The relative role of orbital, CO_2 and ice sheet forcing on Pleistocene climate

C. J. R. Williams*[1], N. S. Lord[2], D. J. Lunt[1], A. T. Kennedy-Asser[1], D. A. Richards[1], M. Crucifix[3], A. Kontula[4], M. Thorne[5], P. J. Valdes[1], G. L. Foster[6] and E. L. McClymont[7]

[1]School of Geographical Sciences, University of Bristol, Bristol, UK.
[2]Fathom, Bristol, UK.
[3]Earth and Life Institute, Georges Lemaître Centre for Earth and Climate Research, Université Catholique de Louvain, Louvain-la-Neuve, Belgium.
[4]Posiva Oy, Eurajoki, Finland.
[5]Mike Thorne and Associates Limited, Quarry Cottage, Hamsterley, Bishop Auckland, Co. Durham, UK.
[6]National Oceanography Centre, University of Southampton, Southampton, UK
[7]Department of Geography, Durham University, Durham, UK

* Corresponding author: Charles JR Williams (c.j.r.williams@bristol.ac.uk)

ABSTRACT

During the Pleistocene (the last ~2.5 million years), Earth's climate has fluctuated between a series of glacials and interglacials, driven by external forcings and mediated by internal feedbacks. Climate models provide a useful tool for addressing the mechanisms associated with these variations; however, the complexity of such models means they require substantial computational resources, meaning they are not suitable for exploring transient orbital-scale variability. Instead, we use a climate model to calibrate a faster statistical model, or emulator, and apply this to the Pleistocene.

The emulator is several orders of magnitude faster than a climate model. The results suggest that the emulator performs well, and agrees with the proxy records over the last 800,000 years, especially concerning the timing of glacial-interglacial cycles. The results also suggest that a combination of the CO_2 and ice components used to drive the emulator provides the largest contribution to the overall signal. The efficiency of this approach allows us to carry out a quasi-transient simulation through the entire Pleistocene, and allows projections of possible future drilling results from deep Antarctic ice cores.

INTRODUCTION

For the last few million years (Myr), the Earth's climate has fluctuated between cooler (glacial) and warmer (interglacial) intervals, as revealed by proxy records of palaeoclimate (e.g. Herbert *et al*. 2010, Jouzel *et al*. 2007, Wang *et al*. 2008). The timing and rate of these glacial-interglacial transitions is ultimately forced by variations in the Earth's orbit around the sun, the three main parameters of which are precession, obliquity and eccentricity (which have periods of ~23 thousand years (kyr), ~41 kyr and both ~96 and 400 kyr, respectively; Berger 1978, Hays *et al*. 1976, Kawamura *et al*. 2007, Lisiecki and Raymo 2007, Milankovitch 1941). In addition to the orbital forcing, key internal feedbacks in the climate system mediate the forcing, including changes in the concentration of atmospheric CO_2 and changes in the extent and thickness of global ice sheets. For the Pleistocene (2.58 million years ago (Ma) to present day), glacial-interglacial fluctuations are evident in a relatively large number of palaeoclimate records, constructed from various climate proxies (see 'Proxy data' in Methods for more details). However, although this period is now well observed in terms of proxies, our understanding of the processes that ultimately drive climate change, and the interplay between forcings and feedbacks, is not well known. As such, there are many important questions that remain unanswered; for example, what proportion of the signals of Pleistocene climate change is driven directly by orbital forcing and how much is associated with CO_2 and/or ice sheet feedbacks, and how does this change through the Pleistocene?

In order to project changes in past climate that are physically-based and spatially resolved, climate models (or more specifically fully-coupled general circulation models, GCMs, or Earth System Models, ESMs) are required, providing a useful tool for addressing such questions by allowing investigation of the driving mechanisms, dynamics, feedbacks, and sensitivity of the climate system associated with variations in orbital forcing, CO_2, and ice sheets (e.g. Abe-Ouchi *et al*. 2013, Erb *et al*. 2015). However, the structural complexity of these models, along with their relatively high spatial and temporal resolution, means that they require substantial computational resources and time to run, and therefore often cannot be run in fully-coupled transient mode over the long timescales (> 100 kyr) over which these forcings and feedbacks vary. Some studies have applied GCMs to conduct transient simulations of past climatic changes, but this has generally been in the form of a number of snapshot simulations (the assumption being that the system is never far from equilibrium with the orbital forcings) in order to 'build' a transient simulation (e.g. Hoogakker *et al*. 2016, Marzocchi *et al*. 2015, Prescott *et al*. 2014); however, even these simulations are typically limited to a single cycle. Alternatively, a limited number of transient GCM simulations have been conducted, such as in the framework of the Deglaciation PMIP experiment (Ivanovic *et al*. 2016) or the PalMod project (PalMod 2023), but these in general simulate at most a single glacial-interglacial cycle. Very few fully transient (i.e. covering many glacial-interglacial cycles) GCM simulations exist; a recent example is that of Timmermann *et al*. (2022), who ran the computationally efficient Community Earth System Model (CESM1.2) for the last 2 Myr, to investigate the climatic effects on five hominin species throughout the Pleistocene. Apart from this, simulations of multiple glacial-intergalcial cycles with a full GCM or ESM are generally not practical, and the number of sensitivity studies that can be performed is limited. An alternative is to use models with a lower complexity, such as Earth system Models of Intermediate Complexity (EMICs), to simulate long-term transient past climate changes (e.g. Loutre and Berger 2000a, Stap *et al*. 2014).

However, EMICs are problematic, because their usefulness is compromised by the need for increased parameterization of physical processes and decreased spatial resolution.

Therefore, to resolve this competing desire for maximum process complexity and sufficient spatial resolution versus practical computing resource requirements and running time, statistical methods are increasingly being developed and applied to simulate long-term changes in climate. One emerging approach is that of emulators, which are statistical models that are calibrated on data from a more complex climate model, such as a GCM. They give a projection of the climate resulting from a certain set of input conditions (climate drivers), along with an estimation of the uncertainty associated with the projection. However, whilst the climate projection is based on the physics of the GCM and is at the same spatial resolution, the requirements of the emulator in terms of computational resources and time are a fraction of those required for the full GCM. This makes them useful for investigating processes and climatic changes occurring over long time scales of hundreds of kyr or longer, thus allowing for comparison to proxy climate data and model validation (e.g. Holden *et al*. 2019, Lord *et al*. 2017). Emulators have been applied in sensitivity analyses of climate to orbital, atmospheric CO_2, and ice sheet configurations (e.g. Araya-Melo *et al*. 2015, Bounceur *et al*. 2015) and to investigate parametric uncertainty in models (e.g. Holden *et al*. 2010, Johnson *et al*. 2015).

Here, we apply a climate emulator to simulate and understand the evolution of long-term climate change in the Pleistocene, compared with proxy records. We expand on the methodology described in Lord *et al*. (2017), which explored drivers of Pliocene climate associated with atmospheric CO_2 concentration and the three main orbital parameters of longitude of perihelion, obliquity and eccentricity. Here, we additionally explore global ice volume (using global sea level (GSL) as a proxy for this), which is critical to explain glacial-interglacial cycles of the Pleistocene (see 'The emulator' in Methods for more details). The emulator is used to simulate global surface air temperature (SAT) and precipitation, and the regional results can be compared to various proxy records (see 'Proxy data' in Methods for more details). Simulations are then performed to explore the sensitivity of climate at different locations to the different drivers (see 'Experimental design' in Methods for more details). Apart from the updates applied here (see 'Updates to the emulator' in Methods for more details, as well as additional text in the

Supplementary Material), the emulator was run in the same way as in Lord *et al*. (2017), Lord *et al*. (2019) and Williams *et al*. (2022). As such, we are able to explore Pleistocene glacial-interglacial cycles in terms of their relative influence from orbital forcing (obliquity, precession, and eccentricity separately), and CO_2, and ice sheet feedbacks.

RESULTS

'All drivers' simulation (E_{11111}): model-data comparison

Figure 1 compares the proxy and emulated temperature data over the last 2.58 Myr at the six sites (for ease of comparison, the same figure is shown in the Supplementary Material, Figure S5, but with uniform y-axes), and when all forcings and feedbacks are included in the emulator (simulation E_{111111}; see 'Nomenclature and factorisation' in Methods for more details). It should be noted that proxy data between 800 kyr BP and 2.58 Myr are only shown for three of these sites (where the temporal resolution of the data is < 5 kyr, mostly 1-2 kyr), because at the other three sites the temporal resolution of existing proxy data during this period was deemed to be too large (> 10 kyr), or in the case of Dome C, proxy data did not exist.

It is interesting to briefly note here the speed of the emulator in producing this 2.58 Myr simulation. Given the temporal resolution is 1 kyr, the emulator took approximately five minutes to run over the 2580 timeslices, calculate comparisons with the proxy data and produce the results. Likewise, concerning the sensitivity simulations (see below), each of which covered the last 800 kyr (i.e. 800 timeslices), the emulator took approximately two minutes to do each simulation, giving a total runtime of just over one hour to complete 32 simulations. As a comparison, for a GCM such as HadCM3 to do the same process (i.e. produce a quasi-transient simulation of 28,180 timeslices in total, comprising one run of 2580 timeslices and then 32 repetitions of 800 timeslices) it would take approximately 280 days, at best (that is, assuming an optimal performance of 100 model years day^{-1}; the reality is often much less). Moreover, the emulator can be run, at this speed, on a standard desktop PC, whereas HadCM3 requires resource-intensive supercomputing infrastructure. This demonstrates the efficacy of the emulator.

Focusing generally on all locations during the most recent 800 kyr, the results suggest that the emulator does a good job of reproducing temperature up variations at these sites. In particular, the timing and duration of many of the interglacials and glacial maxima are similar in the model and proxy reconstructions, and the agreement between the magnitude of variations is generally good (Fig. 1). However, at many of the locations there are some instances when the relative maxima and minima are not well reproduced by the emulator, such as the interglacials at approximately 330 kyr, 240 kyr, and 125 kyr before present (BP), corresponding to Marine Isotope Stage (MIS) 9, 7 and 5, respectively. At times, the emulator demonstrates stronger glacial cooling than the proxy data at a number of locations including at the Last Glacial Maximum (LGM, ~21 kyr BP) and during MIS 6 (~200-130 kyr BP) and MIS 12 (~450 kyr BP).

Focusing on individual sites, at Dome C (Fig. 1a) there is a relatively large amount of SAT variability in the interglacial-glacial cycles, which continues until approximately 1.5 Ma and then decreases towards the early Pleistocene (in line with the other Ocean Drilling Program (ODP) sites). Before the period of relatively good agreement between reconstructed and emulated temperature (i.e. before the last 800 kyr), the emulator appears to be underestimating temperature by as much as 6°C, for example at ODP 982 and 846 (Fig. 1b and e). ODP 722 shows a similar pattern, but with smaller anomalies in both datasets and differences of ~4°C (Fig. 1c). All three of these sites show relatively variable negative SAT emulated anomalies, in contrast to the generally positive sea surface temperature (SST) anomalies from the reconstructions. Moreover, towards the beginning of the Pleistocene (i.e. 2.58-1.5 Ma) the emulator is showing considerably less variability, relative to the last 800 kya, than the reconstructions, especially at the ODP sites (Fig. 1b-f); possible reasons for this are discussed below. At ODP sites 1143 and 846 the emulated and proxy records appear to show less agreement prior to around 500 kyr BP (Fig. 1d and b, respectively). Likewise the SST record from ODP site 1090 (Fig. 1e) shows large temperature fluctuations of ±7°C, significantly larger than the variability predicted by the emulator (which is approximately in line with emulated SST variability at other ODP sites).

At Dome C in particular (Fig. 1a), where proxy data are currently limited by existing drilling methods, the emulator is able to give an indication of what future temperature reconstructions throughout the Pleistocene might resemble there. This is in spite of the uncertainties in the forcing components of the emulator, which increase during this early period. For example, the emulator suggests a series of interglacials at ~950 kyr, ~1100 kyr, 1250 kyr and 2350 kyr, in between a series of glacials which are less severe before ~1500 kyr. If drilling methods can ever be extended, therefore, this gives an indication of what the possible reconstructions might resemble.

Concerning precipitation, the $\delta^{18}O$ data records from China (used as a proxy for precipitation) are compared with emulated precipitation in Figure 2. As with temperature, the emulator appears to be able to reproduce the major variations evident in the $\delta^{18}O$ record reasonably well, particularly with regards to the precession-driven timing of maxima and minima in precipitation. Unlike SAT and SST, however, a similar level of variability continues when the emulator is run further back beyond the proxy data records, and this is true at the three sites considered (Fig. 2). Whilst there do appear to be similarities between the emulated and $\delta^{18}O$ results, direct quantitative comparison between these records is challenging without including an explicit representation of oxygen isotopes in the underlying GCM simulations.

However, there are discrepancies between the reconstructed and emulated temperature and precipitation, particularly towards the beginning of the Pleistocene. The considerably smaller variability from the emulator during this early period, relative to the last 800 kyr, is because the GSL and CO_2 used to drive the emulator are also much less variable during this early period (see the Supplementary Material, Fig. S3). At certain sites, such as in the subantarctic South Atlantic (ODP 1090, Fig. 1f), the proxy record is showing much greater variability than the emulated temperature, which could be the result of a number of factors but is likely because this region is strongly affected by the Benguela current, which is not well simulated in low resolution ocean models, such as the ones used here to calibrate the emulator. The relatively large uncertainty in the emulator at many of the sites at ~400 kyr BP, such as in the North Atlantic (ODP 982, Fig. 1b), is likely due to the fact that, contrary to the transition from interglacial to glacial conditions for which changes in the ice sheets are well sampled, there are only two different Antarctic and Greenland ice sheet configurations for interglacial conditions; modern-day ice (*modice*) and reduced ice (*lowice*). Therefore, the emulator uncertainty in regions where the ice sheets change significantly (whilst in an interglacial state) is relatively high, because the emulator effectively only has two ice sheet configurations to interpolate between. This is not such an issue for most of the interglacials, as many of them have GSL values that are lower than PI, but at MIS 11 GSL is particularly high, suggesting increased melting of one or more of the global ice sheets.

It is important to quantify the model-proxy comparison; to this end, we calculate an Arcsin Mielke measure (M score; see ‘Arcsin Mielke measure’ in Methods for more details) between the emulated and proxy data, for the ‘all drivers’ simulation (E_{11111}) and considering both the 800 kyr and 2.58 Myr time periods; see Table 1. It is important to put these M scores into context. A baseline is to calculate the M score for a constant climate signal with the same mean as the proxy record. As expected, at all sites the modelled variations perform substantially better than a constant signal (see Supplementary Material, Table S2). As a further contextualisation, we calculate an M score between the model and proxy at every possible combination of sites, as shown in Table 1. We assess agreement as ‘good’ for a particular site if the best M score is obtained when the model results at that site are compared with proxies from the same site. Athough some sites (Dome C and ODP 1143) show ‘good’ agreement, the other sites do not. However, this can be explained by the geographical proximity of some of the sites, with the two high latitude sites (ODP 982 and ODP 1090) showing the best match with each other and the two low latitude sites (ODP 722 and 846) showing the best match with each other. As a further test, the above exercise was repeated, but considering only the last 21 kyr (i.e. since the LGM), and similar results were found (not shown).

Single driver experiments: linear factorisation

Following the model-data comparison, the relative impacts of the individual drivers on emulated SAT were investigated. A linear factorisation (see Lunt *et al*. 2021) was initially conducted, in which each individual driver was added in turn, to transition from E_{00000} to E_{11111} (see ‘Nomenclature and factorisation’ in Methods for more details). This is illustrated in Figure 3. Note that this is only one possible linear factorisation, or ‘pathway’; this particular pathway is E_{00000}, E_{10000}, E_{11000}, E_{11100}, E_{11110}, E_{11111} (see ‘Nomenclature and factorisation’ in Methods for more details).

