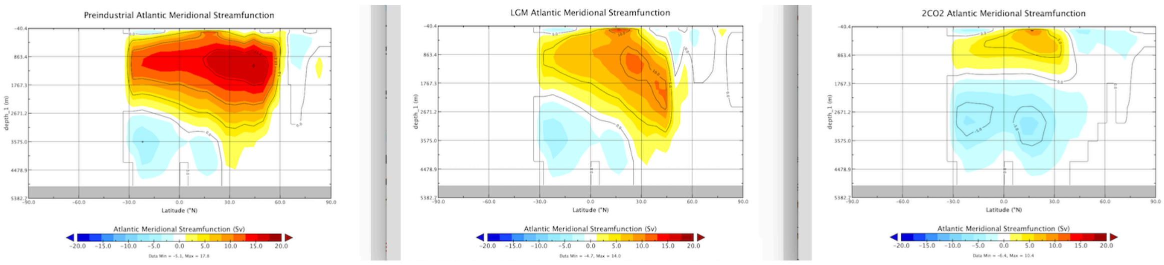
***Interactive comment on* "PALEO-PGEM v1.0: A statistical emulator of Pliocene-Pleistocene climate" *by* Philip B. Holden et al.**
Anonymous Referee #2

We thank the reviewer for this thorough and useful review. Our responses are in bold face and the associated revisions to the manuscript are in italics

First, I apologize to the authors for taking so long to complete this review of "PALEO- PGEM v1.0: A statistical emulator of Pliocene-Pleistocene climate" for publication in GMD. In this paper, the authors describe the building of a statistical emulator from the PLISM-GENIE intermediate complexity model in order to provide a continuous recon- struction of important climate variables over the last 5 millions years, using interpolation to downscale these variables to a higher resolution. This paper logically follows a number of previous articles describing both building of emulators of intermediate complexity models (e.g., Holden et al. 2014) and their application to a range of important scientific questions, such as the links between climate and biodiversity (e.g. Range et al. 2018). It adds in particular a representation of ocean dynamics, which was lacking in previous studies, via the use of PLASIM-GENIE, the coupling of the two models having been described previously (Holden et al. 2018). Overall, the paper is well written and the methodology is sound. I thus recommend publication provided that some important points are taken care of or better discussed:

1. There is no evaluation of the impact of the addition of ocean dynamics on the climate reconstructions, despite being one of the main justifications of the paper (l. 48-49, l. 430). How different would be the results using PLASIM-ENTS rather than PLASIM-GENIE? This leads me to another question: in the PLASIM-GENIE experi- ments with varying ice sheets, is the land/sea mask changed accordingly or are only the height/albedo feedbacks taken into account (with no varying coastlines impacting ocean dynamics)? If the land/sea mask is modified, I wonder how valid are the transition between ocean modes linked to varying sea level in the subsequent emulations of the climate?

   We have added a paragraph to discuss the LGM and 2xCO2 AMOC:
   The climate sensitivity of the optimised parameter set is 3.2°C. The maximum Atlantic overturning is 17.8Sv, at a depth of 1.1km with the 10Sv contour, an indicator of the location of NADW formation, at a latitude of 56°N. Under LGM forcing, Atlantic overturning weakens to a peak of 11.1Sv at a depth of 1.0km and the 10Sv contour shifts southward to 45°N. Under doubled CO_2 forcing, Atlantic overturning weakens substantially to a peak of 7.6Sv at a depth of 0.4km.
   The Atlantic meridional streamfunctions are shown below for your interest - we have not included these plots in the revised manuscript, but would be happy to do so if the reviewers or editor would like.



We cannot meaningfully compare PLASIM-GENIE with PLASIM-ENTS as the two models do not share the same tuned parameter values. However, the results will clearly be affected, for instance by the ocean circulation changes discussed above. While the effects of ocean dynamics on the climate reconstructions could be studied in general, along with an almost unlimited range of other questions regarding the climate dynamics themselves, this is outside the scope of the present paper which is focused on presenting the reconstructions as a tool for interdisciplinary studies.

The land-sea mask is held fixed at the present day. We have clarified this in the text and discussed some of the weaknesses associated with this in the new section 9 (see below).

