PALEO-PGEM v1.0: A statistical emulator of Pliocene-Pleistocene climate

Philip B. Holden[1], Neil R. Edwards[1], Thiago F. Rangel[2], Elisa B. Pereira[2], Giang T. Tran[3] and Richard D. Wilkinson[4]

[1]Earth, Environment and Ecosystems, The Open University, Walton Hall, Milton Keynes MK7 6AA, UK
[2]Departmento de Ecologia, Universidade Federal de Goiás, CP 131, 74.001-970 Goiânia, Goiás, Brazil
[3]GEOMAR Helmholtz Centre for Ocean Research Kiel, Düsternbrooker Weg 20, 24105 Kiel, Germany
[4]School of Mathematics and Statistics, University of Sheffield, Sheffield, UK

Correspondence to: Phil Holden (Philip.holden@open.ac.uk)

Abstract. We describe the development of the "Paleoclimate PLASIM-GENIE emulator" PALEO-PGEM and its application to derive a downscaled high-resolution spatiotemporal description of the climate of the last five million years. The 5-million-year time frame is interesting for a range of paleo-environmental questions, not least because it encompasses the evolution of humans. However, the choice of time-frame was primarily pragmatic; tectonic changes can be neglected to first order, so that it is reasonable to consider climate forcing restricted to the Earth's orbital configuration, ice-sheet state and the concentration of atmosphere CO_2. The approach uses the Gaussian process emulation of the singular value decomposition of boundary-condition ensembles of the intermediate complexity atmosphere-ocean GCM PLASIM-GENIE. Spatial fields of bioclimatic variables of surface air temperature (warmest and coolest seasons) and precipitation (wettest and driest seasons) are emulated at 1,000 year intervals, driven by time-series of scalar boundary-condition forcing (CO_2, orbit and ice-volume), and assuming the climate is in quasi-equilibrium. Paleoclimate anomalies at climate model resolution are interpolated onto the observed modern climatology to produce a high-resolution spatiotemporal paleoclimate reconstruction of the Pliocene-Pleistocene.

1 Introduction

A high-resolution climate reconstruction of the Pliocene-Pleistocene will provide an unprecedented opportunity to advance understanding of many long-standing hypotheses about the origin and maintenance of biodiversity. Climate is among the strongest drivers of biodiversity and has played an important role throughout the history of life on Earth (Svenning et al 2015). Indeed, changes in climate over time have influenced core biological patterns and processes such as diversification, adaptation, species distribution and ecosystem functioning (Svenning et al 2015, Nogués-Bravo et al 2018). However, studies on the relationship between climate and biodiversity are still limited by the lack of high-resolution deep-time spatiotemporal paleoclimatic estimates, as the few studies available are at very sparse time slices (Lima-Ribeiro et al 2015). Thus, a high-resolution spatiotemporal paleoclimate data series of the past 5 million years will be useful to address many pressing questions on biodiversity dynamics. For instance, did the onset of glacial cycles promote more extinctions than recent climate cycles? Do species hold "evolutionary memory" of the warmer temperature of the Miocene? How did biodiversity respond to the increase in strength and frequency of glacial cycles during the Pliocene? Such knowledge is

essential to understand biodiversity patterns and to forecast how organisms will respond to the current anthropogenic climatic change (Nogués-Bravo et al 2018).

Spatio-temporal paleoclimatic estimates are essential to drive process-based models that are capable of exploring
45 causal mechanisms (Nogués-Bravo et al 2018). For instance, a recent ecological coupling study using climate emulation addressed the role of natural climate variability in shaping the evolution of species diversity in South America during the late Quaternary (Rangel et al 2018). That study used a paleoclimate emulator (Holden et al 2015) of the climate model PLASIM-ENTS (Holden et al 2014). The key limitations of the climate emulator were the lack of ocean dynamics in PLASIM-ENTS and the simplified emulation approach which only considered
50 orbital forcing; large-scale approximations were made to account for the effects of time-varying ice sheets and CO_2. Here we address these weaknesses by using boundary-condition ensembles of a fully-coupled Atmosphere-Ocean GCM. However, naïve simulation 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
55 not readily be parallelised. We overcome this intractability by approximating the climate state as being in quasi-equilibrium with the instantaneous forcing, and using statistical emulation.

Emulators are computationally fast statistical representations of process-driven simulators, most useful when application of the simulator would be computationally intractable (Sacks et al 1989, Santner et al 2003, O'Hagan
60 2006). Climate applications of emulation have included the exploration of multi-dimensional parameter input space in order to, for instance, generate probabilistic outputs (Sanso et al 2008, Rougier et al 2009, Harris et al 2013) or calibrate simulator inputs (Sham Bhat et al 2012, Olson et al 2012, Holden et al 2013). Climate emulators have also been developed as fast surrogates of the simulator for use in coupling applications (Castruccio et al 2014, Holden et al 2014). In addition to Rangel et al (2018), coupling applications have included climate change
65 impacts on energy demands (Labriet et al 2015, Warren et al 2018) and adaption to sea-level rise (Joshi et al 2016).

Our statistical methodology is Gaussian process (GP) emulation (Rasmussen 2004) of dimensionally-reduced representation of various climate fields. GP emulators are non-parametric regression models that have become
70 widely used tools in a variety of scientific domains. We train the emulators using boundary-condition ensembles of paleoclimate simulations, driven by variable orbital, CO_2 and ice-sheet forcing, in order to predict spatial fields of bioclimatic variables. This builds on previous studies that have emulated two-dimensional climate fields from CO_2 forcing (Holden and Edwards 2010, Holden et al 2014), orbital forcing (Bounceur et al 2015, Holden et al 2015), from combined CO_2 and ice-sheet forcing (Tran et al 2016) and from combined orbital and CO_2 forcing
75 (Lord et al 2017). Lord et al (2015) additionally considered two ice-sheets states (modern and a reduced Pliocene configuration) but, to our knowledge, these three Pliocene-Pleistocene forcings have not previously been varied together except in the emulation of scalar indices (Araya-Melo et al 2015). Ice-sheet forcing complicates the emulation problem because ice sheets are three-dimensional input fields. Although climate emulators with dimensionally reduced input and output fields have been developed (Holden et al 2015, Tran et al 2018), we

80   simplify the problem by assuming there is an approximate equivalence between the ice sheet state and global sea-level. This reduces the emulation to the more usual problem of relating scalar inputs to high-dimensional outputs.