Focusing generally on all locations, the results suggest that changes in CO_2 and, in particular, ice volume (using GSL as a proxy) generally result in significantly larger temperature variations, of several degrees or more, than those driven by orbital changes alone. This point is made more clearly, and indeed globally, by Figure 4, where a different single pathway is considered but instead of consecutively adding in each forcing component (i.e. linear factorisation), here one component only is removed (i.e. held constant) in each simulation. Hence this becomes E_{11111} minus E_{01111}, E_{11111} minus E_{10111}, etc, always keeping four of the components as varying and then incrementally removing each one. Here, it is clear that the impacts of CO_2 (Fig. 4a) and ice (Fig. 4e) are having a much greater contribution to the overall signal than any of the individual orbital components, with the latter (Fig. 4b-d) contributing less than half of the overall signal across much of the world. In certain places CO_2 and ice are contributing up to all of the overall signal, such as across the tropics for CO_2 (Fig. 4a). When compared to the proxy data, it is only when the CO_2 and ice components are added into the emulator does the temperature match the magnitudes suggested by the reconstructions, qualitatively suggesting that ice and CO_2 are providing the majority contribution to the 'all drivers' simulation (Fig. 3). Secondly, focusing on just CO_2 and ice, the individual contributions are behaving in line with current understanding, such as the ice component having the maximum contribution (accounting for almost all of the overall signal) over the high latitudes in the Northern Hemisphere (Fig. 4e). Lastly, and in line with the maps of M scores, when averaged globally CO_2 and ice are contributing over 70% and 50% to the total, respectively, with the contribution from the orbital parameters being much lower (Fig. 4f).

Single driver experiments: linear-sum/shared-interaction factorisation

To overcome the disadvantages of linear factorisation (see 'Nomenclature and factorisation' in Methods for more details), a linear-sum/shared-interaction factorisation was also conducted, again globally. This is shown in Figure 5, where each driving component is averaged over all 120 pathways (the global mean impact of each driving component, for every individual pathway, is shown in the Supplementary Material, Fig. S7). The most important observation is that Figure 5 is giving a very similar story to Figure 4. Firstly, the impact of the orbital parameters (i.e. obliquity, eccentricity and precession) on the M scores is negligible and they are thus contributing little to the overall signal (Fig. 5b-d). Secondly, the results show that CO_2 and ice are contributing the most to the overall signal, particularly over the tropical landmasses (for CO_2, Fig. 5a) and polar regions (for ice, Fig. 5e). When averaged globally CO_2 is contributing 39% to the overall signal and ice is contributing 56%, whereas the orbital parameters' contributions are all < 5% (Fig. 5f).

The M scores between the emulated and proxy data at the same site, from all 32 simulations, are shown in the Supplementary Material (Table S3); the same data are graphically illustrated in Figure S6. Two points are noteworthy. Firstly, across all simulations Dome C and ODP 1143 are giving the best match (i.e. the largest M score) between emulated temperature and the proxy data, with a mean M score of 330 and 211 respectively, compared to mean M scores < 50 for the other sites. Secondly, only at Dome C is E_{11111} showing the best match with the proxy data (M = 539); at the other sites, the simulations showing the best match are E_{10111} (ODP 982, M = 39), E_{00011} (ODP 722, M = 67), E_{10011} (ODP 1143, M = 445), E_{11001} (ODP 846 and 1090, M = 76 and 9.2, respectively). What is consistent among these, however, is the fact that all of these include varying ice, and all but one of these include both varying ice and CO_2, again agreeing with the above observation that these driving components are dominant.

DISCUSSION

The experiments can be compared to the results of Erb *et al*. (2015), for which idealized simulations were carried out using the Geophysical Fluid Dynamics Laboratory (GFDL) Climate Model 2.1 (CM2.1) fully coupled GCM, which was forced by the same forcings as applied here. The climate response 'fingerprints' for individual forcings were then linearly combined to produce projections of SAT for the last ~400 kyr, and the results compared to proxy climate data from different locations. Two comparisons can be made. Firstly, they find a reasonable fit between the simulated and reconstructed data when considered in general terms (Erb *et al*. 2015), consistent with the results presented here. More specifically, although most of their exact locations differ from those presented here, their locations in the eastern and western equatorial Pacific are relatively close to our ODP sites 846 and 1143, respectively. They find that although the range of simulated temperature variability is reproduced well in the western Pacific, it is underestimated in the eastern Pacific and especially during the simulated interglacials (Erb *et al*. 2015). A similar result is found here, particularly during last 800 kya of the simulation; at ODP

1143 (in the western equatorial Pacific, Fig. 1d) the emulator is mostly capturing the warmth of the various interglacials, but it is often failing to do the same at ODP 846 (in the eastern equatorial Pacific, Fig. 1e). A more direct comparison can be made in the Antarctic, where the model-data comparison carried out by Erb *et al*. (2015) is at Vostok whereas ours is at Dome C. At this location, Erb *et al*. (2015) observe substantial differences between the simulated data and proxy record, which they attribute to obliquity variations and the model not being able to correctly capture the obliquity response. In contrast, the results presented here suggest relatively good agreement at this location, and indeed the best agreement among all the proxy data locations (see Fig. 1a and Table 1), possibly suggesting that the emulator presented here is more successfully capturing these orbital signals. Secondly, Erb *et al*. (2015) conducted a similar forcing breakdown, suggesting that the ice sheets drive greater variation than CO_2, particularly during glacial conditions, consistent with the results presented here. Also at this location, they come to similar conclusions as those presented here, with orbital changes accounting for a limited amount of temperature variability.

There are a number of limitations associated with the methodology presented here. Firstly, discrepancies between the emulator and proxy data could occur due to errors in: i) the emulator, where the response of a climate variable to a forcing has not been fully captured; ii) errors in the underlying GCM, such as changes in atmospheric dust (which, for example, may have an impact on long-term climate but are associated with significant uncertainty and are not accounted for in the GCM); iii) missing processes in the emulator, such as freshwater fluxes (meaning millennial-scale climate change driven by fresh water forcing, such as DO-cycles or Heinrich events, are not directly simulated, although their impact on CO_2 and ice sheets is accounted for in this framework in which CO_2 and ice are prescribed from observations) or non-CO_2 greenhouse gases and aerosols; and iv) uncertainties in the proxy data, such as their calibration or age models (e.g. Lawrence *et al*. 2009, Li *et al*. 2011, Lisiecki and Raymo 2005) or seasonal biases (e.g. Herbert *et al*. 2010). Secondly, the GCM simulations on which the emulator is calibrated use the ICE-5G ice sheet reconstruction of Peltier (2004). However, a number of other reconstructions of the ice sheets also exist and demonstrate some differences to the ICE-5G reconstructions (see 'Updates to the emulator' in Methods for more details). Lastly, the emulator is based on the results of a single GCM, HadCM3, whereas other models may produce different results, due to variations in model structure and the parameterisations used.

SUMMARY AND CONCLUSIONS

To summarise, SAT and precipitation is presented here for the entire Pleistocene (2.58 Myr), produced using a statistical emulator calibrated on a large ensemble of HadCM3 simulations, and forced by orbital, atmospheric CO_2, and ice volume variations. The emulator results at various locations were compared to a number of proxy records for palaeoclimate taken from those same locations, including Antarctic temperature data, SST reconstructed from ocean sediment cores, and $\delta^{18}O$ records measured from cave speleothems in China. An additional set of simulations was performed to test the sensitivity of climate to the individual forcings.

Several conclusions can be made. Firstly, the model-data comparison results suggest a good level of agreement between the reconstructed and emulated temperature and precipitation, particularly during the last 800 kyr of the Pleistocene when the proxy data are most reliable and particularly in terms of the timing and duration of many of the interglacial and glacial periods. The timings and relative minima and maxima in precipitation are also well reproduced by the emulator. This is in spite of there being some discrepancies for both temperature and precipitation, however, such as instances when relative minima and maxima are not in agreement with the proxy data; these are likely to be the result of a combination of errors in the emulator, the underlying GCM, and the proxy data. Considering the entire Pleistocene, the emulator performs reasonably well compared to available proxy data, despite greater uncertainty from both sources. Therefore, the emulator is able to give a projection of what certain proxy data, such as SAT at Dome C in Antarctica, might resemble if deeper drilling occurs. Secondly, sensitivity testing of the proxy data gives some confidence in the emulated results, suggesting that the emulator has at least some regional skill in terms of magnitude and cyclicity of climate change. This, coupled with the general agreement (where proxy records exist) as the emulator is run further back in time, indicates that it is appropriate to use the emulator to explore the relative importance of the various drivers that contribute to Pleistocene change. Lastly, when all 32 simulations are analysed, to determine the driving components contributing the most to the overall signal, all results appear to suggest that the ice and CO_2 components are dominant.

Although, as might be expected, the 'all drivers' simulation does not provide the best match with proxy data at all of the locations, the simulations that do provide the best match all have varying ice (and most have varying CO_2) in common. The only place where the 'all drivers' simulation does provide the best match with proxy data is at Dome C (which also gives the highest M scores compared to the other locations), and this is likely because of the dominant ice forcing component, which would have the clearest signal over this region. When averaged globally, although the three orbital components, either individually or combined, do provide some contribution to the overall signal (1.9%, 2.3% and 0.5% for obliquity, eccentricity and precession, respectively), they are overshadowed by contributions from CO_2 and ice (38.8% and 56.4%, respectively).

METHODS

The emulator

The emulator used here was based on that presented in Lord *et al*. (2017), which is a statistical representation of a more complex model, in this case a GCM. It should be noted that, we refer to the groups of climate model 'experiments' as 'ensembles', we refer to the emulator runs as 'simulations' and we refer directly to 'the GCM' when discussing calibration of the emulator. The underlying principle is that a relatively small number of experiments are carried out initially using the GCM, which sample the multidimensional climate forcing input space (in this case, five dimensions consisting of three orbital dimensions, a CO_2 dimension, and a GSL dimension). The emulator is calibrated on these experiments, with the aim of being able to interpolate the GCM results such that it can provide a prediction of the output that the GCM would produce if it were run using any particular combination of input parameters (climate forcings), along with an estimation of the uncertainty associated with the prediction. The emulator can be validated by comparing results with additional GCM results not included in the calibration, and any set (or sequence of sets) of forcing conditions (orbit, CO_2, GSL), within the range of conditions on which it has been calibrated, can then be applied (see Lord *et al*. 2017). For each single combination of values for these climate forcings, the emulator simulates the corresponding state of a particular climate variable, in this case mean annual SAT and mean annual precipitation (MAP). It should not, however, be extrapolated to conditions outside the range on which it has been calibrated.

In this study, as in Lord *et al*. (2017), a principal component analysis (PCA) Gaussian process (GP) emulator is used, based on Sacks *et al*. (1989), with the subsequent Bayesian treatment of Kennedy and O'Hagan (2000) and Oakley and O'Hagan (2002), and a PCA approach associated with Wilkinson (2010). Climate data for the entire global grid is used to calibrate the principal component (PC) emulator, as opposed to calibrating separate emulators based on data for individual grid boxes. This approach is taken because, for past climate, the global response overall is of interest, rather than just the response at specific locations individually. It also means that the results are consistent across all locations.

For a full description of the theoretical basis for the emulator, see Lord *et al*. (2017), which also provides details of its calibration, optimisation, and limitations. To avoid repetition, brief descriptions of these topics is provided in the Supplementary Material, along with an evaluation of the performance of the updated emulator, and its associated limitations. Here, only detailed information about key updates to the emulator are given, along with the experimental design and description of the proxy data.

Updates to the emulator

We updated the emulator described in Lord *et al*. (2017) to allow it to simulate Pleistocene climate by: i) including an additional forcing term of ice sheet volume (using GSL as a proxy for this); and ii) adding additional GCM simulations to the calibration that included various ice sheet configurations. Two separate emulators were used, each of which were calibrated on different GCM simulations; one suitable for glacial conditions (ice sheets greater in volume than the present-day) and one for interglacial conditions (ice sheets equal or less in volume than present-day). For timesteps that had a GSL value of equal to or higher than 0 m (equivalent to present-day interglacial conditions), the interglacial emulator was used, whilst timesteps with a GSL lower than 0 m used the glacial emulator. This approach was taken to ensure that there was no blending of state-specific climate variations across different climate states. For example, under glacial conditions development of the Laurentide ice sheet is accompanied by strong cooling over northern North America. However, this does not mean that during interglacial conditions there should be a strong warming in the same region, since the data are presented as an anomaly from interglacial present-day conditions (where no ice sheet is present). The use of separate emulators therefore ensures that the different glacial states are characterized separately and correctly; checks were carried out to ensure that there were no discontinuities between the climate projections from the different emulators.

As in Lord *et al*. (2017), the emulator for interglacial conditions was calibrated on the *modice* (modern-day ice sheets) and *lowice* (reduced Pliocene ice sheets) GCM simulations described in the previous section and shown in the Supplementary Material (Table S1). Global temperatures during the Pliocene were higher than they are at present, meaning that in the ice sheet reconstruction of the Greenland ice sheet (GrIS) is limited to high elevations in the Eastern Greenland Mountains, and no ice is present over Western Antarctica. Large regions of the East Antarctic ice sheet (EAIS) show minimal changes or slightly increased surface elevation, although there is substantial loss of ice in the Wilkes and Aurora subglacial basins (Haywood *et al*. 2016). Similar patterns of ice retreat have been simulated in response to future warming scenarios for the GrIS (Greve 2000, Huybrechts and de Wolde 1999, Ridley *et al*. 2005, Stone *et al*. 2010) and WAIS (Huybrechts and de Wolde 1999, Winkelmann *et al*. 2015). These two ice sheet configurations therefore represent climate states ranging from interglacial conditions to relatively severe global warming conditions accompanied by significant ice sheet retreat.

New for this study, the emulator was also configured for glacial conditions. The 60-member *modice* ensemble was used, plus the *LGC* ensemble (see Table S1). The reconstructed ice sheet extents are based on the ICE-5G model of Peltier (2004). These reconstructions were selected because they are based on palaeo data (GSL and/or ice-sheet extent) and include the range of associated data that was required for this work. In addition, for glaciations prior to the last glacial cycle, there is very little or no palaeo data available that enables the 3-D reconstructions which are required here. There are, however, other reconstructions (see, for example, Schmidt *et al*. 2014), which could result in changes to the simulated climate. For example, some studies have suggested a lower maximum elevation for the Laurentide ice sheet during glacial conditions (e.g. Abe-Ouchi *et al*. 2015, Kleman *et al*. 2013). Modelling studies have suggested that the topography of the ice sheet can have a significant impact on mean sea level pressure, which affects the location of the wind-driven gyre circulation in the subpolar North Atlantic, and on the strength of the AMOC, resulting in warming in the North Atlantic (Colleoni *et al*. 2016, Pausata *et al*. 2011). Storm tracks in the North Atlantic have also been shown to be affected, with associated impacts on precipitation and snowfall in northern Europe (e.g. Colleoni *et al*. 2016, Kageyama and Valdes 2000, Lofverstrom *et al*. 2014). Combined, the *modice* and *LGC* ensembles capture changes in climate and ice sheet extent ranging from interglacial states to full glacial conditions.

Experimental design

The emulator was used to simulate global changes in temperature and precipitation over firstly the last 800 kyr (the late Pleistocene) and secondly the last 2.58 Myr (the entire Pleistocene). The late Pleistocene was initially selected because a range of high-resolution palaeo records exist, of multiple variables and in multiple locations, providing both forcing data for the emulator and proxy climate reconstructions which can be compared to the emulator results. However, some of the ODP sites provide temperature reconstructions going further back in time and so the emulator was also run back to match these. The climate variables were emulated at 1 kyr intervals over the time period.