2. Related to the above, the anomaly method for the downscaling approach supposes that the biases between an observed climate state Ct and the emulated climate Et remain somehow constant through time. This is a very crude assumption given, among others, the documented changes in ocean circulation dynamics across the last 5 Ma. How well is the ocean circulation represented in the model? And ocean circulation changes, e.g. between the LGM and MH?

See above re ocean circulation. We agree that the downscaling assumes the modelled bias remains fixed through time, and have included this in an extended discussion of weaknesses (section 9 reproduced below), which also discusses aspects of ocean circulation, in response to both reviewers.

9 Limitations of the approach

PALEO-PGEM is to our knowledge the first attempt to provide a detailed spatiotemporal description of the climate of the entire Pliocene-Pleistocene period. It is essential to understand the main limitations of our modelling framework, discussed below, some of which may induce large errors or uncertainties in specific applications, or even rule out certain applications completely. For all practical purposes and for the foreseeable future, substantial uncertainties exist in any paleoclimate reconstruction as a result of incomplete knowledge, computing limitations and irreducible climatic noise. Ideally, these uncertainties should be quantified in relation to any reconstruction and their implications propagated through the analysis. Our approach provides an estimate of inherent uncertainty derived from the emulation step of the reconstruction and thus underestimates the full uncertainty, but nevertheless in some aspects remains comparable to the uncertainty in state-of-the-art reconstructions of particular periods as measured by the variance across ensembles of PMIP simulations.

Compared to state-of-the-art models, PLASIM-GENIE is a relatively low resolution, intermediate complexity climate model. This implies that processes operating at spatial and temporal scales below the native resolution of the climate model cannot be properly represented, although certain aspects of spatial variation are reintroduced in a highly idealised way by the downscaling process. The temporal effects of dynamical processes operating at sub-millenial timescales are further filtered out by the approximation inherent in the emulator construction that the climate is in quasi-equilibrium with the forcing, which is then only resolved at 1000-year time intervals.

In applications where (downscaled) time-slice simulations are adequate and are available from higher complexity models and/or multi-model ensembles (Section 7), these would normally be preferable to PALEO-PGEM as errors and biases will generally be smaller, particularly in high latitudes, regions of steep topography, close to coastlines or in known regions of locally extreme climate. We note that HadCM3 climate simulations (Singarayer et al 2017), downscaled to 1° resolution are available back to 120 kaBP (Saupe et al 2019), which would provide preferable (or supplementary) climate data for applications restricted to this time-domain.

The emulator uncertainty captures much of the uncertainty seen in multi-model intercomparisons (Figures 3 and 4), but PALEO-PGEM cannot fully represent model uncertainty, because it is derived from a single configuration of a single model. Most clearly in this respect, the 90% uncertainty range of climate sensitivity (3.8 ± 0.6°C) is understated relative to multi-model estimates of 3.2 ± 1.3°C (Flato et al 2013). Some significant biases in spatial patterns are also apparent, most clearly temperature biases in high southern latitudes.

Emulator forcing is limited to orbit, CO_2 and ice sheets. Ice meltwater forcing is not considered so that millennial variability, especially important in North Atlantic, is neglected. The land-sea mask and orography are held fixed, so that ocean circulation changes driven by changing gateways (e.g. the closing Panama isthmus, with implications for the thermohaline circulation) are neglected and feedbacks driven by changing orography are neglected, especially important in regions of rapid tectonic uplift.

The representation of ice sheets applies Peltier 5G deglaciation ice sheets (Peltier 2004), assuming a fixed relationship between global sealevel reconstructions (derived from benthic oxygen isotopes) and the spatial form and extent of ice sheets. This approximation neglects the substantial asymmetry between build-up and decay phases of ice sheets and assumes that ice sheets were located similarly in all previous Pliocene-Pleistocene glaciations, which may not have been the case. Particular caution is therefore essential when applying the climate reconstruction at locations near to the margins of ice sheets.

We apply a downscaling approach because spatial climate gradients can be critically important for ecosystem dynamics, especially in mountainous regions which are poorly resolved at native climate model resolution (Rangel et al 2018). The downscaling approximation assumes that the lapse rate within a downscaled grid cell does not change with time, but it does capture the first order effect of topographic complexity by assuming a *constant present-day lapse rate. Similarly, the downscaling cannot capture feedbacks between atmospheric circulation and high resolution topography, which could alter the patterns of rain shadowing. However, for many applications, it is preferable to neglect this second order feedback than to neglect the first order effect of a rain shadow that could not be resolved at native climate model resolution (e.g. the Atacama), which downscaling imposes through the baseline climatology. Other simplifications include the implicit assumptions of fixed mountain glaciers and ecotone distributions. In short, the high-resolution reconstructions should not be interpreted as a faithful reconstruction of high-resolution climate, but serve to introduce a more realistic degree of spatial variability.*

3. I am a bit disappointed by the lack of actual comparison to data over the last 5 Ma. For instance, how does it represent the mid-Pliocene period (and compare to the extensive database of PlioMiP exercises) or glacial-interglacial T and P variations? And transient events like MIS M2?