The motivation for our approach is to generate spatiotemporal climate fields for use in dynamic coupling applications that need temporal variability and therefore cannot use snapshot AOGCM simulations. To this end,
85   we need forcing time series that extend back 5 million years and have sufficient temporal resolution to capture orbitally forced climate variability. For PALEO-PGEM v1.0 we use the sea-level reconstructions of Stap et al (2017) for the whole period and their CO_2 reconstruction prior to 800,000 BP (when ice core records are not available).

90   **2 The model PLASIM-GENIE**

PALEO-PGEM was built from quasi-equilibrium simulations of the intermediate complexity AOGCM PLASIM-GENIE (Holden et al 2016), a coupling of the spectral atmosphere model Planet Simulator (PLASIM, Fraedrich 2012) to the Grid-Enabled Integrated Earth system model (GENIE, Lenton et al 2006). The component modules, coupling and preindustrial climatology are described in detail in Holden et al (2016). PLASIM-GENIE is not flux
95   corrected. The moisture flux correction required in the Holden et al (2016) tuning was removed during a subsequent calibration (Holden et al 2018). PLASIM-GENIE has been applied to studies on Eocene climate (Keery et al 2018) and climate-carbon cycle uncertainties under strong mitigation (Holden et al 2018).

We applied PLASIM-GENIE at a spectral T21 atmospheric resolution (5.625 degrees) with 10 vertical layers, and
100 a matching ocean grid with 16 logarithmically spaced depth levels. We enabled the ocean BIOGEM (Ridgwell et al 2007) and terrestrial ENTS (Williamson et al 2006) carbon-cycle modules, as described in Holden et al (2018). We do not consider ocean biogeochemistry outputs here, but these are available should applications arise.

The 2000-year spun-up simulations required for emulation were performed with atmosphere-ocean gearing
105 enabled (Holden et al 2018). In geared mode, PLASIM-GENIE alternates between conventional coupling (for 1 year) and a fixed-atmosphere mode (for 9 years), reducing spin-up time by an order of magnitude, to roughly four days CPU.

3 Experimental overview
110 We first provide a summary of the entire approach in five steps, as illustrated schematically in Figure 1. Each step is described in more detail in Section 4.

i) Ensemble calibration: We previously developed a 69-member ensemble of plausible parameter sets using 'history matching' (see, e.g., Williamson et al 2013). Applying any of these parameter sets to PLASIM-GENIE
115 gives a reasonable climate-carbon cycle simulation of the present day, as evaluated by ten large scale metrics; all 69 parameter sets produce simulated outputs that lie within the ten history match acceptance ranges listed in Table 1. This step has been published elsewhere (Holden et al 2018).

ii) Model selection: We do not address parametric uncertainty in PALEO-PGEM, and so required a single

120 favoured PLASIM-GENIE parameter set. One of the 69 history matched parameter sets was identified by picking the parameter set whose simulator output had the largest likelihood (defined in Section 4.1) and this "optimised" parameter set was used in all subsequent simulations. We require PALEO-PGEM to describe glacial states and so, as part of the calibration, we performed an additional ensemble with the 69 parameter sets forced by Last Glacial Maximum (LGM) boundary conditions. The calibration considered simulated LGM cooling in addition

125 to the ten present day metrics (Table 1).

iii) Paleoemulator construction: PALEO-PGEM was constructed via a two-stage process, in both stages applying Gaussian process emulation to a singular value decomposition of the outputs of a PLASIM-GENIE simulation ensemble (c.f. Wilkinson 2010, Bounceur et al 2015, Holden et al 2015, Lord et al 2017). The first

130 stage emulated the simulated climate response to variable orbital and CO_2 forcing, while the second stage emulated the incremental climate anomaly due to the presence of glacial ice sheets. The motivation for this two-stage approach was to impose physical meaning on the decomposition by isolating the ice-sheet forced components from the orbital and CO_2 forced components. Note that we do not assume a linear superposition of the forcing components, and interactions between ice sheets, CO_2 and orbit are represented in the second stage

135 (see Section 4.2). All simulations used the optimised parameter set, and varied only the climate forcing.

iv) Paleoclimate emulation: Forcing time series of orbital parameters, atmospheric CO_2 concentration and sea-level (as a proxy for ice-sheet volume) were applied to the two-stage emulator to generate emulated climates at the native climate model resolution.

140

v) Downscaling. The emulated climates were converted to anomalies with respect to the emulated preindustrial state and interpolated onto a high-resolution grid. These interpolated anomalies were applied to the observed climatology to derive a high-resolution paleoclimate reconstruction at 1000 year intervals from 5MaBP.