The palaeo data used to force the emulator are illustrated in the Supplementary Material (Fig. S3). For the first 800 kya, orbital data (Fig. S3a) were calculated using the method of Laskar *et al*. (2004). A composite record of observed atmospheric CO_2 was used (Fig. S3b), which was measured from the Dome C ice cores from Antarctica (Bereiter *et al*. 2015). The sea level stack reconstructed from $\delta^{18}O$ data from ocean sediment cores produced by Spratt and Lisiecki (2016) was used in order to provide the GSL index used as the fifth forcing parameter to the emulator (Fig. S3c), which represents changes in global ice sheet volume. We equate a GSL index value of 0.0 with the modern ice sheet volume, and a GSL value of -125 metres with the ice volume of the LGM. For the remainder of the Pleistocene, the emulator was run with orbital parameters calculated in the same way as in the 800 kyr simulations (Laskar *et al*. 2004). However, CO_2 and GSL were from different sources: i) CO_2 was taken from the Van de Wal *et al*. (2011) record, which is a model-derived CO_2 signal that is consistent with global sea level reconstructions (De Boer *et al*. 2010, Van de Wal *et al*. 2011); and ii) GSL was taken from the De Boer *et al*. (2010) study, who used 1-D ice sheet models to reconstruct sea level over the last

20 Myr. These two records were spliced with the existing 800 kyr records (see above) during the overlapping 5 kyr (i.e. 795-800 ka). Otherwise, the emulator was run in the same way as discussed here and in the Supplementary Material.

Concerning the individual simulations, firstly the temperature and precipitation emulators were forced with all of the forcings for the full 2.58 Myr, and the climate response at the different sites with proxy data compared (where proxy data were available). A set of sensitivity simulations, covering just the last 800 kyr, was then carried out in order to investigate the sensitivity of the sites, firstly considering a shorter time period (the last 21 kya) and secondly with the proxy data set as either a constant average, constant PI, constant valid first timeslice (as some of the proxy data contain missing timeslices), constant LGM and randomly generated values. Finally, again for just the last 800 kyr, the emulator was run with a single forcing or subset of the forcings being incrementally varied, following the multi-variate factorisation methods of Lunt *et al*. (2021), whilst the other forcings were held constant at PI values.

Nomenclature and factorisation

For the single forcing simulations presented here, the nomenclature uses a binary system of five digits representing (in order) CO_2, obliquity, eccentricity, precession and ice, where 0 = constant PI values and 1 = varying values. Hence E_{00000} (the ‘no drivers’ simulation) has all drivers set to constant PI values, E_{10000} has varying CO_2 but everything else is set to constant PI values, and so on, up to E_{11111} where all forcings are varying (the ‘all drivers’ simulation). This gave a total of 32 simulations, capturing all possible combinations of forcings and enabling a calculation of all possible pathways, of which there were 120, within the factorisation (see below).

The linear factorisation method consists of incrementally adding in one driver at each step, to create a single pathway from a ‘no drivers’ simulation to an ‘all drivers’ simulation (Lunt *et al*. 2021); one pathway is presented in ‘Single driver experiments: linear factorisation’ in the Results. However, a disadvantage of this method is that the order in which the drivers are added at each step will determine the overall contribution of the driver. Moreover, when there are more than two dimensions (in our case there are five, corresponding to the five driving components), the linear factorisation method only possesses three out of the four desirable properties of uniqueness, symmetry, purity and completeness (see Lunt *et al*. 2021 for more details). Other multi-variate factorisation methods were therefore developed by Lunt *et al*. (2021), which do possess all of the above four properties, and therefore following an initial linear factorisation here we also use the linear-sum/shared-interaction method. Unlike linear factorisation, this method compares all possible pathways within the factorisation, averaging across each dimension to get to the mean contribution from each driver (Lunt *et al*. 2021).

Proxy data

A number of different reconstructions of palaeoclimate were used for comparison to the emulator results. These were taken from several different locations, shown in the Supplementary Material (Fig. S4), and were selected due to their high temporal resolution and coverage of the Pleistocene. These records consist of:

- Reconstructed Antarctic SAT, derived from Deuterium data from the Dome C ice core (Jouzel *et al*. 2007).
- Reconstructed SST data based on δ^{18}O and alkenone data from ocean sediment cores collected as part of the ODP. Records were collected from the North Atlantic (ODP site 982; Lawrence *et al*. 2009), the Arabian Sea (ODP 722; Herbert *et al*. 2010), the South China Sea (ODP 1143; Li *et al*. 2011), the Eastern Equatorial Pacific (ODP 846; Herbert *et al*. 2010), and the sub-Antarctic South Atlantic (ODP 1090; Martinez-Garcia *et al*. 2010).
- δ^{18}O data measured from cave speleothems, which is taken to be a proxy for variations in the strength of East Asian monsoon rainfall. Records from three caves were used to provide full coverage of the last 640 kyr, namely the Hulu Cave in eastern China (Wang *et al*. 2008), the Sanbao Cave in central China (Wang *et al*. 2001), and the Dongge cave in southern China (Cheng *et al*. 2016).

Palaeo SST data is used for comparison with emulated SAT because multiple records exist from varying global locations, which cover hundreds of thousands of years at a sufficient resolution to capture orbital cycles. The offset between SST and SAT at a single location is small compared with the local glacial-interglacial cyclicity. Each proxy data set was compared to emulated mean annual SAT or precipitation for the appropriate grid box for each site. Temperature and precipitation data are generally shown as anomalies compared to PI conditions at the different locations, taken from each proxy data set in the case of SST, and the PI control simulation in the case of SAT and precipitation.

Arcsin Mielke measure

In order to obtain some quantitative metric as to the 'goodness of fit' and skill of the emulator, for each site we generate a skill score for temperature (SAT and SST) using the non-dimensional arcsin Mielke measure, M (Watterson 1996, Watterson *et al*. 2014). This is the mean square error (MSE), non-dimensionalised by the variance (Watterson *et al*. 2014). For emulated values x and proxy data y:

$$M = (2/\pi)\arcsin\left\{1 - \text{mse}/\left[V_x + V_y + (G_x - G_y)^2\right]\right\} \times 1000 \quad \textbf{Eq. (1)}$$

where V is the variance and G is the mean. The resulting skill score has a maximum possible value (for MSE = 0) of 1000, while a score of 0 or negative values indicates no skill (Watterson *et al*. 2014). For the model-data comparison, M is calculated between the emulated (using the E_{11111} simulation) and proxy data at the same site (such as ODP 1090 versus proxy ODP 1090) and at every possible combination of sites (such as emulated ODP 1090 versus proxy ODP 1143). For the sensitivity tests, M is calculated between the emulated (using the E_{11111} simulation) and the modified (to be constant) proxy data at the same site. For the linear and linear-sum/shared-interaction factorisation, M is calculated between the E_{11111} simulation and all other simulations (i.e. treating the E_{11111} simulation as 'observations'). Whereas the figures presented here show anomalies (relative to the PI), the M scores were calculated using the emulated and proxy absolute values.

DATA AVAILABILITY

Code for the wrapper of the GP package is available from https://github.com/cwilliams2020-new/Pleistocene_emulator

REFERENCES

Abe-Ouchi, A., Saito, F., Kageyama, M., Braconnot, P., Harrison, S. P., Lambeck, K., . . . Takahashi, K. (2015). Ice-sheet configuration in the CMIP5/PMIP3 Last Glacial Maximum experiments. *Geoscientific Model Development, 8*(11), 3621-3637. doi:10.5194/gmd-8-3621-2015

Abe-Ouchi, A., Saito, F., Kawamura, K., Raymo, M. E., Okuno, J., Takahashi, K., & Blatter, H. (2013). Insolation-driven 100,000-year glacial cycles and hysteresis of ice-sheet volume. *Nature, 500*(7461), 190-+. doi:10.1038/nature12374

Annan, J. D., & Hargreaves, J. C. (2013). A new global reconstruction of temperature changes at the Last Glacial Maximum. *Climate of the Past, 9*(1), 367-376. doi:10.5194/cp-9-367-2013

Araya-Melo, P. A., Crucifix, M., & Bounceur, N. (2015). Global sensitivity analysis of the Indian monsoon during the Pleistocene. *Climate of the Past, 11*(1), 45-61. doi:10.5194/cp-11-45-2015

Armstrong, E., Valdes, P., House, J., & Singarayer, J. (2016). The role of CO_2 and dynamic vegetation on the impact of temperate land-use change in the HadCM3 coupled climate model. *Earth Interactions, 20*, 1-20. doi:10.1175/Ei-D-15-0036.1

Bereiter, B., Eggleston, S., Schmitt, J., Nehrbass-Ahles, C., Stocker, T. F., Fischer, H., . . . Chappellaz, J. (2015). Revision of the EPICA Dome C CO_2 record from 800 to 600 kyr before present. *Geophysical Research Letters, 42*(2), 542-549. doi:10.1002/2014gl061957

Berger, A. (1978). Long-term variations of daily insolation and Quaternary climatic changes. *Journal of the Atmospheric Sciences, 35*(12), 2362-2367. doi:10.1175/1520-0469(1978)035<2362:Ltvodi>2.0.Co;2

Bounceur, N., Crucifix, M., & Wilkinson, R. D. (2015). Global sensitivity analysis of the climate-vegetation system to astronomical forcing: an emulator-based approach. *Earth System Dynamics, 6*, 205-224. doi:10.5194/esd-6-205-2015

Cheng, H., Edwards, R. L., Sinha, A., Spotl, C., Yi, L., Chen, S. T., . . . Zhang, H. W. (2016). The Asian monsoon over the past 640,000 years and ice age terminations. *Nature, 534*(7609), 640-+. doi:10.1038/nature18591

Colleoni, F., Wekerle, C., Naslund, J. O., Brandefelt, J., & Masina, S. (2016). Constraint on the penultimate glacial maximum Northern Hemisphere ice topography (approximate to 140 kyrs BP). *Quaternary Science Reviews, 137,* 97-112. doi:10.1016/j.quascirev.2016.01.024

Cox, P. M., Betts, R. A., Jones, C. D., Spall, S. A., & Totterdell, I. J. (2002). Modelling vegetation and the carbon cycle as interactive elements of the climate system. In R. Pearce (Ed.), *Meteorology at the Millennium* (pp. 259-279). San Diego CA, USA: Academic Press.

Crowley, T. J. (1992). North Atlantic Deep Water Cools the Southern Hemisphere. *Paleoceanography, 7*(4), 489-497. doi:10.1029/92pa01058

De Boer, B., van de Wal, R. S. W., Bintanja, R., Lourens, L. J., & Tuenter, E. (2010). Cenozoic global ice-volume and temperature simulations with 1-D ice-sheet models forced by benthic $\delta^{18}O$ records. *Annals of Glaciology*, 51, 23-33, https://doi.org/10.3189/172756410791392736

Dowsett, H. J., Dolan, A., Rowley, D., Moucha, R., Forte, A. M., Mitrovica, J. X., . . . Haywood, A. (2016). The PRISM4 (mid-Piacenzian) paleoenvironmental reconstruction. *Climate of the Past*, *12*(7), 1519-1538. doi:10.5194/cp-12-1519-2016

Erb, M. P., Jackson, C. S. As 333 & Broccoli, A. J. (2015). Using Single-Forcing GCM Simulations to Reconstruct and Interpret Quaternary Climate Change. *Journal of Climate, 28*(24), 9746-9767. https://doi.org/10.1175/JCLI-D-15-0329.1

Greve, R. (2000). On the response of the Greenland ice sheet to greenhouse climate change. *Climatic Change*, *46*(3), 289-303. doi:10.1023/A:1005647226590

Hays, J. D., Imbrie, J., & Shackleton, N. J. (1976). Variations in the earth's orbit: Pacemaker of the Ice Ages. *Science, 194,* 1121-1132.

Haywood, A. M., Dowsett, H. J., Dolan, A. M., Rowley, D., Abe-Ouchi, A., Otto-Bliesner, B., . . . Salzmann, U. (2016). The Pliocene Model Intercomparison Project (PlioMIP) Phase 2: scientific objectives and experimental design. *Climate of the Past, 12*(3), 663-675. doi:10.5194/cp-12-663-2016

Herbert, T. D., Peterson, L. C., Lawrence, K. T., & Liu, Z. H. (2010). Tropical ocean temperatures over the past 3.5 million years. *Science*, *328*(5985), 1530-1534. doi:10.1126/science.1185435

Holden, P. B., Edwards, N. R., Oliver, K. I. C., Lenton, T. M., & Wilkinson, R. D. (2010). A probabilistic calibration of climate sensitivity and terrestrial carbon change in GENIE-1. *Climate Dynamics, 35*(5), 785-806. doi:10.1007/s00382-009-0630-8

Holden, P. B., Edwards, N. R., Rangel, T. F., Pereira, E. B., Tran, G. T., & Wilkinson, R. D. (2019). PALEO-PGEM v1.0: a statistical emulator of Pliocene-Pleistocene climate. *Geoscientific Model Development, 12*(12), 5137-5155. doi:10.5194/gmd-12-5137-2019

Hoogakker, B. A. A., Smith, R. S., Singarayer, J. S., Marchant, R., Prentice, I. C., Allen, J. R. M., Anderson, R. S., Bhagwat, S. A., Behling, H., Borisova, O., Bush, M., Correa-Metrio, A., de Vernal, A., Finch, J. M., Fréchette, B., Lozano-Garcia, S., Gosling, W. D., Granoszewski, W., Grimm, E. C., Grüger, E., Hanselman, J., Harrison, S. P., Hill, T. R., Huntley, B., Jiménez-Moreno, G., Kershaw, P., Ledru, M.-P., Magri, D., McKenzie, M., Müller, U., Nakagawa, T., Novenko, E., Penny, D., Sadori, L., Scott, L., Stevenson, J., Valdes, P. J., Vandergoes, M., Velichko, A., Whitlock, C., and Tzedakis, C. (2016). Terrestrial biosphere changes over the last 120 kyr. *Climate of the Past*, 12, 51-73, https://doi.org/10.5194/cp-12-51-2016

Huybrechts, P., & de Wolde, J. (1999). The dynamic response of the Greenland and Antarctic ice sheets to multiple-century climatic warming. *Journal of Climate, 12*(8), 2169-2188. doi:10.1175/1520-0442(1999)012<2169:Tdrotg>2.0.Co;2

Ivanovic, R. F., Gregoire, L. J., Kageyama, M., Roche, D. M., Valdes, P. J., Burke, A., Drummond, R., Peltier, W. R., and Tarasov, L. (2016). Transient climate simulations of the deglaciation 21-9 thousand years before present (version 1) - PMIP4 Core experiment design and boundary conditions, *Geoscientific Model Development*, 9, 2563-2587, https://doi.org/10.5194/gmd-9-2563-2016

Johnson, J. S., Cui, Z., Lee, L. A., Gosling, J. P., Blyth, A. M., & Carslaw, K. S. (2015). Evaluating uncertainty in convective cloud microphysics using statistical emulation. *Journal of Advances in Modeling Earth Systems, 7*(1), 162-187. doi:10.1002/2014ms000383

Jouzel, J., Masson-Delmotte, V., Cattani, O., Dreyfus, G., Falourd, S., Hoffmann, G., . . . Wolff, E. W. (2007). Orbital and millennial Antarctic climate variability over the past 800,000 years. *Science, 317* (5839), 793-796. doi:10.1126/science.1141038