We have added comparisons on glacial-interglacial variability (c.f. Koehler et al 2010), last interglacial transients (cf. Bakker et al 2013) and the Mid-Pliocene (c.f. Haywood et al 2013).

7.3 Glacial-interglacial variability

The emulated global temperature change over the last 800,000 years is plotted in Figure 5, reflecting the familiar glacial cycles and compared to the observationally based global temperature reconstruction of Koehler et al (2010). Ten separate emulators were built (following the steps described in Section 5 applied to annual average temperature) and the mean prediction time-series for all ten emulators are plotted.

The Last Glacial Maximum cooling across these ten emulators is 4.3 ± 0.3°C, which compares to 4.5 ± 0.3°C when emulated values were drawn randomly from a single emulator. The emulated estimates are lower than the simulated LGM cooling of 5.9°C (Table 1) and may reflect bias in the ice-sheet emulator under the extreme of LGM forcing; the ice-sheet emulator was only able to explain 81% of the variance of cold season temperatures (Table 3). However, the seasonal patterns of emulated change are reasonable (Figure 4) and the annual average cooling is well-centered on the 3.1 to 5.9°C range simulated by the CMIP5/PMIP3 and PMIP2 ensembles (Masson Delmotte et al 2013).

Maximum warming of 0.3 ± 0.1°C is emulated in the Last Interglacial (Marine Isotope Stage 5), peaking at 125 kaBP. This is consistent with CMIP estimates of 0.0 ± 0.5°C, but lower than data-based estimates of ~1 to 2°C (Masson Delmotte et al 2013). Maximum warming in Marine Isotope Stage 11 is 0.1 ± 0.2°C, peaking at 401 kaBP.

7.4 Last Interglacial transients

Zonally-averaged emulated temperature changes are compared with the Last Interglacial transient model inter-comparison of Bakker et al (2013) in Figure 5 and Table 5. The latitudinal temporal trends are well captured by the emulator, considering the inter-model *spread of Bakker et al (2013). Notably, temperatures in Jun-Jul-Aug generally peak earlier (~125 kaBP) than temperatures in Dec-Jan-Feb (~120 kaBP). Maximum warming of ~2 to 3°C is emulated in northern summer mid-high latitudes, peaking at 126kaBP, and consistent with inter-model estimates in the range 0.3 to 5.3°C, peaking between 125 and 128kaBP. Eight of the emulated peak warming estimates are consistent within the 1σ multi-model uncertainty ranges, and the remaining two are consistent within 2σ multi-model uncertainty (Table 5). The clearest difference is seen in Antarctic summer, where cooling of up to 5°C is emulated, significantly greater than in any of the models.*

7.4 The Mid-Pliocene warm period

The emulated climate of the Mid-Pliocene warm period is plotted in Figure 6. The only emulator forcing is CO_2 increased to 405ppm, as assumed in the model inter-comparison of Haywood et al (2013). Ice-sheets are fixed at present day, in contrast to Haywood et al (2013) where the boundary conditions included a reduced West Antarctic Ice Sheet.

Ensemble-averaged emulated warming is 1.6 ± 0.2°C and global precipitation change 0.10 ± 0.01 mm/day. These compare to multi-model estimates of 1.8 to 3.6°C and precipitation changes of 0.09 to 0.18 mm/day in Experiment 2 (the coupled atmosphere-ocean configuration) of Haywood et al (2013). Emulated high latitude warming of ~4°C is low-biased, but within the wide multi-model uncertainty range of ~3 to 14°C. Similarly, the emulated peak precipitation change of ~0.3 mm/day near the Equator is low biased, but within the multi-model range of ~0 to 1.3mm/day.