145 **4 The simulation ensembles**

4.1 The optimised parameter set θ^*

Given computational constraints we chose to neglect parametric uncertainty in PALEO-PGEM, and selected a single 'optimised parameter set' for all simulations. Earlier work (Holden et al 2018) had developed a calibrated ensemble of 69 plausible PLASIM-GENIE parameter sets through a history matching approach. In summary,

150 these authors built and applied emulators of seven scalar metrics (items 1-7 in Table 1) to search for plausible input space. They considered hundreds of millions of potentially valid model parameterisations, each selected randomly by drawing from priors for 32 varied input parameters (Table 2). Each of these 32-element parameter vectors were applied to the seven emulators in turn and 200 of them were selected to maximize a criterion that combined the distance of candidate points to the other points already in the design (to ensure the design points

155 fully span the input space) and the probability (according to the emulator) of reasonably simulating the observational targets: global average surface air temperature, global vegetation carbon, global soil carbon, Atlantic overturning circulation strength, Pacific Ocean overturning circulation strength, global average dissolved ocean oxygen concentration and global average calcium carbonate flux to the ocean-floor. The 200 parameter sets

were applied to simulation ensembles of the preindustrial state and transient historical CO_2 emissions-forcing
160   (1805 to 2005). Finally, 69 of these parameter sets were selected as acceptable on the basis of the seven pre-
industrial metrics and three additional metrics that relate only to the transient simulations (items 8-10 in Table 1):
emissions-forced CO_2 concentration in 1870 and 2005, and transient warming (from 1865 to 2005).

In addition to these ten plausibility tests of Holden et al (2018), we also required the optimized model to exhibit
165   a reasonable response to glacial ice sheets. We therefore performed an additional 69-member PLASIM-GENIE
ensemble, applying Last Glacial Maximum forcing of 180ppm CO_2 concentration, 'ICE-5G' LGM ice sheets
(Peltier 2004) and the LGM orbital configuration of Berger (1989), with eccentricity 0.0019, obliquity 22.949°
and longitude of the perihelion at vernal equinox 114.4°.

170   For each of j=1, …, 69 parameter combinations, we calculate a score P_j which indicates how successful simulation
j was, in terms of matching the observations for each of the eleven metrics. These are tabulated in the "Calibration"
column of Table 1, where μ_i denotes the observational estimate for metric i and σ_i an estimate of uncertainty,
cognizant of both observational and model error.

175    $$P_j = \prod_{i=1,11} e^{-(g_i(\theta_j)-\mu_i)^2/2\sigma_i^2} \tag{1}$$

where $g_i(\theta_j)$ is is the output of the simulator corresponding to the ith metric when it is run at parameter setting
θ_j. The optimised parameter set θ^* was selected to be the ensemble member with the highest score, equivalent to
minimizing a weighted sum of squared errors. This optimised parameter set was used in all simulations that follow.
180   The optimized output metrics are provided in Table 1, and the input parameter values in Table 2. The most notable
bias is the cold LGM when compared to observational target, though the optimised model lies within the 3.1 to
5.9°C ranges simulated by the CMIP5/PMIP3 and PMIP2 ensembles (Masson Delmotte et al 2013).

4.2 Ensemble design

185   Our approach to emulating climate output fields relies on dimension reduction using the singular value
decomposition. This is a statistical technique which rotates the data onto a new orthogonal coordinate system, so
that the first coordinate is in the direction of maximum variance in the data, the second coordinate is then in the
direction of maximum variance conditional on being orthogonal to the first coordinate, etc. The new coordinates
are often called principal components (or empirical orthogonal functions), and whilst they are orthogonal, they
190   are not expected to cleanly isolate distinct physical processes. In order to impose a physical separation of the
components, and therefore to enforce a clean response to a distinct forcing, we chose to build the emulator as a
two-stage process. We first decomposed and emulated the smoothly varying climate response to changing orbit
and CO_2 concentration with fixed present-day ice sheets (the 'E1' emulator). We then separately emulated the
incremental climate response to a change in ice-sheet state under the same orbital and CO_2 forcing (the
195   'E2'emulator) so that the final emulation is the sum of these two components.

To build the E1 and E2 emulators, two separate 50-member boundary-condition ensembles were performed (BC1
and BC2) with the optimized parameter set. The statistical design of both ensembles was the same 5x50 maximin

200 latin hypercube (MLH,) varying the three orbital parameters, the CO_2 concentration and the ice sheet state. The only difference between the two ensembles was that the fifth hypercube variable, reserved for ice sheets, was ignored for the BC1 ensemble and the present-day ice-sheet configuration imposed for all BC1 simulations. The BC1 ensemble is designed to simulate the model response to orbit and CO_2 forcing only, while the BC2 ensemble simulates the different response driven by the presence of glacial ice sheets under the same set of choices of orbital and CO_2 forcing.

205

The sampling strategy for the orbital variables (eccentricity e, the longitude of the perihelion at the vernal equinox ω and obliquity ε) followed Araya-Melo et al (2015), uniformly sampling $e \sin \omega$ and $e \cos \omega$ in the range -0.05 to 0.05 and ε in the range 22° to 25°. This transformation was chosen because the insolation at any point in space and time of year is generally well approximated as a linear combination of these terms. Carbon dioxide was varied

210 uniformly in log space, in the range log(160 ppm) to log(1000 ppm). For ice sheets, relevant only to the BC2 ensemble, four states were allowed in the training ensemble, being the Peltier Ice-5G ice sheets (Peltier 2004) at 10, 13 15 and 20ka. These times were chosen as they correspond to well-spaced ice-volume intervals as evidenced by benthic $\delta^{18}O$ (Lisiecki and Raymo 2007). These times correspond to sea-level falls of 29, 45, 64 and 107m relative to modern in the Stap et al (2017) reconstruction that we use to force the time series emulation (Section

215 6).

In contrast to Araya-Melo et al (2015), we did not restrict input space to exclude combinations of high CO_2 and high glaciation levels, preferring instead to use all BC1 ensemble members (i.e. including those with high CO_2) in the BC2 ice sheet anomaly ensemble. This maintained the maximin and orthogonal properties of the MLH

220 design, and moreover avoided any risk of extrapolation outside of training input space during the Pliocene. Present day (~400ppm) CO_2 levels can be associated with significant (~50m) sea-level falls according to the Stap et al (2017) reconstructions (see Figure 2). However, the trade-off for this simplicity is that realistic input space during glacial periods was less well sampled than it would be for a more targeted ensemble of the same size (c.f. Araya-Melo et al 2015).