Kageyama, M., & Valdes, P. J. (2000). Impact of the North American ice-sheet orography on the Last Glacial Maximum eddies and snowfall. *Geophysical Research Letters, 27*(10), 1515-1518. doi:10.1029/1999gl011274

Kawamura, K., Parrenin, F., Lisiecki, L., Uemura, R., Vimeux, F., Severinghaus, J. P., . . . Watanabe, O. (2007). Northern Hemisphere forcing of climatic cycles in Antarctica over the past 360,000 years. *Nature, 448*(7156), 912-914. doi:10.1038/Nature06015

Kennedy, M. C., & O'Hagan, A. (2000). Predicting the output from a complex computer code when fast approximations are available. *Biometrika, 87*(1), 1-13. doi:10.1093/biomet/87.1.1

Kennett, J. P., & Stott, L. D. (1991). Abrupt deep-sea warming, palaeoceanographic changes and benthic extinctions at the end of the Paleocene. *Nature, 353*(6341), 225-229. doi:10.1038/353225a0

Kleman, J., Fastook, J., Ebert, K., Nilsson, J., & Caballero, R. (2013). Pre-LGM Northern Hemisphere ice sheet topography. *Climate of the Past, 9*(5), 2365-2378. doi:10.5194/cp-9-2365-2013

Laskar, J., Robutel, P., Joutel, F., Gastineau, M., Correia, A. C. M., & Levrard, B. (2004). A long-term numerical solution for the insolation quantities of the Earth. *Astronomy & Astrophysics, 428*(1), 261-285. doi:10.1051/0004-6361:20041335

Lawrence, K. T., Herbert, T. D., Brown, C. M., Raymo, M. E., & Haywood, A. M. (2009). High-amplitude variations in North Atlantic sea surface temperature during the early Pliocene warm period. *Paleoceanography, 24*. doi:10.1029/2008pa001669

Li, L., Li, Q. Y., Tian, J., Wang, P. X., Wang, H., & Liu, Z. H. (2011). A 4-Ma record of thermal evolution in the tropical western Pacific and its implications on climate change. *Earth and Planetary Science Letters, 309*(1-2), 10-20. doi:10.1016/j.epsl.2011.04.016

Lisiecki, L. E., & Raymo, M. E. (2005). A Pliocene-Pleistocene stack of 57 globally distributed benthic $\delta^{18}O$ records. *Paleoceanography, 20*(1). doi:10.1029/2004pa001071

Lisiecki, L. E., & Raymo, M. E. (2007). Plio-Pleistocene climate evolution: trends and transitions in glacial cycle dynamics. *Quaternary Science Reviews, 26*(1-2), 56-69. doi:10.1016/j.quascirev.2006.09.005

Lofverstrom, M., Caballero, R., Nilsson, J., & Kleman, J. (2014). Evolution of the large-scale atmospheric circulation in response to changing ice sheets over the last glacial cycle. *Climate of the Past, 10*(4), 1453-1471. doi:10.5194/cp-10-1453-2014

Lord, N. S. (2017). *Projecting long-term past and future climate change within the context of post-closure performance assessments for disposal of radioactive waste.* (PhD Thesis). University of Bristol, UK, Bristol, UK.

Lord, N. S., Crucifix, M., Lunt, D. J., Thorne, M. C., Bounceur, N., Dowsett, H., . . . Ridgwell, A. (2017). Emulation of long-term changes in global climate: Application to the mid-Pliocene and future. *Climate of the Past, 13*, 1539-1571. doi:10.5194/cp-13-1539-2017

Lord, N. S., Lunt, D. J. & Thorne, M. C. (2019). Modelling changes in climate over the next 1 million years. *SKB Public Report TR-19-09*, https://www.skb.com/publication/2494175

Lord, N. S., Ridgwell, A., Thorne, M. C., & Lunt, D. J. (2016). An impulse response function for the "long tail" of excess atmospheric CO_2 in an Earth system model. *Global Biogeochemical Cycles, 30*(1), 2-17. doi:10.1002/2014gb005074

Loutre, M. F. (1993). *Paramètres orbitaux et cycles diurne et saisonnier des insolations.* (PhD). Université catholique de Louvain, Louvain-la-Neuve, Belgium.

Loutre, M. F., & Berger, A. (2000a). No glacial-interglacial cycle in the ice volume simulated under a constant astronomical forcing and a variable CO_2. *Geophysical Research Letters, 27*(6), 783-786. doi:10.1029/1999gl006081

Lunt, D. J., Chandan, D., Haywood, A. M., Lunt, G. M., Rougier, J. C., Salzmann, U., Schmidt, G. A. & Valdes, P. J. (2021). Multi-variate factorisation of numerical simulations. *Geoscientific Model Development*. 14, 4307-4317. https://doi.org/10.5194/gmd-14-4307-2021.

Martinez-Garcia, A., Rosell-Mele, A., McClymont, E. L., Gersonde, R., & Haug, G. H. (2010). Subpolar Link to the Emergence of the Modern Equatorial Pacific Cold Tongue. *Science, 328*(5985), 1550-1553. doi:10.1126/science.1184480

Marzocchi, A., Lunt, D. J., Flecker, R., Bradshaw, C. D., Farnsworth, A., & Hilgen, F. J. (2015). Orbital control on late Miocene climate and the North African monsoon: insight from an ensemble of sub-precessional simulations. *Climate of the Past, 11*(10), 1271-1295. doi:10.5194/cp-11-1271-2015

Milankovitch, M. (1941). *Kanon der Erdbestrahlung und seine Anwendung auf das Eiszeitenproblem* (J. Israel Program for Scientific Translations, Trans. Vol. 33). Washington D.C.: U.S. Department of Commerce and National Science Foundation.

Oakley, J., & O'Hagan, A. (2002). Bayesian inference for the uncertainty distribution of computer model outputs. *Biometrika, 89*(4), 769-784. doi:10.1093/biomet/89.4.769

PalMod (2023). *From the Last Interglacial to the Anthropocene - Modelling a Complete Glacial Cycle*. https://www.palmod.de/home

Pausata, F. S. R., Li, C., Wettstein, J. J., Kageyama, M., & Nisancioglu, K. H. (2011). The key role of topography in altering North Atlantic atmospheric circulation during the last glacial period. *Climate of the Past, 7*(4), 1089-1101. doi:10.5194/cp-7-1089-2011

Peltier, W. R. (2004). Global glacial isostasy and the surface of the ice-age earth: The ice-5G (VM2) model and grace. *Annual Review of Earth and Planetary Sciences, 32*, 111-149. doi:10.1146/annurev.earth.32.082503.144359

Prescott, C. L., Haywood, A. M., Dolan, A. M., Hunter, S. J., Pope, J. O., & Pickering, S. J. (2014). Assessing orbitally-forced interglacial climate variability during the mid-Pliocene Warm Period. *Earth and Planetary Science Letters, 400*, 261-271. doi:10.1016/j.epsl.2014.05.030

Ridley, J. K., Huybrechts, P., Gregory, J. M., & Lowe, J. A. (2005). Elimination of the Greenland ice sheet in a high CO_2 climate. *Journal of Climate, 18*(17), 3409-3427. doi:10.1175/Jcli3482.1

Sacks, J., Welch, W. J., Mitchell, T. J., & Wynn, H. P. (1989). Design and analysis of computer experiments. *Statistical Science, 4*(4), 409-423. doi:10.1214/ss/1177012413

Schmidt, P., Lund, B., Naslund, J. O., & Fastook, J. (2014). Comparing a thermo-mechanical Weichselian Ice Sheet reconstruction to reconstructions based on the sea level equation: aspects of ice configurations and glacial isostatic adjustment. *Solid Earth, 5*(1), 371-388. doi:10.5194/se-5-371-2014

Singarayer, J. S., & Valdes, P. J. (2010). High-latitude climate sensitivity to ice-sheet forcing over the last 120 kyr. *Quaternary Science Reviews, 29*(1-2), 43-55. doi:10.1016/j.quascirev.2009.10.011

Spratt, R. M., & Lisiecki, L. E. (2016). A Late Pleistocene sea level stack. *Climate of the Past, 12*(4), 1079-1092. doi:10.5194/cp-12-1079-2016

Stap, L. B., van de Wal, R. S. W., de Boer, B., Bintanja, R., & Lourens, L. J. (2014). Interaction of ice sheets and climate during the past 800 000 years. *Climate of the Past, 10*(6), 2135-2152. doi:10.5194/cp-10-2135-2014

Stone, E. J., Lunt, D. J., Rutt, I. C., & Hanna, E. (2010). Investigating the sensitivity of numerical model simulations of the modern state of the Greenland ice-sheet and its future response to climate change. *Cryosphere*, *4*(3), 397-417. doi:10.5194/tc-4-397-2010

Timmermann, A., Yun, KS., Raia, P. *et al*. (2022). Climate effects on archaic human habitats and species successions. *Nature*, 604, 495-501. https://doi.org/10.1038/s41586-022-04600-9

Valdes, P. J., Armstrong, E., Badger, M. P. S., Bradshaw, C. D., Bragg, F., Crucifix, M., . . . Williams, J. H. T. (2017). The BRIDGE HadCM3 family of climate models: HadCM3@Bristol v1.0. *Geoscientific Model Development*. doi:10.5194/gmd-2017-16

Van de Wal, R. S. W., de Boer, B., Lourens, L. J., Köhler, P. & Bintanja, R. (2011). Reconstruction of a continuous high-resolution CO2 record over the past 20 million years. *Climate of the Past*, 7, 1459-1469, https://doi.org/10.5194/cp-7-1459-2011

Wang, Y. J., Cheng, H., Edwards, R. L., An, Z. S., Wu, J. Y., Shen, C. C., & Dorale, J. A. (2001). A high-resolution absolute-dated Late Pleistocene monsoon record from Hulu Cave, China. *Science, 294*(5550), 2345-2348. doi:10.1126/science.1064618

Wang, Y. J., Cheng, H., Edwards, R. L., Kong, X. G., Shao, X. H., Chen, S. T., . . . An, Z. S. (2008). Millennial- and orbital-scale changes in the East Asian monsoon over the past 224,000 years. *Nature*, *451*(7182), 1090-1093. doi:10.1038/nature06692

Watterson, I. G. (1996). Non-dimensional measures of climate model performance. *International Journal of Climatology*. 16. https://doi.org/10.1002/(SICI)1097-0088(199604)16:4<379::AID-JOC18>3.0.CO;2-U

Watterson, I. G., Bathols, J.& Heady, C. (2014). What Influences the Skill of Climate Models over the Continents? *Bulletin of the American Meteorological Society*. 95 (5). doi: https://doi.org/10.1175/BAMS-D-12-00136.1

Wilkinson, R. D. (Ed.) (2010). *Bayesian calibration of expensive multivariate computer experiments*. Chichester, UK: John Wiley & Sons, Ltd.

Williams, C. J. R., Lunt, D. J., Kennedy-Asser, A. T. & Lord, N. S. (2022). Uncertainties in modelled climate changes at Forsmark over the next 1 million years. *SKB Public Report TR-22-02*, https://www.skb.com/publication/2501108

Winkelmann, R., Levermann, A., Ridgwell, A., & Caldeira, K. (2015). Combustion of available fossil fuel resources sufficient to eliminate the Antarctic Ice Sheet. *Science Advances, 1*(8). doi:10.1126/sciadv.1500589

ACKNOWLEDGMENTS

The work described in this paper was carried out in the framework of a project funded by Posiva Oy and SKB, funding CJRW and NSL. Some of the earlier development work was carried out in the framework of a project funded by RWM. This work was carried out using the computational facilities of the Advanced Computing Research Centre, University of Bristol - http://www.bris.ac.uk/acrc/.

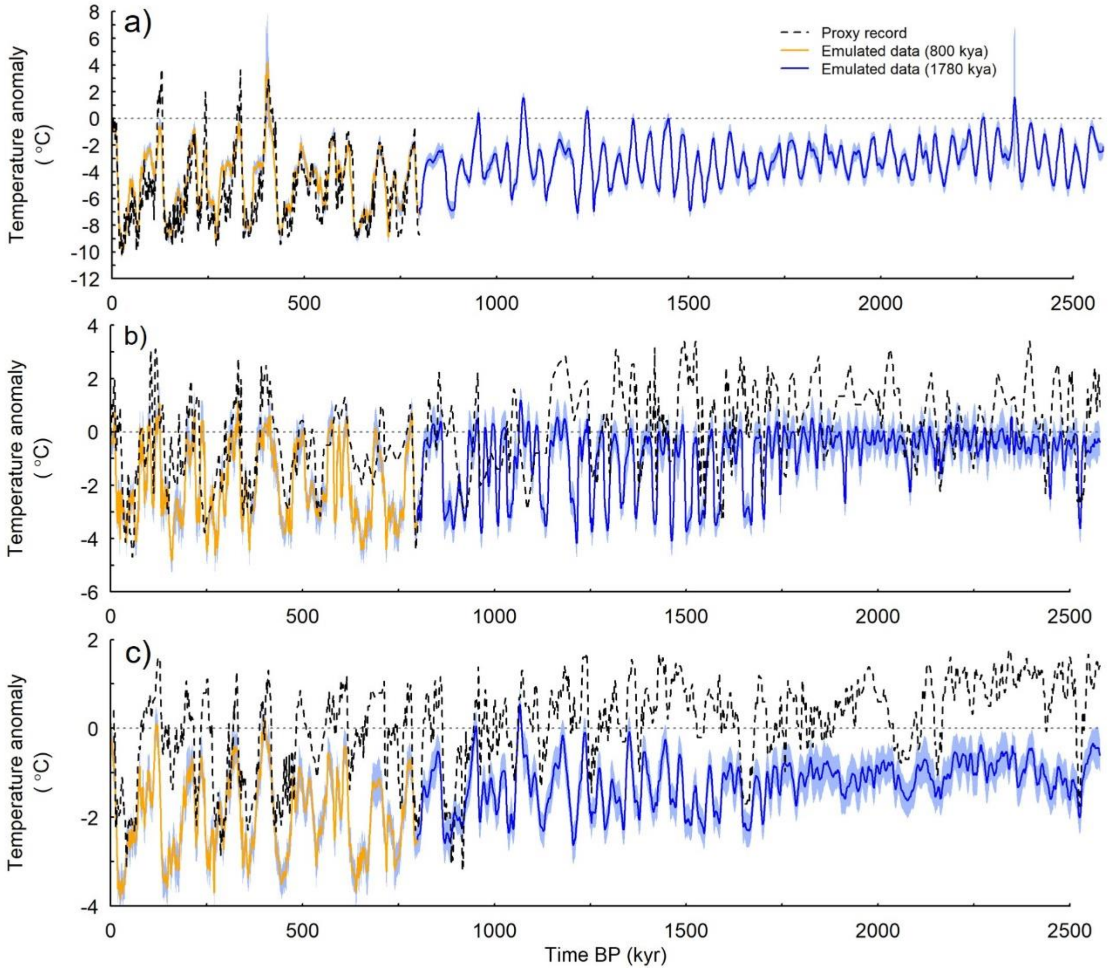
TABLES

		Emulated					
		Dome C	**ODP 982**	**ODP 722**	**ODP 1143**	**ODP 846**	**ODP 1090**
Proxy sites	**Dome C**	539.413	1.08866	0.45785	0.390459	0.410267	0.578972
	ODP 982	0.506231	38.406 *27.9499*	10.2852 *7.79169*	4.37246	6.14715 .	65.8864
	ODP 722	0.250364	2.61711 *2.40548*	54.0366 *34.7502*	179.624	224.687 *115.407*	2.10363
	ODP 1143	0.22107	2.12009	26.021	443.81	79.2869	1.64023
	ODP 846	0.267516	3.56242 *4.20367*	414.315 *300.402*	25.5858	74.9473 *175.953*	3.31721
	ODP 1090	1.11522	28.568	2.10431	1.46071	1.83273	9.18878

Table 1. M scores between emulated and proxy data, for E_{11111}, at every possible combination of sites. Values not in italics are calculated over the last 800 kya, values in italics are calculated over the full 2.58 Myr (at sites with proxy data extending back this far).