	60°N-90°N	30°N-60°N	30°S-30°N	60°S-30°S	90°S-60°S
DJF peak warming °C	1.7 (-5.8 to 1.2)	0.1 (-0.8 to 2.1)	0.4 (0.6 to 1.2)	0.4 (-0.7 to 1.0)	-0.3 (-1.3 to 2.3)
DJF year of peak warming BP	124 (118 to 124)	119 (117 to 121)	119 (116 to 119)	119 (119 to 121)	119 (116 to 118)
JJA peak warming °C	2.4 (0.3 to 3.7)	3.0 (0.7 to 5.3)	0.7 (0.3 to 2.5)	0.2 (-0.7 to 1.0)	-0.3 (-1.3 to 2.3)
JJA year of peak warming BP	126 (125 to 128)	126 (126 to 129)	125 (127 to 130)	118 (124 to 130)	119 (126 to 129)

Table 5. Last Interglacial peak warming (°C) and year of peak warming (BP) compared to the model inter-comparison $\pm 1\sigma$ ranges of Bakker et al (2013). Emulated data are provided for Dec-Fan-Feb and Jun-Jul-August, compared to January and July data in the model inter-comparison, and comparisons are provided for five latitude bands.

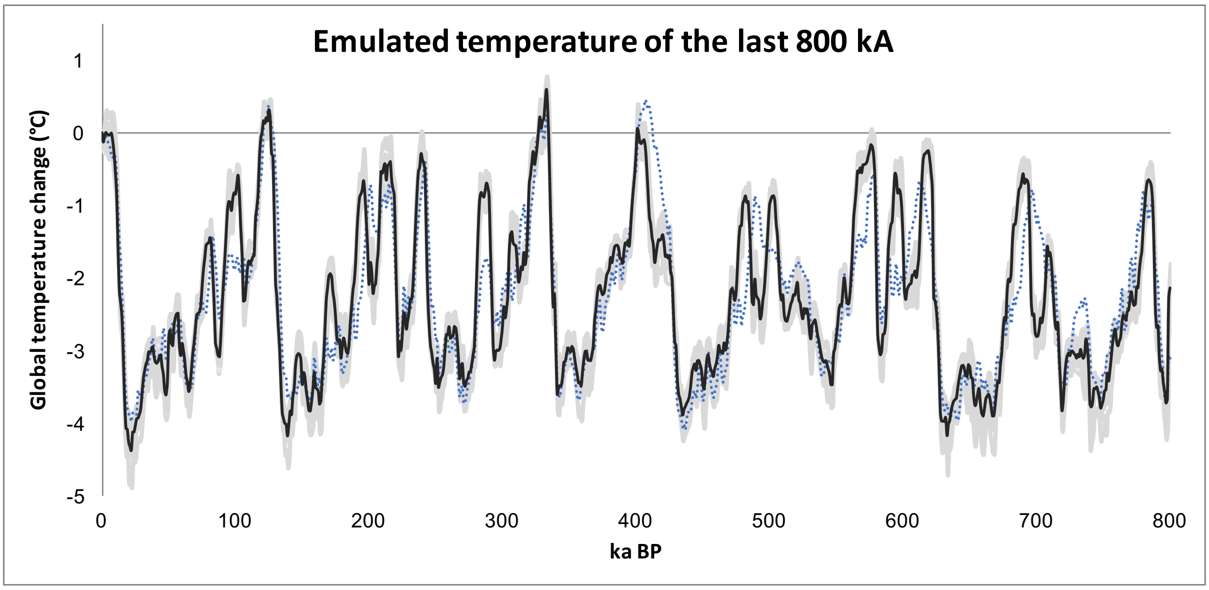


Figure 5: Emulated global temperature over the last 800,000 years. An emulator was built ten times and the mean prediction time series of each emulator are plotted as grey lines, with the mean of these plotted as the single black line. The blue dotted line is the observationally based reconstruction of Koehler et al (2010). "Inter-emulator" variability compares to emulated LGM ensemble cooling (i.e. when drawing principal components scores randomly from a single emulator) of 4.5 ± 0.3 °C.