225

5 Emulator construction

Emulators were built for four bioclimatic variables: the mean temperature of the warmest and coolest quarters and the mean daily precipitation of the wettest and driest quarters. Each variable was calculated on a grid-point basis as the maximum and minimum of the DJF, MAM, JJA and SON seasons. These emulated variables were chosen

230 as being of bioclimatic relevance (c.f. Rangel et al 2018), and suitable for a wide range of ecological and impact coupling applications, defining the extremes of climate experienced over each grid-cell during a (decadally-averaged) annual cycle. Emulators of DJF and JJA temperature and precipitation were also built for validation purposes (Section 6.1).

235 We derived emulators from inputs of $e \sin \omega$, $e \cos \omega$, ε, $\log(CO_2)$ and sea level S, each normalized on the range -1 to 1. Sea level provides a proxy for ice-sheet volume, and hence ice-sheet state (under the assumption of an invariant correspondence between ice-sheets and sea level). This neglects the asymmetry of ice sheets under glaciation and deglaciation. The E1 emulator was built from the outputs of the BC1 ensemble (after centering the

data, by subtracting the ensemble mean field M from each simulation before singular value decomposition). The
E2 emulator was built from the anomaly outputs BC2-BC1. For E2, we appended the training data with a synthetic
50-member ensemble with the hypercube inputs repeated except that sea level was randomly assigned to be
between -25m and +100m. In these synthetic data, no simulations were performed, but instead all the climate
anomalies were set to zero, equivalent to performing a second ice-sheet forced ensemble with a present-day ice
sheets (and therefore with no anomaly by construction). This was needed so that the ice-sheet anomaly emulator
can be used when glacial ice sheets are absent (i.e. sea level greater that -25m) i.e. when the ice-sheet emulated
anomaly (E2) is trained to be zero and the emulation is determined only the orbit and CO_2 emulator (E1). Note
that this approach neglected the loss of Antarctic and/or Greenland ice compared to modern that is implicit when
paleo sea level exceeded the present day.

All emulators were built following the "one-step emulator" algorithm described by Holden *et al.* (2015),
summarized briefly here. For each ensemble member, we formed the 2048-element vector which describes the 64
× 32 output field to be emulated. The vectors for the N ensemble members were combined into a (2048 × N)
matrix Y describing the entire ensemble output of that variable. The matrices Y used to train the E1 emulators
comprised decadal-averaged outputs of the BC1 ensemble, and these matrixes were centered by subtracting the
ensemble mean field. The matrices for the E2 emulators were constructed from the decadal-averaged anomalies
BC2-BC1. This separation of the forcing elements is a key difference with earlier work; every BC1 member has
an identical BC2 member with the same inputs except for the incremental ice-sheet forcing, which cleanly isolates
the emulation of ice-sheet forcing from the orbital and CO_2 forcing.

Singular value decomposition was performed to reduce the dimensionality of the simulation fields:

$$Y = LDR^T \qquad\qquad (2)$$

where L is the (2048×N) matrix of left singular vectors ("components"), D is the N × N diagonal matrix of the
square roots of the eigenvalues and R is the N × N matrix of right singular vectors ("component scores"). This
decomposition produced a series of orthogonal components, ordered by the percentage of variance explained. We
truncated the decompositions, considering only the first ten components. Each of the ten retained sets of scores
thus comprised a vector of N coefficients, representing the projection of each simulation onto the respective
component. As each simulated field is a function of the input parameters, so are the coefficients that comprise the
scores, so that each component score can be emulated as a scalar function of the input parameters to the simulator.

We used Gaussian process (GP) emulation (Rasmussen 2004) in preference to stepwise linear regression. The
principal motivation for using this more sophisticated approach was that GPs are highly flexible non-parametric
regression models which have greater modelling power than linear models. An additional motivation was that GP
emulation provides both a central estimate and an estimate of uncertainty, and therefore provides us with a means
to generate uncertain climate emulations in the absence of parametric uncertainty. It is important to note that
emulator uncertainty is entirely distinct from (and therefore incremental to) parametric uncertainty.

Gaussian Process models are generalized models, but nevertheless require some user choices, the most important being the choice of covariance function. In order to evaluate the optimal covariance function, we considered the metric P, see Section 4.3.1 of Holden et al (2014):

$$P = \sum_{c=1,10} R_c^2 \, V_c \tag{3}$$

where R_c^2 is the cross-validated R-squared score of the emulator of principal component c, and V_c is the variance explained by that component, summed across the leading ten components. The metric is designed to quantify the percentage of the spatial variance explained by the emulator, capturing the unexplained variance due to principal component truncation (only ten components are considered) and to the emulation itself (i.e. unexplained variance of the simulated component scores under cross-validation).

Table 3 tabulates the cross-validated metric P for the eight emulators (i.e. four bioclimatic variables, two forcing categories) considering a series of different covariance functions, which are the alternatives available in the *DiceKriging* R package (Roustant et al 2012). The power exponential and exponential were found to give the best performance when averaged across the eight emulators. We chose the power exponential covariance function as the default as it consistently outperformed the exponential covariance function in the (more difficult) precipitation emulators.

6 Emulator application

The emulators generate a paleoclimate as

$$E(e, \omega, \varepsilon, CO_2, S) = M + E1(e, \omega, \varepsilon, CO_2) + E2(e, \omega, \varepsilon, CO_2, S) \tag{4}$$

where M is the simulation mean field that was subtracted to center the ensemble before decomposition (Section 5). To generate a paleoclimate time series, we therefore require time series of the boundary condition inputs $e, \omega, \varepsilon, CO_2$ and S.