803

804 **FIGURES**



805

806

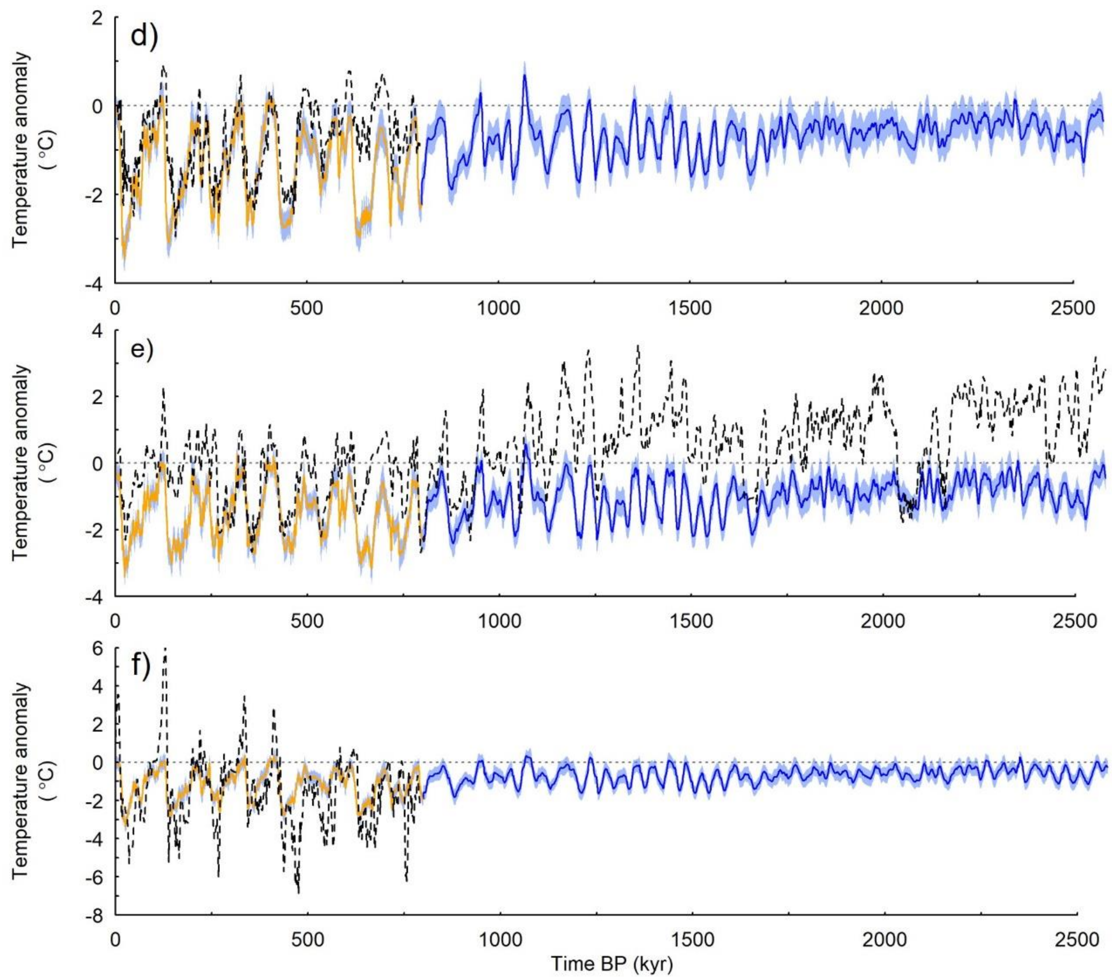


Figure 1. Timeseries of temperature anomaly (compared to a pre-industrial control simulation, °C) for the last 2.58 Myr at six locations, reconstructed from proxy data, where available (dashed black lines), and modelled every 1 kyr using the emulator E_{11111} simulation (solid lines, either orange for the last 800 kya or blue for the remaining 1780 kya): a) SAT from Dome C, Antarctica; b) ODP 982, North Atlantic, where SST is reconstructed from a $\delta^{18}O$ record from ocean sediment cores; c) ODP 722, Arabian Sea, reconstructed from an alkenone record from ocean sediment cores; d) ODP 1143, South China Sea, reconstructed from a $\delta^{18}O$ record from ocean sediment cores; e) ODP 846, Eastern Equatorial Pacific, reconstructed from an alkenone record from ocean sediment cores; f) ODP 1090, Subantarctic South Atlantic, reconstructed from an alkenone record from ocean sediment cores. Temperature is shown as an anomaly compared to the pre-industrial SST from the data set. Light blue error bands represent the emulated grid box posterior variance (1 SD). Note differing y-axes.

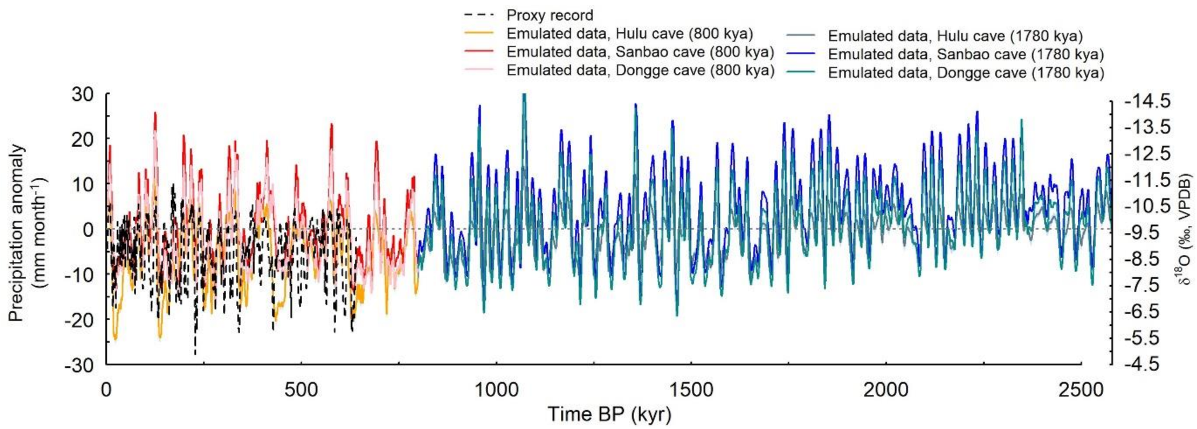
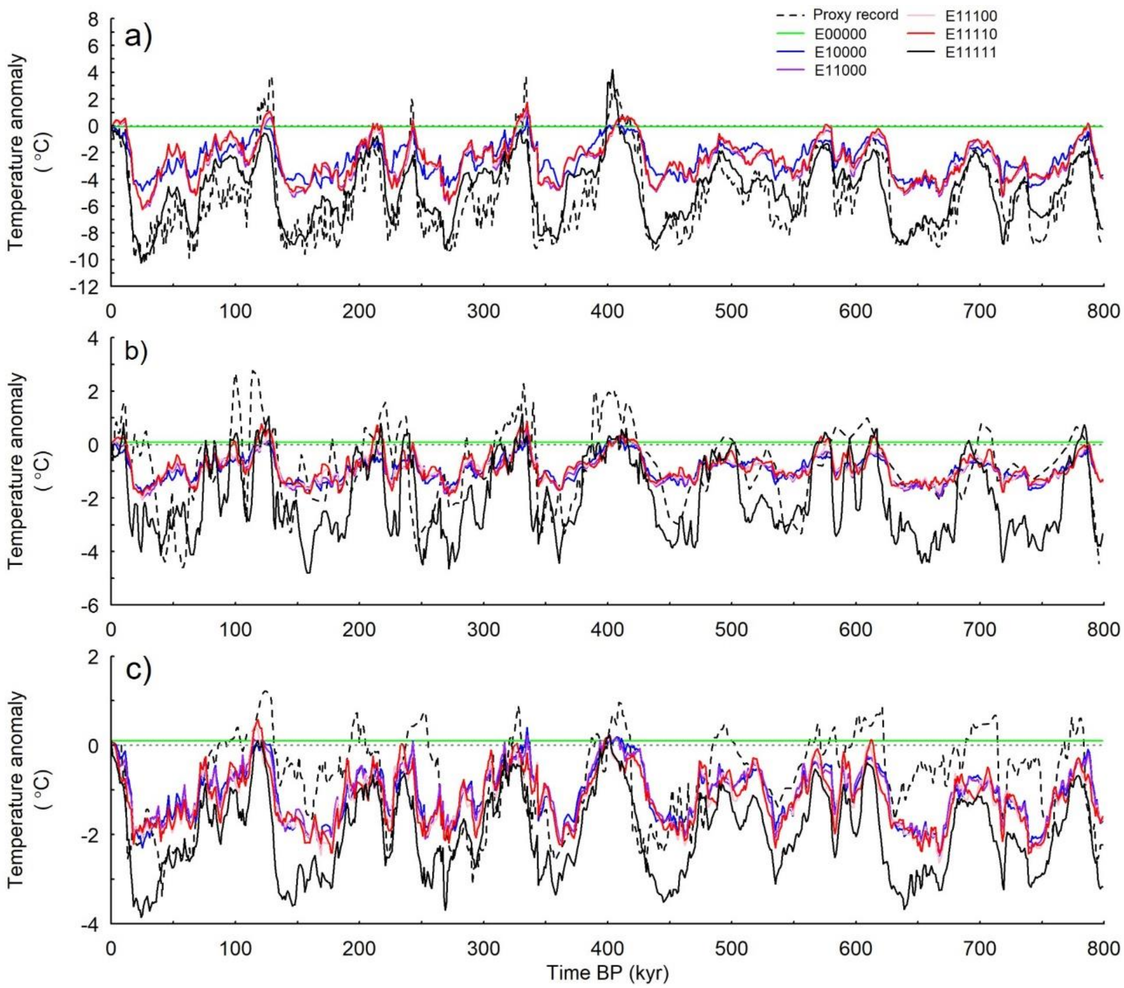


Figure 2. Timeseries of MAP anomaly (compared to a pre-industrial control simulation, mm month^{-1}) for the last 2.58 Myr in China, reconstructed from proxy data (dashed black line) and modelled every 1 kyr using the emulator E_{11111} simulation (solid lines, coloured separately for the last 800 kya and the remaining 1780 kya). Proxy data are from a composite $\delta^{18}O$ record (‰, VPDB) covering the last 640 kyr, constructed from records from the Hulu, Sanbao, and Dongge caves, where cave speleothem $\delta^{18}O$ data are taken to be a proxy for variations in the strength of the East Asian monsoon. To enable comparison, the Hulu and Dongge proxy records are plotted 1.6‰ more negative to account for their higher values (Cheng *et al*. 2016, Wang *et al*. 2001). Emulated anomalies are shown at the grid boxes containing the Hulu (grey/orange solid line), Sanbao (blue/red solid line), and Dongge (turquoise/pink solid line) caves.


a)
b)
c)
Proxy record
E00000
E10000
E11000
E11100
E11110
E11111
Temperature anomaly (°C)
Time BP (kyr)
0
100
200
300
400
500
600
700
800

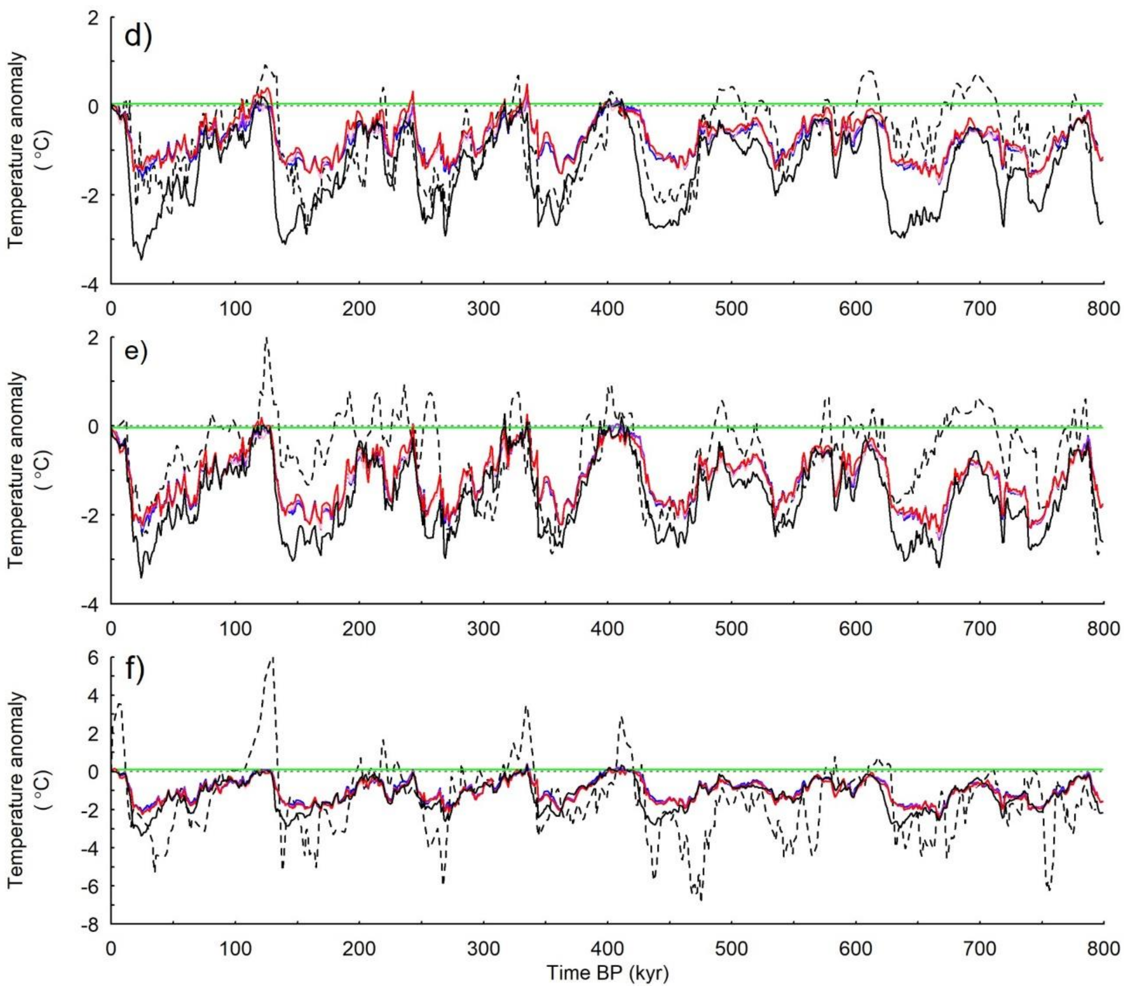


Figure 3. Timeseries of temperature anomaly (compared to a pre-industrial control simulation, °C) for the last 800 kyr at six locations, reconstructed from proxy data (dashed black lines) and modelled every 1 kyr using the emulator, which was forced by each simulation following the linear factorisation of Lunt *et al*. (2021). Other forcings were held constant at pre-industrial values: a) Dome C, Antarctica, where temperature is shown as an anomaly relative to the mean temperature of the last millennium; b) ODP 982, North Atlantic, reconstructed from a $\delta^{18}O$ record from ocean sediment cores; c) ODP 722, Arabian Sea, reconstructed from an alkenone record from ocean sediment cores; d) ODP 1143, South China Sea, reconstructed from a $\delta^{18}O$ record from ocean sediment cores; e) ODP 846, Eastern Equatorial Pacific, reconstructed from an alkenone record from ocean sediment cores; f) ODP 1090, Subantarctic South Atlantic, reconstructed from an alkenone record from ocean sediment cores. SST is shown as an anomaly compared to the pre-industrial SST from the data set. Note differing y-axes.