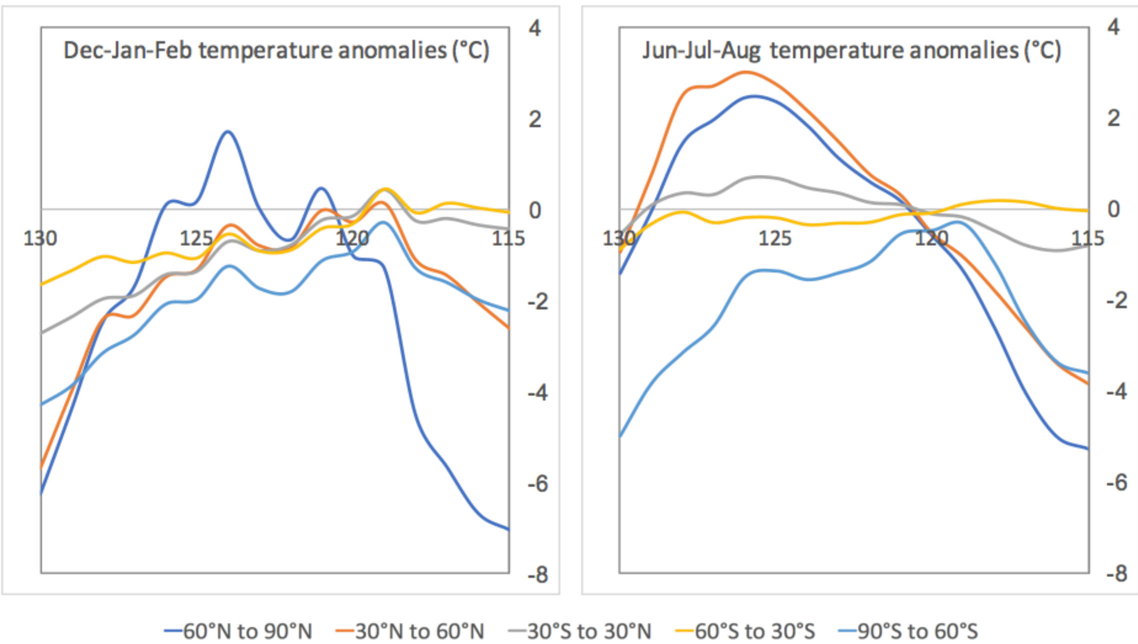


Figure 6: Emulated Last Interglacial temperature anomalies with respect to pre-industrial. Data are provided for Dec-Jan-Feb and Jun-Jul-Aug averaged over five latitude bands c.f. Figures 2 and 3 of Bakker et al (2013).

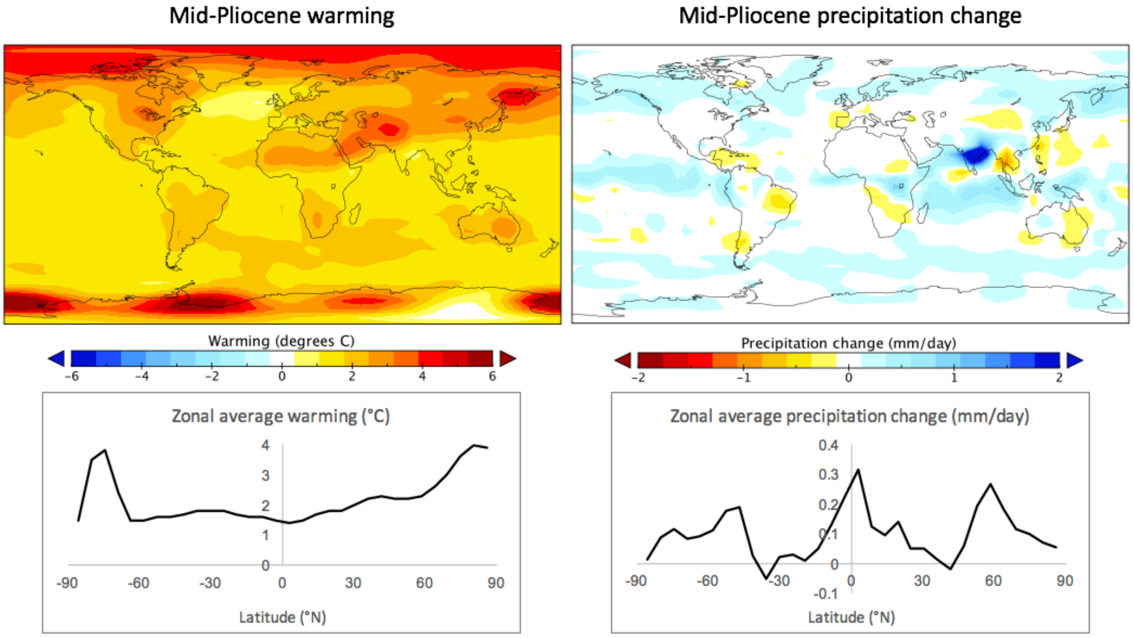


Figure 6: Emulated Mid-Pliocene temperature and precipitation anomalies with respect to pre-industrial. The ice-sheet and orbital inputs are set to preindustrial, and the emulated change is driven by an assumed CO_2 concentration of 405ppm