For the orbital parameter inputs, we applied the 5 million-year calculation of Berger and Loutre (1999). We used CO_2 from Antarctic ice cores for the last 800,000 years (Luethi et al 2008). Prior to 800,000 BP, and for the entire sea-level record, we used the CO_2 and sea-level reconstructions of Stap et al (2017). These authors used a zonally averaged energy balance model coupled to a 6-level ocean model, a thermodynamic sea-ice model and to one-dimensional mass-balance modules for each of the five major Cenozoic ice sheets (East and West Antarctica, Greenland, Laurentide and Eurasian). The Stap model is forced with benthic $\delta^{18}O$ records, and uses an inversing routine to de-convolve the temperature and ice-volume components of the isotope signal and generate a self-consistent time series of CO_2 and sea-level (ice volume).

Figure 2 plots the forcing time series and an illustrative application of the emulator, for which we emulated time-varying annual mean surface air temperature field and plot its area-weighted global average through time.

6.1 Validation of emulation of spatial fields

320   In order to validate the emulators, we performed a series of ensemble experiments with Mid-Holocene (MH) and Last Glacial Maximum (LGM) CO_2, ice-sheets and orbital forcing. These time slices have been well-studied in Paleo-Modelling Inter-comparison Projects (PMIP) and are well suited to explore variability driven by all three forcings. The MH response is dominantly forced by orbit, while the LGM response is dominantly forced by CO_2 and ice-sheet state. To assist comparison with readily available PMIP2 data (Braconnot et al 2007), we here

325   emulate seasonal (DFJ and JJA) fields rather than seasonal (MAX and MIN) fields.

In each of the comparisons (Figures 3 and 4) we consider emulators with two and ten principal components (for each of the E1 and E2 emulators). The appropriate number of components to include is a somewhat subjective choice, although it can be useful to optimize by maximizing the explained variance under cross-validation (c.f.

330   Holden et al 2015). Here, we did not attempt to optimize the number of components, but instead treat the truncation as a variable input, so that model complexity can be varied, thereby enabling different emulator realizations (Figures 3 and 4) which may be useful to introduce additional uncertainty in coupling applications.

A further source of uncertainty arises from the emulation of the component scores. Gaussian process emulation

335   quantifies this by providing a mean prediction and an estimate of the uncertainty associated with that prediction. We generated two 200-member emulation ensembles with respective (MH or LGM) forcing. The 200 ensemble members differ because we do not assume the mean prediction for the emulated component scores, but instead draw randomly from the posterior distributions. In figures 3 and 4 these ensembles are summarised with mean and standard deviation fields. (We note that for applications in which climate uncertainty is not addressed, it is

340   appropriate to use the mean predictions of principal component scores to generate the best estimate.)

Fig 3 top panels compare emulated MH surface temperature (anomalies relative to preindustrial) with the PMIP2 OAV (coupled atmosphere-ocean-vegetation) ensemble. In northern winter DJF, high latitude warming is apparent in the emulated ensemble mean, although of uncertain sign (variability > mean). Cooling is apparent

345   over all other land regions. In northern summer JJA, robust warming is apparent at mid to high latitudes, while changes of variable signs are apparent in the tropics, with cooling apparent over the Sahel, India and SE Asia. Each of these features is also found in the PMIP ensemble.

Fig 3 lower panels compare emulated MH precipitation with the PMIP2 OAV ensemble. In DJF, significant drying

350   is emulated over central and northwestern South America, southern Africa, eastern Asia and northern Australia, while wetter conditions are emulated over northeastern South America. In JJA the largest changes are seen as a strengthening of the Asian monsoon precipitation, and significantly wetter conditions are also seen over the Sahel and western South America. These changes all reflect a general agreement with PMIP2.

355   Fig 4 upper panels compare the emulated Last Glacial Maximum temperatures with the PMIP2 OA (ocean-atmosphere) ensemble. We neglect the OAV LGM ensemble because it has only two simulations. LGM cooling is dominated by cooling of up to ~40°C over the northern hemisphere glacial ice sheets. The most significant differences are apparent in the emulated uncertainty, which is understated by a factor of roughly two relative to

PMIP. This is expected because the emulator is built from a single parameterization of PLASIM-GENIE and
360   therefore does not capture uncertain climate sensitivity. We note that by applying the principles of invariant
temperature pattern scaling (Tebaldi and Arblaster, 2014), the temperature uncertainties due to neglected climate
sensitivity could be approximated by inflating the variance of the principal component scores.

Fig 4 lower panels compare emulated Last Glacial Maximum precipitation with the PMIP2 OA ensemble. In DJF,
365   the drying apparent in central Africa, northern America and the Amazon are captured by the emulator, while JJA
drying at northern latitudes and in the Asian and African monsoon regions, and increased precipitation in South
America are common to the emulator and the PMIP2 ensemble. The most significant difference is the increase of
DJF precipitation emulated in central South America, which is not present in the PMIP ensemble mean, although
we note that the PMIP2 simulations display change of uncertain sign.

370

7 Downscaling

A spatial resolution higher than the native resolution of the underlying climate model may be required for paleo-
applications given the scale dependency of many patterns and processes (e.g. Rahbek 2005), such as scale-
dependent climate heterogeneity (Rangel et al 2018). We address this need by interpolating the low-resolution
375   climate model anomalies onto fine-resolution climatological data. This approach is widely-used in climate impact
assessment (e.g. Osborn et al 2016), and has also been applied in paleo-applications in anthropology (Melchionnaa
et al 2018) and ecology (Rangel et al 2018).