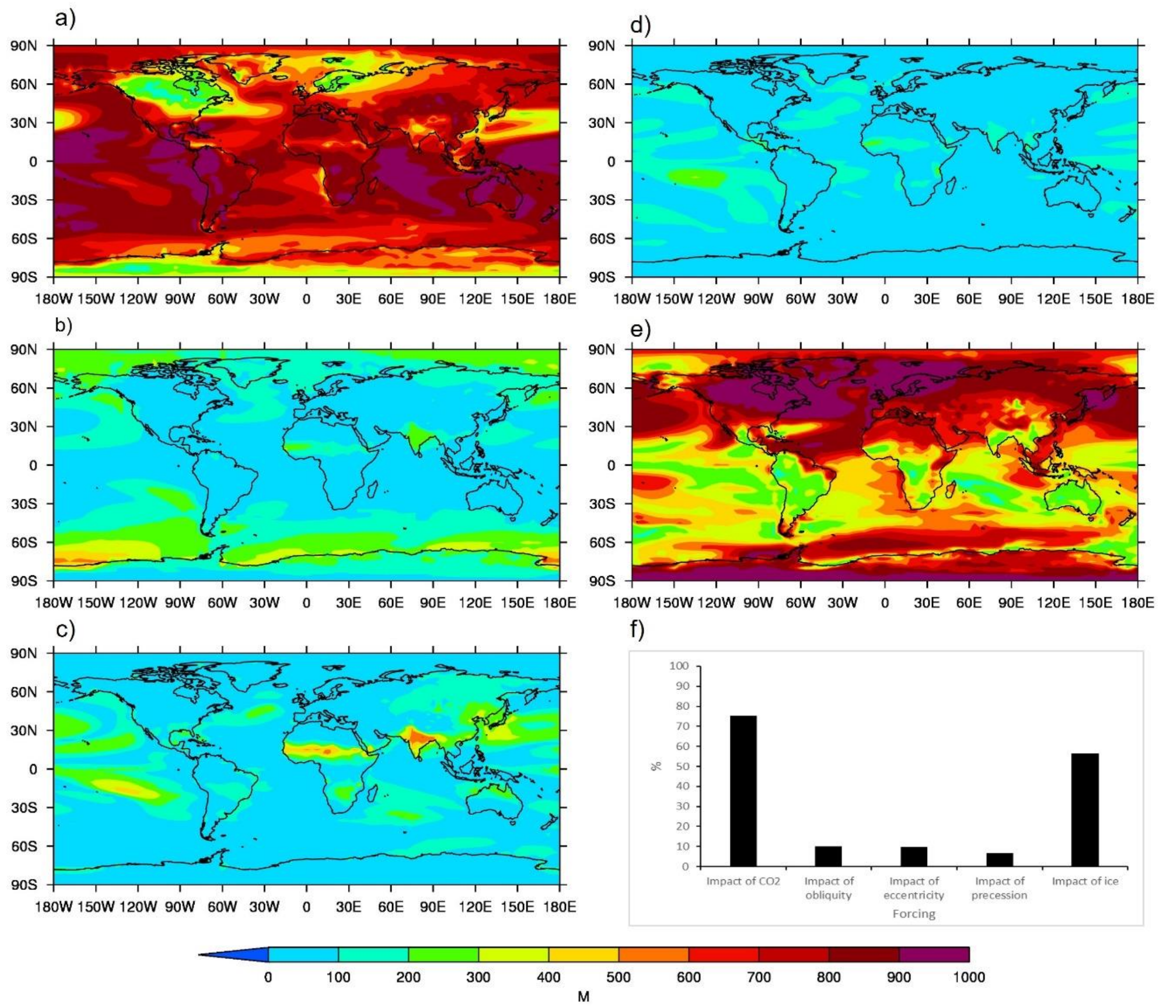


Figure 4. M score differences between emulated temperature from the E_{11111} simulation and that from five other simulations, in which one driving component only is incrementally removed: a) E_{11111}-E_{01111}; b) E_{11111}-E_{10111}; c) E_{11111}-E_{11011}; d) E_{11111}-E_{11101}; e) E_{11111}-E_{11110}; f) Global means of each difference, expressed as a percentage.

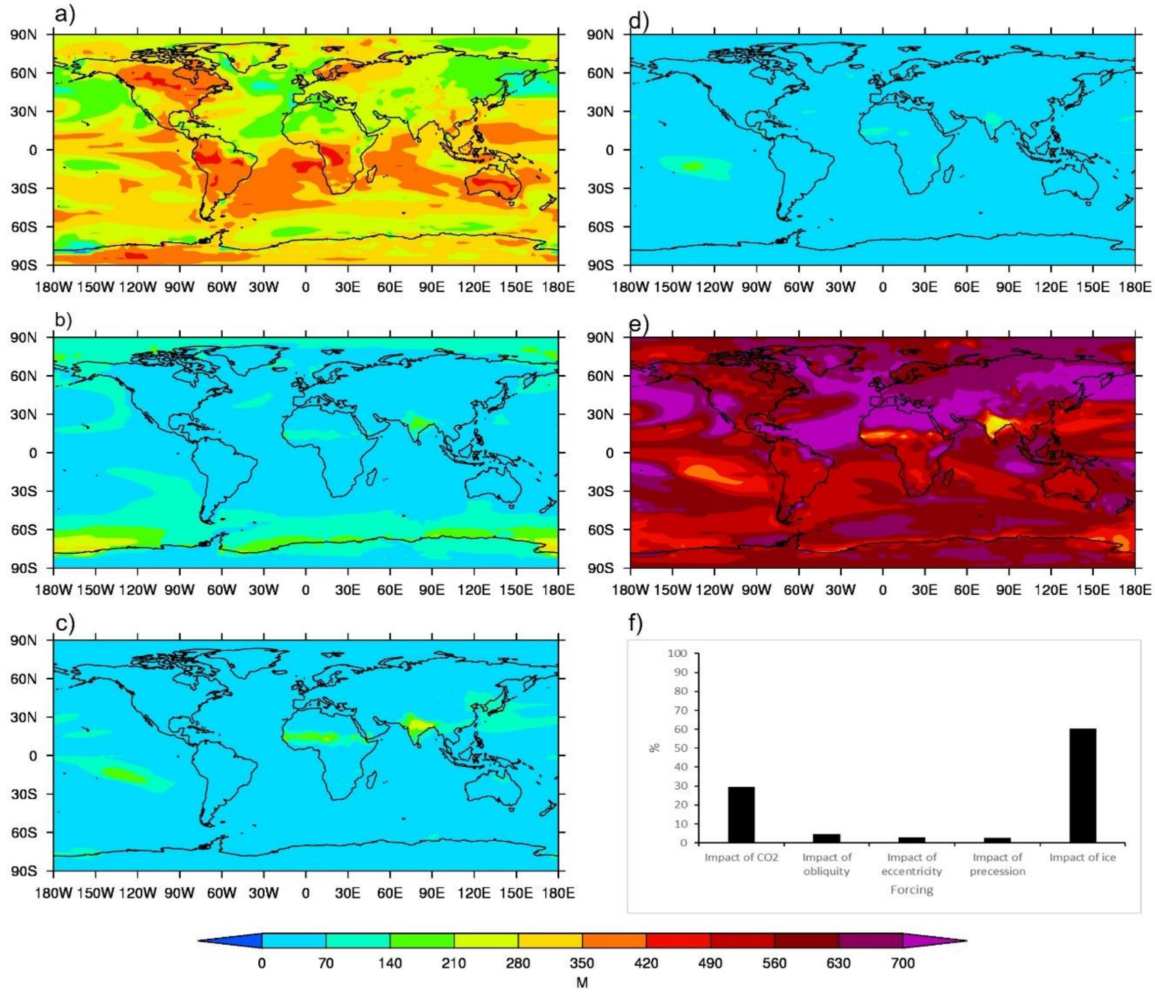


Figure 5. M scores between emulated temperature from the E_{11111} simulation and that from individual simulations following the linear-sum/shared interaction factorization of Lunt *et al*. (2021), averaging over all 120 pathways for each driving component: a) CO_2; b) Obliquity; c) Eccentricity; d) Precession; e) Ice; f) Global means, expressed as a percentage.

Supplementary Material

TEXT

Brief Description of Previously-Published Emulator

The emulator presented in Lord *et al*. (2017) was calibrated on GCM simulations run using the HadCM3 climate model, a fully coupled atmosphere-ocean GCM developed by the UK Met Office. The specific model setup is denoted HadCM3B-M2.1aE and is described in Valdes *et al*. (2017). The GCM is coupled to the land surface scheme MOSES2.1 (Met Office Surface Exchange Scheme), which is in turn coupled to the dynamic vegetation model TRIFFID (Top-down Representation of Interactive Foliage and Flora Including Dynamics) (Cox *et al*. 2002). TRIFFID calculates the global distribution of vegetation based on five plant functional types: broadleaf trees, needleleaf trees, C3 grasses, C4 grasses and shrubs. The horizontal resolution of the atmosphere component is 2.5° latitude by 3.75° longitude with 19 vertical levels, whilst the ocean has a resolution of 1.25° by 1.25° and 20 vertical levels.

When compared with the latest generation of GCMs, such as those included in the IPCC Fifth Assessment Report (IPCC 2013), HadCM3 can no longer be considered as state-of-the-art. However, it is relatively computationally efficient which makes it useful for running experiments that cover relatively long periods of time (of several centuries or longer), as well as for running ensembles with a large number of ensemble members, as is required in this study. For this reason, the model is still widely used in climate research, both in palaeoclimatic studies (e.g. Prescott *et al*. 2014) and in projections of future climate (e.g. Armstrong *et al*. 2016, Lord *et al*. 2017). It has also previously been used in conjunction with a statistical emulator to investigate climate sensitivity (Araya-Melo *et al*. 2015).

Two 60-member ensembles of HadCM3 simulations were used by Lord *et al*. (2017), with input parameters varied across four dimensions: atmospheric CO_2 concentration and the three main orbital parameters of longitude of perihelion (ϖ), obliquity (ε) and eccentricity (e). The ranges of orbital and CO_2 values sampled in the ensembles (see Table S1) included those appropriate for wide range of palaeo and future climate states (Lord *et al*. 2016). Eccentricity and longitude of perihelion were combined under the forms $e\sin\varpi$ and $e\cos\varpi$ given that, in general at any point in the year, insolation can be approximated as a linear combination of these two terms and obliquity (Loutre 1993). Latin hypercube sampling was used in order to optimally fill the 4-dimensional input space.

Two different fixed ice sheet configurations were used, consisting of changes to the Greenland (GrIS) and West Antarctic (WAIS) ice sheets only. One represents their present-day extents, denoted the "*modice*" ensemble, and the other their reduced extents, based on the PRISM4 mid-Piacenzian reconstructions (Dowsett *et al*. 2016), denoted the "*lowice*" ensemble. An additional suite of simulations was also included, produced by (Singarayer and Valdes 2010) (denoted the "*LGC*" ensemble), and also run using HadCM3. These additional simulations are a series of snapshot simulations covering the last glacial cycle (LGC; 120 kyr BP to present day), forced by changes in orbit, atmospheric CO_2 concentration, and ice sheet evolution. Each of the 60 sampled combinations of orbital and CO_2 values therefore had two sets of GCM output; one with *modice* ice sheets and one with *lowice* ice sheets. For more information about the GCM ensembles, and the sampling and GCM simulation post-processing methods used, see Lord *et al*. (2017).

Methodology

To summarise the methodology (see Lord *et al*. 2017 for full details), the input parameter data (denoted $\boldsymbol{D}$ in Lord *et al*. (2017)) for the interglacial emulator was a 120 × 5 matrix ($n \times p$), consisting of 120 GCM simulations (n) and five input factors (ε, $e\sin\varpi$, $e\cos\varpi$, CO_2, and GSL; denoted p). On the other hand, the input data for the glacial emulator was a 122 × 5 matrix. The matrix containing the output data from the GCM (denoted $\boldsymbol{Y}$) had dimensions of 96 × 73 × 120 (longitude × latitude × n) for the interglacial emulator, and 96 × 73 × 122 for the glacial emulator. A PC analysis was performed on each set of output data, the results of which were then used to calibrate the two emulators. Five correlation length hyperparameters (δ) were used, one for each of the five input factors, which describe the smoothness of the climate response in the GCM data to the input conditions. A nugget term (υ) was also used, which accounts for any non-linearity in the output response to the inputs, non-explicitly specified inactive inputs, and the effects of lower-order PCs that are excluded from the emulator. The optimal values for these hyperparameters and the number of PCs retained were calculated during calibration and evaluation of the emulator, discussed in the next section. The GCM data used in this study are mean annual SAT and MAP, although these were each emulated separately using different emulators.

Calibration and Evaluation of Methodology

Four emulators were created in total, two of which were calibrated (and projected) on SAT (an interglacial and a glacial version), and two of which were calibrated (and projected) on precipitation. Before being used, the two SAT emulator configurations were optimised via the method described in Lord *et al*. (2017). This involved generating an ensemble of emulators with different numbers of PCs retained (the majority of the PCs were discarded as they only account for a very small amount of total variation in the GCM data) and different values for a number of hyperparameters used in the emulator, including the correlation length hyperparameters (δ) and the nugget parameter (ν). The performances of the different emulators were compared in order to identify the optimal number of PCs to retain and the hyperparameter values to use. It was demonstrated by Lord (2017) that the optimisation can be carried out on SAT, and then the optimised configuration applied to other climate variables (e.g. precipitation) with no significant loss of performance. Hence, this approach was adopted here.

In order to optimise the two emulators, each was calibrated on the GCM SAT data from its respective ensemble(s) of GCM simulations. The input factors ($\ln(CO_2)$, ε, $e\sin\varpi$, $e\cos\varpi$, and GSL) were standardised prior to the calibration being performed; each was centred in relation to its column mean, and then scaled based on the standard deviation (SD) of the column. Different emulator configurations were tested by varying the number of PCs retained, ranging from 5 to 20, and for each emulator configuration, the correlation length scales δ and nugget ν were optimised by maximisation of the penalized likelihood (see Lord *et al*. 2017 for further details). This optimisation was carried out in log space, ensuring that the optimised hyperparameters would be positive.

The performance of the optimised emulators was evaluated, using a leave-one-out approach and by comparing global SAT produced by the emulator for the LGM to a reconstruction based on proxy data and GCMs. The performance of each emulator was assessed using a leave-one-out cross-validation approach, in which a series of emulators was constructed and used to predict one left out experiment each time. For example, for the interglacial emulator (120 experiments), 120 separate emulators were calibrated with one experiment left out of each. This left out experiment was then reproduced using the corresponding emulator and the results were compared with the actual experiment results. The number of grid boxes for each experiment calculated to lie within different SD bands, and the root mean squared error (RMSE) averaged across all the 120 emulators were used as performance indicators to compare the different selected values for retained PCs and hyperparameters.

The two emulator configurations that performed best were selected as the final two optimised emulators. It was found that the optimised interglacial emulator retained 15 PCs (accounting for 90% of the total variance), and had length scales δ of 2.792 (ε), 1.310 ($e\sin\varpi$), 1.664 ($e\cos\varpi$), 0.523 (CO_2), and 10.000 (GSL), and a nugget of 0.000. The optimised glacial emulator retained 15 PCs (accounting for 81% of the total variance), and had length scales δ of 6.908 (ε), 7.499 ($e\sin\varpi$), 5.460 ($e\cos\varpi$), 1.003 (CO_2), and 0.290 (GSL), and a nugget of 0.050.