4. I think a concise but very general reminder of what is an emulator and of the theoretical basis behind could be useful for non specialized readers. In the present version, this reduces to one sentence (l. 58-59) because the Sections 3 and 4 either use too many reference to previous papers (Section 3) or are probably too specialized already (Section 4).

We have extended this as below:
Our methodology uses principal component analysis to project spatial fields of model output onto a lower dimensional space of the dominant simulated patterns of change and then derives regression relationships between the simulator inputs and the coefficients of the dominant patterns. The method is analogous to the widely-used pattern-scaling technique (Tebaldi and Arblaster 2014), which assumes that an invariant pattern of simulated change can be scaled by global warming. Our approach extends this by including several (here ten) principal components for each climate variable, thereby allowing us to capture nonlinear patterns of change. The regression approach we use involves Gaussian process (GP) emulation (Rasmussen 2004).

5. The discussion on the limits of the approach and of the (numerous) uncertainties and approximations made should be expanded, in particular if the aim is to provide a widely available climatic reconstruction that is then used to force ecological niche modelling or biodiversity models (l. 36-42). Because this requires at least some confidence in the ability of the model to reproduce “true” (absolute) paleoclimatic conditions and variability (see points 1-3 above).

See new section 9 (above)

Other minor details l. 52 “naïve simulation would not be possible for an application of this ambition”. I do not understand this sentence.

We have revised this as below (full paragraph included for context)

However, simulation alone would not be possible for an application of this ambition. We use the computationally-fast low-resolution AOGCM PLASIM-GENIE (Holden et al 2016), but even with this relatively simple model a five million-year transient simulation would demand ~300 CPU years of computing, which could not readily be parallelised. We overcome this intractability by using statistical emulation.

l. 272-274. “GPs are highly flexible non-parametric regression models which have greater modelling power than linear models”. Please clarify.

Clarified with
Linear models live in a finite dimensional space defined by polynomial functions of the covariates. Gaussian processes live in a much richer space of functions.

l. 393. “Warm biases are more modest”. On the basis of the figure, this is hard to believe as there are regions with a warm bias as large as the cold bias of other regions.

Clarified with:
Warm biases are more modest except for the Tibetan Plateau and Andes where the lapse rate cooling in these narrow mountain chains is poorly resolved by the climate model (but corrected for by the downscaling approach described below).

Fig. 3 and 4. Please use the same color scale and range as the PMIP ensemble to ease comparisons. Why is the Southern Ocean cooling rather than warming in the MH simulation?

We have removed the 2-component analysis to simplify the presentation, and have increased font sizes for improved clarity. Unfortunately we were unable to access the PMIP2 netcdf files, so could not use the same colour scales. In general we use the same ranges in both plots, but in some cases we prefer to leave these different for adequate contrast in the legend scale. Most notably the SD of LGM temperature is only 5 degrees in PALEO-PGEM, compared to 20 degrees in PMIP2, reflecting the fact that we do not sample uncertain climate sensitivity. We have added a note in the caption to emphasise this.

Re emulated Southern Ocean cooling in the MH, we have added the text.

The most significant difference is Antarctic cooling of ~3°C in PALEO-PGEM, which contrasts with a warming signal in the ensemble mean of PMIP2 (although we note DJF Antarctic cooling of 0.5°C was simulated in HadCM3M2). A significant cold Antarctic bias is also apparent during the Last Interglacial (Section 7.4). High southern latitudes are poorly modelled by PLASIM-GENIE. The preindustrial state exhibits a warm Antarctic bias, with greatly understated sea ice, a slow Antarctic Circumpolar Current and weak, northerly shifted zonal winds (Holden et al 2016), which are likely to be associated with well-known *difficulties of resolving Southern Ocean wind stress at low meridional resolution (Tibaldi et al 1990, Schmittner et al 2010).*