Downscaling can be performed in any given grid. Here we illustrate downscaling on a global hexagonal grid build
380   on a geodesic dome, because it minimizes geographic distortions in shape, area and distance that are common to
map projections. The hexagonal grid is composed of 17,151 *quasi* equal-area cells of $6,918 \pm 859$ km^2 whose area
variation is not spatially structured.

The four present-day (preindustrial) emulated bioclimatic variables $E0$ were linearly interpolated onto the
385   geodesic grid. All emulations used the mean prediction and the E1 and E2 emulators were both truncated at ten
principal components. Contemporary observations of the bioclimatic variables $C0$ were derived from WorldClim
(Hijmans et al 2005), which provides temperature and precipitation estimates at 1 km^2 resolution, interpolated
from temporally averaged measurements (1950 to 2000) from ~15,000-50,000 weather stations globally
(depending upon the variable). The raw emulated climate data $E0$ and the difference with observed climatology
390    $E0 - C0$ are illustrated in figure 5

The emulated climatology is reasonable, accepting the low resolution of the underlying climate model. Cold biases
are generally confined to northern-winter high latitudes. Warm biases are more modest, and are most apparent in
the Tibetan Plateau and Andes; the lapse rate cooling in these narrow mountain chains is poorly resolved by the
395   climate model. Excess precipitation bias is mostly apparent in the (wet-season) monsoon regions. Deserts are
generally well resolved in the emulator, a notable exception being the hyper-arid Atacama, which is an orography-
driven feature that cannot be captured at low resolution. Conversely, orography-driven precipitation is understated
in the Tibetan plateau. Precipitation is also understated in the Sahel.

400   We apply anomaly adjustments to derive downscaled emulated climate fields through time Ct. This approach preserves the high-resolution spatial heterogeneity of climatology. In the case of temperature this is straightforward. Emulated anomalies $Et - E0$ are interpolated onto the hexagonal grid and applied additively, i.e. $Ct = C0 + (Et - E0)$. For precipitation, the situation is more complex. In arid regions that are not well captured by the emulator, a multiplicative anomaly approach is preferable $Ct = C0 \times (Et/E0)$, preserving hyper-arid

405   (topographically-forced) desert, and preventing unphysical negative precipitation when $Et - E0 < 0$. Conversely, in wet regions that are understated by the emulator, a multiplicative anomaly approach can create unphysically high precipitation, but an additive approach ensures a physically reasonable solution. A pragmatic solution to this is to apply an additive precipitation anomaly when $E0 < C0$, and a multiplicative precipitation anomaly when $E0 > C0$. This approach is well-behaved, noting that the additive and multiplicative anomalies are

410   equivalent when $E0 = C0$. Consider, when $E0 < C0$,

$$Ct = C0 + (Et - E0) > Et \tag{5}$$

and the additive anomaly partially compensates for the low bias in emulated climatological precipitation.

415   Conversely, when $E0 > C0$,

$$Ct = C0 \times (Et/E0) < Et \tag{6}$$

and the multiplicative anomaly partially compensates for the high bias in emulated climatological precipitation.

420

The present-day climatology and downscaled emulated LGM climate are illustrated in Figure 6. An animation of the entire 5,000,000-year reconstruction is provided as supplementary material.

8 Conclusions and summary

425   We have used dimensionally reduced emulators of the intermediate complexity AOGCM PLASIM-GENIE, downscaled onto high resolution observed climatology, to generate a high resolution transient climate reconstruction of the last 5 million years. The reconstruction substantially improves on a previous emulated reconstruction (Rangel et al 2018) in the following ways

430   i) The underlying climate model is a fully coupled AOGCM. Rangel et al (2018) used PLASIM-ENTS (Holden et al 2014) which has a slab ocean and therefore neglected ocean circulation feedbacks.
ii) The new simulation ensembles considered climate forcing by orbit, CO_2 and ice-sheets. Rangel et al (2018) considered only orbit forcing, with large scale adjustments to crudely approximate the effects of CO_2 and ice sheets.

435   iii) We use Gaussian process emulation. Rangel et al (2018) used linear regression emulation, which cannot capture complex (non-linear) relationships between inputs and outputs.

440 These improvements allow us to provide a global emulation; the previous emulation was inappropriate for northern hemisphere due to the crude approximation of the response to ice sheet forcing. Additionally, we were able to extend the emulation back to 5 million years; the previous emulation was limited by the length of an existing 800,000-year transient GENIE simulations (Holden et al 2010) for CO_2 and ice sheets forcing. Finally, the use of GP emulation allows uncertainty estimates that we show in Figures 3 and 4 can be used to provide a reasonable proxy for model error, neglected in our single-parameterisation boundary condition ensembles.

445 The limitations of the reconstruction arise from the underlying climate model (low resolution, intermediate complexity), the approximated boundary conditions (in particular the use of only five ice-sheet states), uncertainties in the forcing time series (especially for sea level and CO_2), the assumption of quasi-equilibrium (so that e.g. millennial variability is neglected) and the limitations of downscaling. We note that the emulations and associated uncertainty compare favorably to existing ensembles of simulations with higher complexity models

450 (Figures 3 and 4). We note further that reconstructing climates with different forcing time series is straightforward. Future improvements are anticipated by including a representation of changing topography. For instance, the Andes have uplifted by 25 to 40% of their 3,700m present day elevation over the last 5 million years (Gregory-Wodzicki 2000) and Himalayan uplift has been associated with intensification of the Asian monsoon about 3.6 to 2.6 Myr ago (Zhisheng et al 2001). Ensembles that address changing orography, land sea masks and ocean

455 gateways, will improve the simulated climate and allow the extension of the emulation further back in time, to periods in which it would be unreasonable to ignore tectonically driven change.