The results of the evaluation of the emulators are shown in Figure S1. The results suggest that the emulators perform relatively well, similar to the results from Lord *et al*. (2017), the emulators in which did not include glacial GCM simulations. The calibration and evaluation shows that the emulators are able to reproduce the left out ensemble simulations reasonably well, with no obvious systematic errors in their predictions. These emulator configurations were optimised on SAT data, and these same configurations (i.e. same number of retained PCs and hyperparameter values) were then also applied to emulate the other variables. This approach ensures that the results for the variables are consistent with each other.

The SAT anomaly (compared to PI) at the LGM, predicted using the emulator is shown in Figure S2b, and compared to that produced directly by HadCM3 (Figure S2a) and to the SAT anomaly reconstruction of Annan and Hargreaves (2013) (Figure S2c). The latter reconstruction was produced by combining a range of proxy records of climate from different parts of the globe with an ensemble of simulations performed with various climate models as part of the PMIP2 project (Braconnot *et al*. 2007). It can be seen that the large-scale features of SAT are similar between the three projections, with the most significant regions of cooling located over the Laurentide ice sheet in North America, and to a lesser extent the Fennoscandian ice sheet, in line with current understanding. There are, however, some discrepancies between the two modelled (HadCM3 and emulator) projections and the Annan and Hargreaves (2013) reconstruction. Discrepancies include the extent of cooling over the Laurentide and Fennoscandian ice sheets being larger in the modelled projections than in the reconstruction, and high latitudes (Arctic and Antarctic regions) also being slightly colder in these simulations compared to that of Annan and Hargreaves (2013). This enhanced cooling observed in the emulator at this time is highlighted by a global mean cooling value of 5.1°C, more than one degree cooler than the 4°C estimated by Annan and Hargreaves (2013). However, the temperatures over ice sheets in the Annan and Hargreaves (2013) reconstruction are uncertain, and are largely based on model results rather than direct proxies. If this reconstruction is taken to be correct, the emulated SAT projections during glacial conditions could therefore be considered as being somewhat cold-biased at high latitudes.

TABLES

Ensemble	**Number of simulations**	**Input parameters**			
		Orbital		**Atmospheric CO_2 (ppmv)**	**Ice sheets**
modice	60	ε	22.2 : 24.4	250 : 1901	Modern
		$E\sin\varpi$	-0.055 : 0.055		
		$E\cos\varpi$	-0.055 : 0.055		
lowice	60	ε	22.2 : 24.4	250 : 1901	PRISM4 Pliocene
		$E\sin\varpi$	-0.055 : 0.055		
		$E\cos\varpi$	-0.055 : 0.055		
LGC	62	LGC		LGC	ICE-5G LGC

Table S1. Input parameter set-ups for the modice, lowice and LGC (120-0 kyr BP) GCM ensembles, including sampling ranges.

		Sensitivity simulations using real and constant values					
		Real data	**Average**	**PI**	**First**	**LGM**	**Random**
Proxy sites	**Dome C**	**539.4**	**0.65**	**0.002**	**0.14**	**0.48**	**-0.04**
	ODP 982	**38.41**	**0.03**	**0.04**	**0.04**	**0.05**	**-0.05**
	ODP 722	**54.04**	**-0.007**	**0.002**	**0.008**	**0.08**	**0.33**
	ODP 1143	**443.8**	**0.8**	**0.001**	**0.54**	**0.35**	**0.01**
	ODP 846	**74.95**	**0.12**	**0.001**	**0.16**	**-0.03**	**0.11**
	ODP 1090	**9.19**	**0.003**	**0.005**	**0.005**	**0.005**	**-0.04**

Table S2. M scores between emulated and proxy data, for E_{11111} (at the same sites) over the last 800 kya, where the Real data column shows the M scores between the actual emulated and proxy data (i.e. the same as the diagonals in Table 1), and the other columns are the M scores between the emulated data and constant values in the proxy data.

Simulation	Proxy sites					
	Dome C	**ODP 982**	**ODP 722**	**ODP 1143**	**ODP 846**	**ODP 1090**
E00000	0.344583	0.071945	0.264328	0.403474	0.047963	0.024133
E00001	325.261	27.7248	63.6199	167.498	10.6718	1.74508
E00010	4.03982	2.90694	16.0747	8.07642	1.97647	0.173069
E00011	328.355	28.7124	66.703	169.715	10.3691	1.74599
E00100	3.77609	0.548882	11.6131	13.147	2.67353	-0.0894
E00101	326.23	28.4583	58.2413	166.702	10.6164	1.69666
E00110	7.62055	3.02833	24.5126	20.4247	4.56978	0.024247
E00111	329.283	29.4227	60.9621	168.94	10.3232	1.69328
E01000	38.7483	-0.27807	-4.11324	-2.14263	0.027017	0.288044
E01001	383.839	26.8872	62.7225	166.181	11.0075	2.1857
E01010	41.9192	2.42696	11.3616	5.34173	1.80154	0.436297
E01011	386.796	27.8879	65.8405	168.536	10.6925	2.1887
E01100	41.545	1.44793	7.15014	9.6323	2.21534	0.082721
E01101	384.695	27.4674	57.523	165.488	10.9575	2.13155
E01110	43.6772	3.88329	19.2364	16.6587	3.98593	0.179558
E01111	387.557	28.432	60.2619	167.853	10.6527	2.13091
E10000	385.871	18.1169	54.8322	221.551	41.6643	7.04518
E10001	507.61	37.4965	56.0221	444.006	74.776	8.70527
E10010	392.119	20.3506	60.5384	227.997	43.1684	7.20301
E10011	512.302	38.1194	57.4068	444.824	74.1388	8.69715
E10100	380.992	17.8834	48.4845	225.739	40.7859	6.80144
E10101	515.739	38.3551	52.7542	443.517	74.5436	8.65662
E10110	387.294	19.141	53.7477	232.299	42.3356	6.96029
E10111	520.419	38.973	54.0126	444.408	73.8976	8.64482
E11000	447.936	17.7805	52.4718	212.647	40.7789	7.34341
E11001	526.012	37.0082	55.9842	443.182	75.8682	9.25034
E11010	454.152	19.9811	58.2291	218.914	42.2258	7.51479
E11011	531.088	37.6387	57.3765	444.125	75.1832	9.24499
E11100	442.738	18.2586	46.7905	216.161	39.6713	7.01987
E11101	534.343	37.7883	52.7782	442.807	75.6474	9.19732
E11110	448.031	19.612	51.9976	222.615	41.1935	7.17428

E11111	539.413	38.406	54.0366	443.81	74.9473	9.18878

Table S3. M scores between emulated and proxy data (at the same sites) over the last 800 kya, from all 32 simulations.

FIGURES

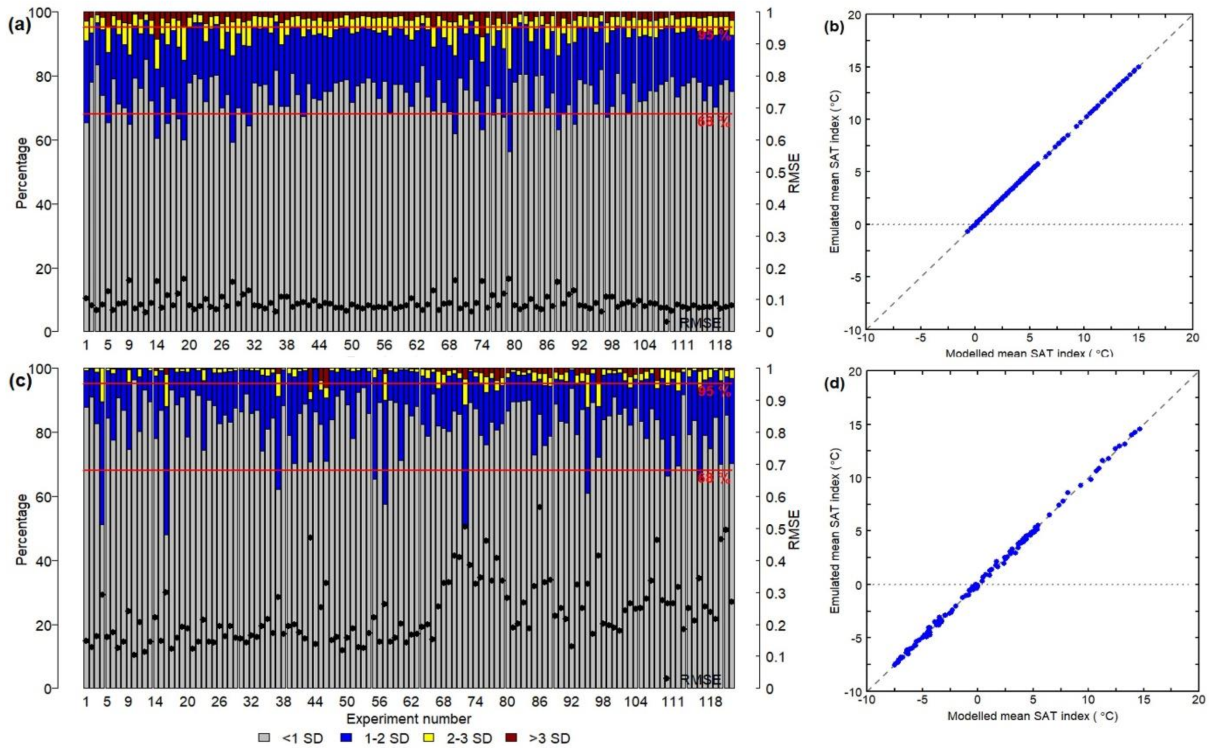


Figure S1. Evaluation of emulator performance for the interglacial emulator (top panel) and glacial emulator (bottom panel), both calibrated on SAT data. (a)+(c) Bars give the percentage of grid boxes for which the emulator predicts the SAT of the left-out experiment to within 1, 2, 3 and more than 3 SD. Also shown is the RMSE for the experiments (black circles). Red lines indicate 68 and 95 %. (b)+(d) Mean annual SAT index (°C) calculated by the emulator and the GCM. The 1:1 line (dashed) is included for reference. Note: this is the mean value for the GCM output data grid assuming all grid boxes are of equal size, hence not taking into account variations in grid box area. SAT is shown as an anomaly compared with the pre-industrial control simulation.

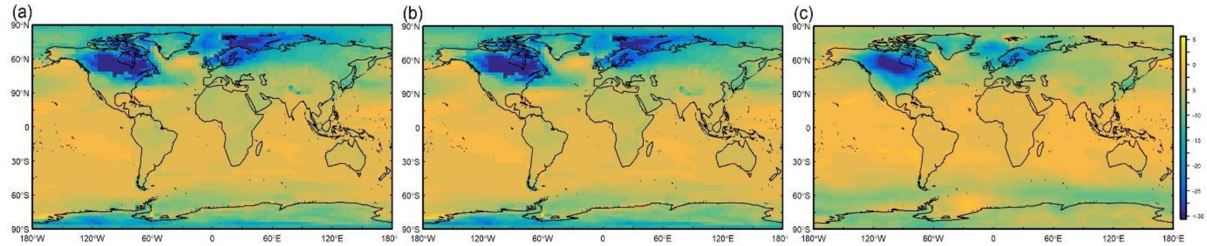


Figure S2. Maps of SAT anomaly (compared to pre-industrial; °C) at the Last Glacial Maximum (21 kyr BP) as projected by: a) HadCM3; b) the emulator; c) Annan and Hargreaves (2013), using a combination of proxy climate data and multi-model regression.

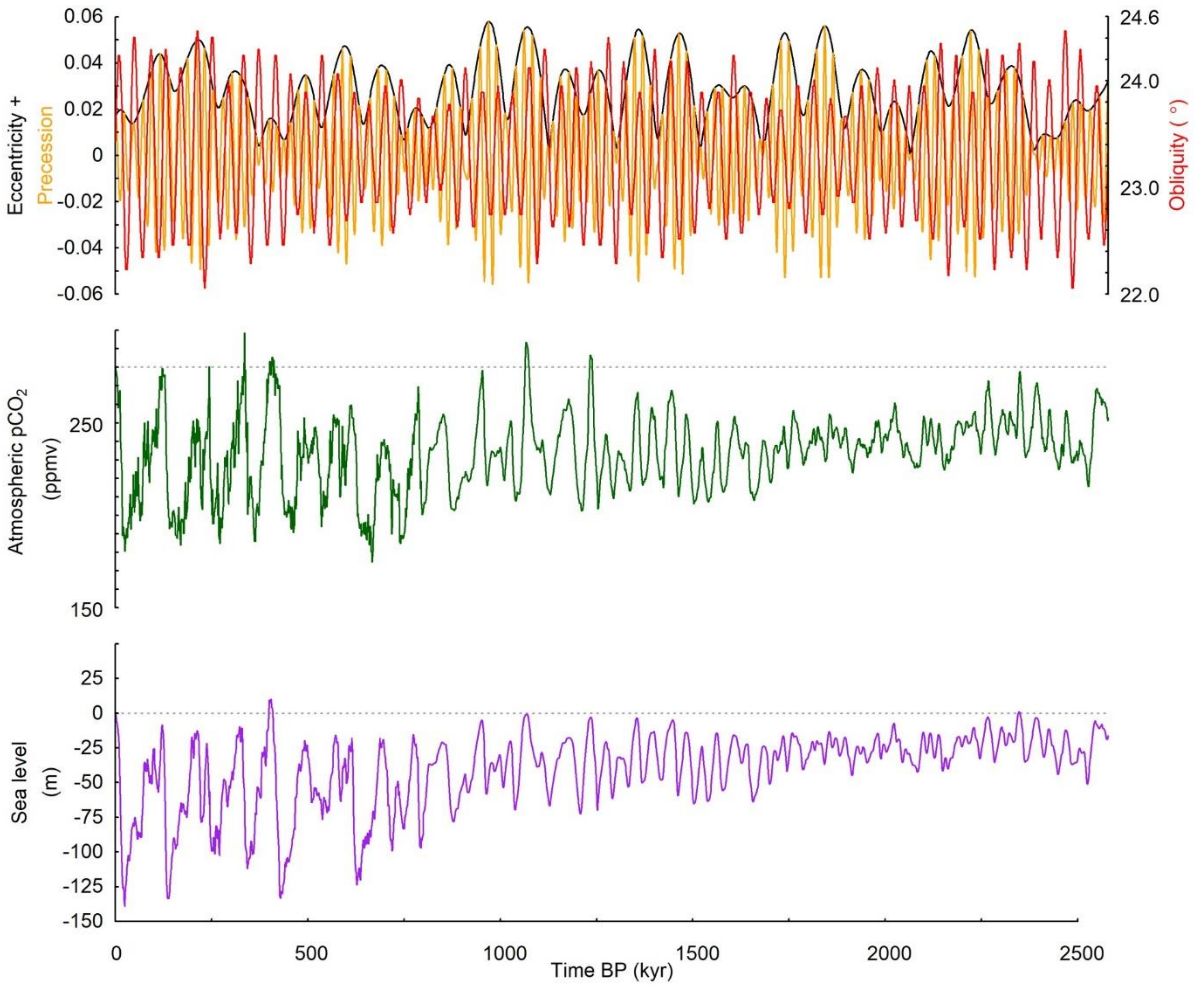


Figure S3. Climate forcing data used as input to the emulator for the last 2.58 Myr: a) Orbital variations (Laskar *et al*. 2004), showing eccentricity and precession on the left axis, and obliquity on the right axis; b) Atmospheric CO_2 concentrations, constructed from composite records from Antarctic ice cores (Bereiter *et al*. 2015) over the last 800 kya and based on a model-derived CO_2 signal for the remaining Pleistocene (van der Wal *et al*. 2011), where the PI CO_2 is also shown (grey dotted line); c) Reconstructed global sea level, derived from ocean sediment core $\delta^{18}O$ data (Spratt and Lisiecki 2016) over the last 800 kyr and model-derived sea level (De Boer *et al*. 2010) for the remaining Pleistocene, both of which are shown as an anomaly compared with PI (grey dotted line).