8 Code availability

The supplementary information contains the following

460 PALEO-PGEMv1.0_5M_1Ka.mp4   Animation of the four bioclimatic variables over 5Ma

PALEO-PGEMv1.0.R   R code to build and run the emulators.

R input files

ensemble.dat   ensemble input design for the BC1/BC2 ensembles

5000_1000_forcing.dat   time series forcing for 5Ma at 1kyr intervals

465 MH_forcing.dat   mid Holocene ensemble forcing

LGM_forcing.dat   Last Glacial Maximum forcing

area.dat   grid cell areas for area weighting

data subdirectories

data   outputs of the BC1 PLASIM-GENIE ensemble

470 icedata   outputs of the BC2 PLASIM-GENIE ensemble

supporting spreadsheets

ensemble   supporting calculations for the ensemble design

5000ka_forcing   supporting calculations for the time series forcing

475 PALEO-PGEMv1.0.R was saved with settings to emulate DJF temperature and produce a 5Ma time series using the GP mean prediction (no emulator uncertainty), ten principal components and a power exponential covariance function. Each of these settings can be changed as documented in the code. The code outputs the area-weighted

average to screen, and three data sets to file: emul.dat (the full spatiotemporal output), mean.dat and SD.dat (the mean and standard deviation of the emulated fields, most relevant when code is set to generate an ensemble e.g.

480   with MH or LGM forcing).

Author contributions

PBH, NRE and TFR developed the concept. PBH performed the PLASIM-GENIE simulations, using boundary conditions developed by GTT. PBH, RDW and NRE developed the GP emulators. TFR and EBP developed the

485   downscaling, with advice from PBH and NRE. PBH wrote the manuscript with contributions from all authors.

Competing interests

The authors declare that they have no conflict of interest.

490

References

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

Araya-Melo, P.A., Crucifix, M. and Bounceur, N: Global sensitivity analysis on the Indian monsoon during the

495   Pleistocene, Clim. Past 11, 45-61. doi:10.5194/cp-11-45-2015, 2015.

Berger, A. and Loutre, M.-F.: Parameters of the Earths orbit for the last 5 Million years in 1 kyr resolution. *PANGAEA*, https://doi.org/10.1594/PANGAEA.56040, 1999.

Bondeau, A., et al. Modelling the role of agriculture for the 20th century global terrestrial carbon balance. *Glob. Change Biol.* **13**, 679–706, 2007.

500   Bounceur, N., Crucifix, M., and Wilkinson, R. D.: Global sensitivity analysis of the climate-vegetation system to astronomical forcing: an emulator-based approach, Earth Syst. Dynam., 6, 205– 224, https://doi.org/10.5194/esd-6-205-2015, 2015.

Braconnot, P., et al., 2007: Results of PMIP2 coupled simulations of the mid-Holocene and Last Glacial Maximum - Part 1: experiments and large-scale features. Clim. Past, 3, 261–277.

505   Castruccio, S., McInerney, D.J., Stein, M.L., Liu Crouch, F., Jacob, R.L. and E. J. Moyer, E.J.L Statistical emulation of climate model projections based on precomputed GCM runs. Journal of Climate, 27, 1829–1844, 2014.

Feely, R.A. et al. Impact of anthropogenic CO_2 on the $CaCO_3$ system in the oceans *Science* **305**, 362-366, 2005.

Fraedrich, K.: A suite of user-friendly global climate models: hysteresis experiments, Eur. Phys. J. Plus, 127, 1–

510   9, https://doi.org/10.1140/epjp/i2012-12053-7, 2012.

Gregory-Wodzicki, K.M. Uplift history of the Central and Northern Andes: a review. GSA Bulletin, 7, 1091-1105, 2000.

Harris, G.R., Sexton, D.M., Booth, B.B., Collins, M., and Murphy, J.M.: Probabilistic projections of transient climate change, Climate Dynamics, 40, 2937–2972, 2013.

515   Hijmans, R.J., Cameron, S.E., Parra, J.L., Jones, P.G., Jarvis, A. Very high resolution interpolated climate surfaces for global land areas. *International Journal of Climatology* **25**, 1965-1978, 2005.

Holden, P.B. and Edwards, N.R.: Dimensionally reduced emulation of an AOGCM for application to integrated assessment modelling, GRL, **37**, L21707, doi:10.1029/2010GL045137, 2010.

520 Holden P.B., Edwards, N.R., Wolff, E.W., Lang, N.J., Singarayer, J.S., Valdes, P.J and Stocker, T.F.: Interhemispheric coupling, the West Antarctic Ice Sheet and warm Antarctic interglacials, Clim. Past 6, 431–443 doi: 10.5194/cp-6-431-2010, 2010.

Holden, P.B., Edwards, N.R., Gerten, D., and Schaphoff, S.: A model-based constraint on CO_2 fertilisation, Biogeosciences, 10, 339–355, 2013.

Holden. P.B., Edwards, N.R., Garthwaite, P.H., Fraedrich, F., Lunkeit, F., Kirk, E., Labriet, M., Kanudia, A. and

525 Babonneau, F.: PLASIM-ENTSem v1.0: a spatio-temporal emulator of future climate change for impact assessment, Geosci. Model Dev., 7, 433-451, doi:10.5194/gmd-7-433-2014, 2014.

Holden, P.B, Edwards, N.R., Garthwaite, P.H. and Wilkinson, R.D. Emulation and interpretation of high-dimensional climate outputs, *J. App. Stats.* **42**, 2038-2055, 2015.

Holden, P.B. et al. PLASIM–GENIE v1.0: a new intermediate complexity AOGCM *Geosci. Mod. Dev.* **9** 3347-

530 3361, 2016.