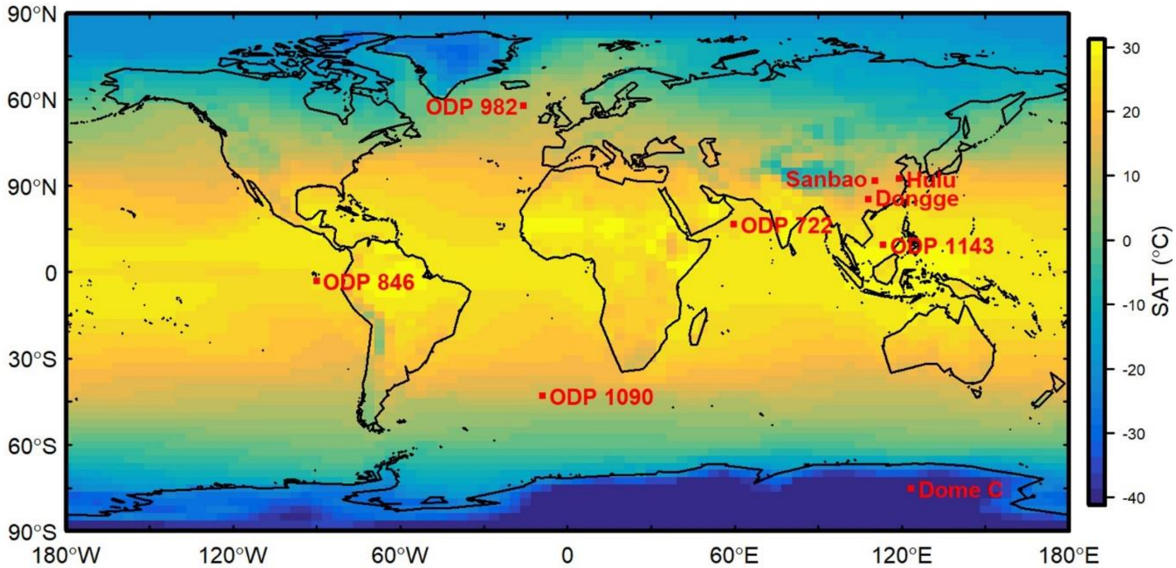
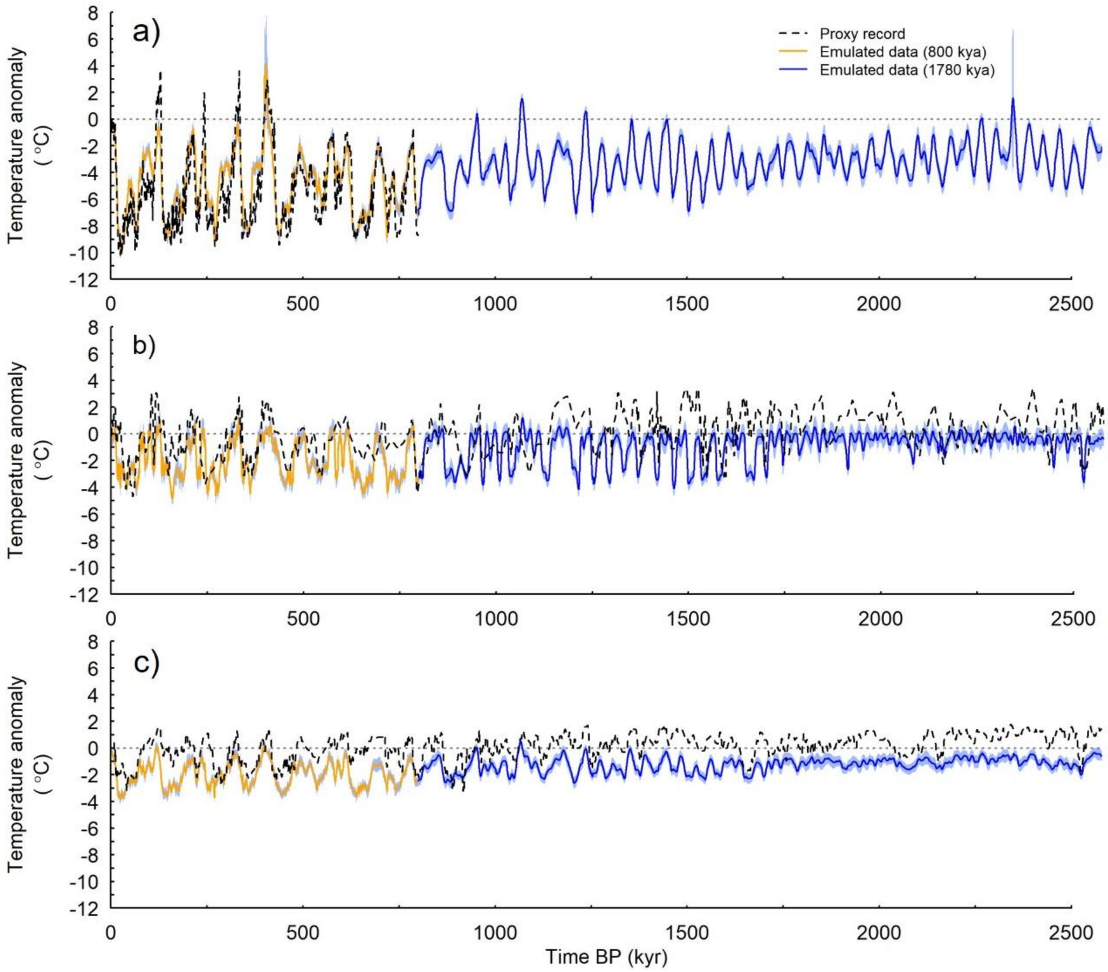
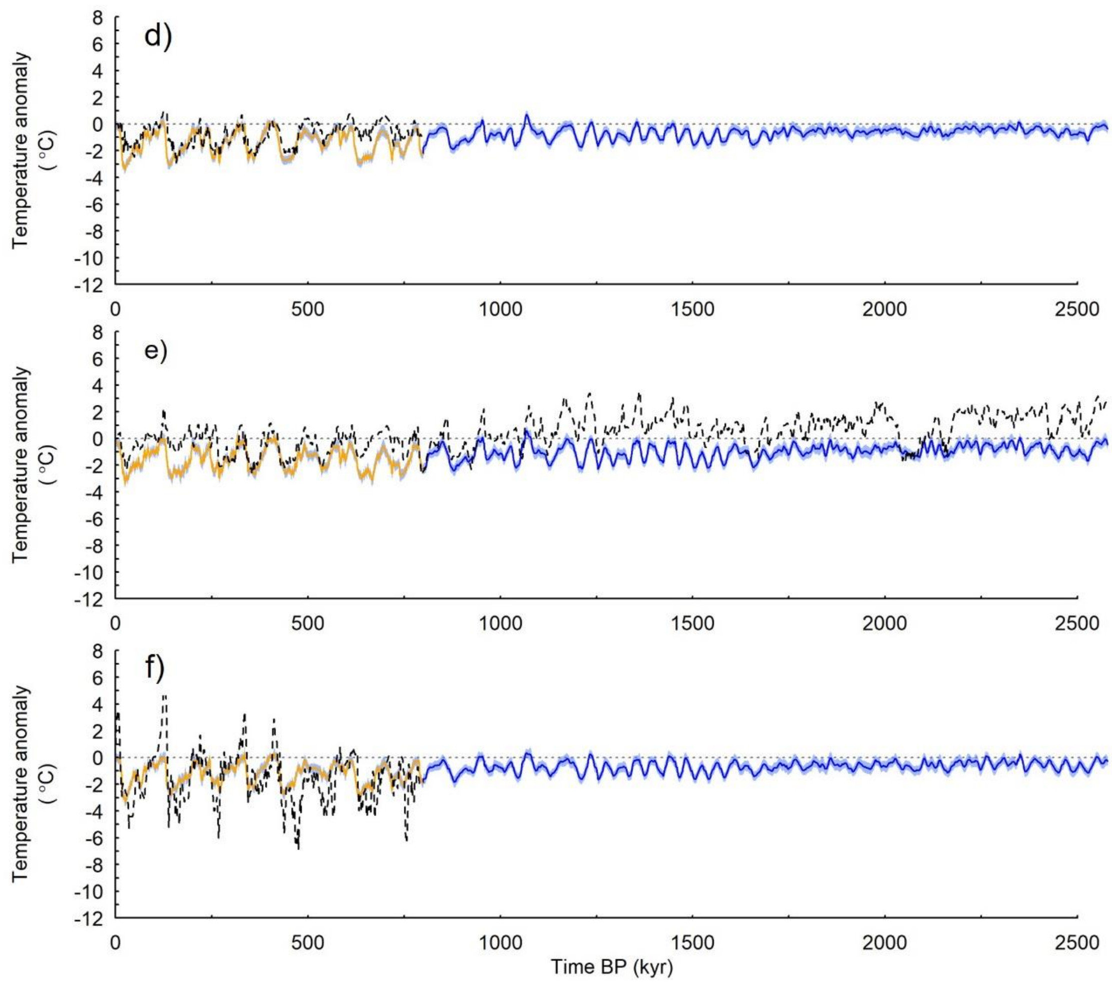


Figure S4. Map highlighting the source locations of the proxy data records. Temperature records are from the North Atlantic (ODP 982), the Arabian Sea (ODP 722), the South China Sea (ODP 1143), the Eastern Equatorial Pacific (ODP 846), the Subantarctic South Atlantic (ODP 1090), and Antarctica (Dome C). East Asian Monsoon records are from China, located in the east (Hulu Cave), centre (Sanbao Cave), and south (Dongge Cave) of the country. Background shading illustrates pre-industrial SAT (°C) taken from a HadCM3 control simulation.


a)
b)
c)
Proxy record
Emulated data (800 kya)
Emulated data (1780 kya)
Temperature anomaly (°C)
8
6
4
2
0
-2
-4
-6
-8
-10
-12
0
500
1000
1500
2000
2500
Time BP (kyr)

1047



1048

1049 **Figure S5**. Same as Figure 1, but using same y-axis.

1050

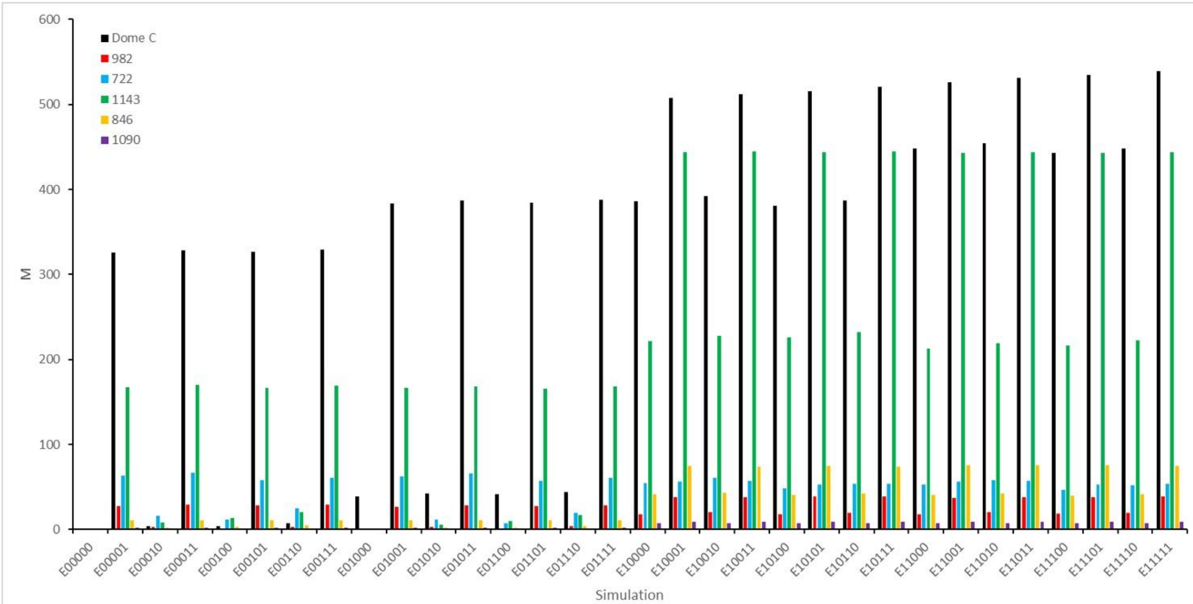


Figure S6. M scores between emulated and proxy data (at the same sites) over the last 800 kyr, from all 32 simulations.

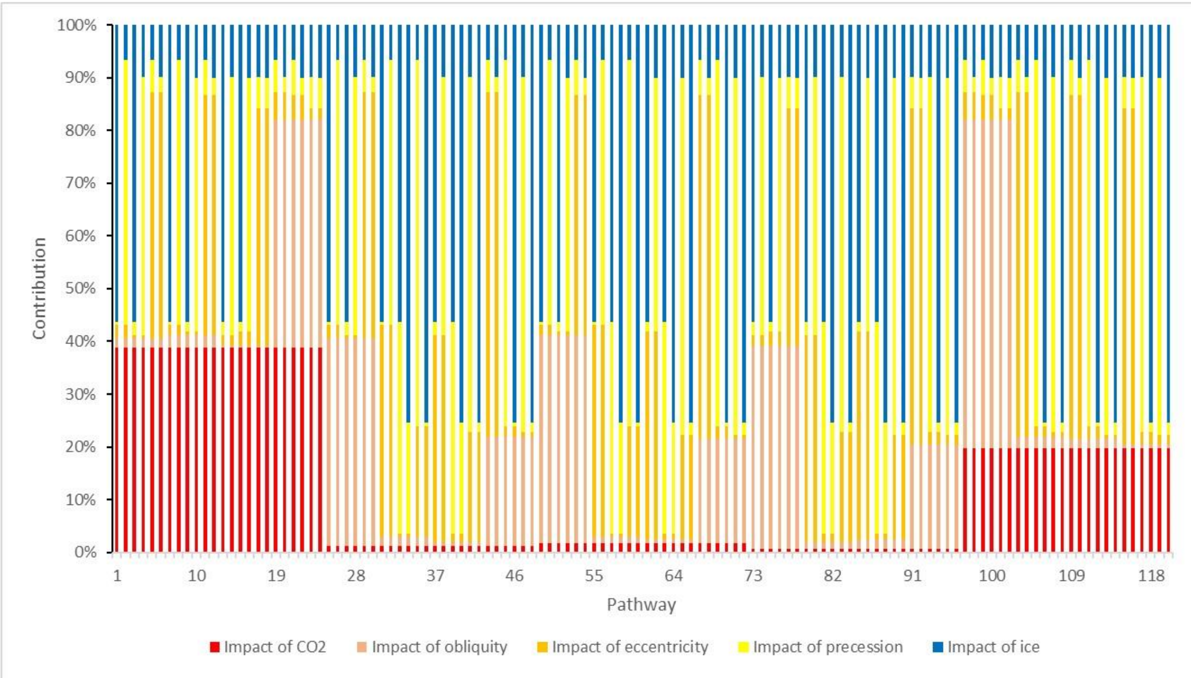


Figure S7. Global mean M scores for each driving component, expressed as a percentage, between emulated temperature from the E_{11111} simulation and that from individual simulations, following the linear-sum/shared interaction factorization of Lunt *et al*. (2021), for all 120 pathways individually.