Holden, P.B., Edwards, N.R., Ridgwell, A., Wilkinson, R.D., Fraedrich, K., Lunkeit, F., Pollitt, H.E., Mercure, J.-F., Salas, P., Lam, A., Knobloch, F., Chewpreecha U., and J. E. Viñuales, J.E.: Climate-carbon cycle uncertainties and the Paris Agreement, Nature Climate Change, 8, 609-613, doi:10.1038/s41558-018-0197-7, 2018.

535 IPCC: Summary for Policymakers. In: *Climate Change 2013: The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change* Stocker, T.F., D. Qin, G.-K. Plattner, M. Tignor, S.K. Allen, J. Boschung, A. Nauels, Y. Xia, V. Bex and P.M. Midgley (eds.). Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA, 2013.

Joshi, S.R., Vielle, M., Babonneau, F., Edwards, N.R and Holden P.B.: Physical and Economic Consequences of

540 Sea-Level Rise: A Coupled GIS and CGE Analysis Under Uncertainties, *Environmental and Resource Economics*, 65, 813-839, 2016.

Jones, P. D., New, M., Parker, D. E., Martin, S., & Rigor, I. G. Surface air temperature and its changes over the past 150 years. *Rev. Geophys.* **37**, 173–199, 1999.

Kanzow, T. et al. Seasonal Variability of the Atlantic Meridional Overturning Circulation at 26.58∘N, *J. Clim.* **23**,

545 5678–5698, 2010.

Keeling, C.D. et al. Atmospheric CO_2 and $^{13}CO_2$ exchange with the terrestrial biosphere and oceans from 1978 to 2000: observations and carbon cycle implications in *A History of Atmospheric CO_2 and its effects on Plants, Animals, and Ecosystems* editors, Ehleringer, J.R., T. E. Cerling, M. D. Dearing, Springer Verlag, New York, 2005.

550 Keery, J.S., Holden, P.B. and Edwards, N.R.: Sensitivity of the Eocene climate to CO_2 and orbital variability, Climate of the Past, 14, 2018.

Konkright, M. E et al. World Ocean Atlas 2001: Objective Analysis, Data, Statistics and Figures, CD- ROM Documentation, National Oceanographic Data Center, Silver Spring, MD, 2002.

Labriet, M., Joshi, S.R., Babonneau F., Edwards, N.R., Holden, P.B., Kanudia, A., Loulou, R and Vielle, M.

555 Worldwide impacts of climate change on energy for heating and cooling, Mitig. Adapt. Strateg. Glob. Change 20, 1111-1136, doi10.1007/s11027-013-9522-7, 2015.

Lenton, T. M., Williamson, M. S., Edwards, N. R., Marsh, R., Price, A. R., Ridgwell, A. J., Shepherd, J. G., Cox, S. J., and The GENIE team: Millennial timescale carbon cycle and climate change in an efficient Earth system model, Clim. Dynam., 26, 687–711, 2006.

560 Lima-Ribeiro, M. S., Varela, S., González-Hernández, J., Oliveira, G. de, Diniz-Filho, J. A. F., and Terribile, L. C.: Ecoclimate: a database of climate data from multiple models for past, present, and future for macroecologists and biogeographers. Biodiversity Informatics, 10, 1–21, 2015.

Lisiecki, L. E. and Raymo, M. E.: A Pliocene-Pleistocene stack of 57 globally distributed benthic δ18O records, Paleoceanography, 20, PA1003, doi:10.1029/2004PA001071, 2005.

565 Lord, N.S., Crucifix, M., Lunt D.J., Thorne, M.C., Bounceur, N., Dowsett, H., O'Brien, C.L., and Ridgwell A.: Emulation of long-term changes in global climate: application to the late Pliocene and future *Clim. Past***, 13,** 1539–1571, https://doi.org/10.5194/cp-13-1539-2017, 2017.

Luethi, D., Le Floch, M., Bereiter, B., Bluner, T., Barnola, J.-M., Siegenthaler, U., Raynaud, D., Jouzel, J., Fischer, H., Kawamura, K., and Stocker, T. F.: High-resolution carbon dioxide concentration record 650,000–

570 800,000 years before present, Nature, 453, 379–382, 2008.

Masson-Delmotte V.M., et al 2013: Information from Paleoclimate Archives. In: *Climate Change 2013: The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change* [Stocker, T.F., D. Qin, G.-K. Plattner, M. Tignor, S.K. Allen, J. Boschung, A. Nauels, Y. Xia, V. Bex and P.M. Midgley (eds.)]. Cambridge University Press, Cambridge, United Kingdom and New

575 York, NY, USA.

Melchionna, M., Di Febbraro, M., Carotenuto, F., Rook, L, Mondanaro, A., Castiglione, S., Serio, C., Vero, V.A., Tesone, G., Piccolo, M., Diniz-Filho, J.A.F., and Raia, P. Fragmentation of Neanderthals' pre-extinction distribution by climate change, Palaeogeography, Paleoclimatology, Palaeoecology, 496, 146-154, 10.1016/j.palaeo.2018.01.031, 2018.

580 Nogués-Bravo, D., Rodríguez-Sánchez, F., Orsini, L., de Boer, E., Jansson, R., Morlon, H., Fordham, D.A., and Jackson, S.T.: Cracking the Code of Biodiversity Responses to Past Climate Change. Trends in Ecology & Evolution, 10.1016/j.tree.2018.07.005, 1–12, 2018.

Olson, R., Sriver, R., Goes, M., Urban, N.M., Matthews, H.D., Haran, M., and Keller, K.: A climate sensitivity estimate using Bayesian fusion of instrumental observations and an Earth System model, Journal of Geophysical

585 Research: Atmospheres (1984–2012), 117 (D4), 2012.

O'Hagan, A.: Bayesian analysis of computer code outputs: a tutorial. Reliability Engineering & System Safety, 91, 1290–1300, 2006.

Osborn, T.J., Wallace, C.J., Harris, I.C., Melvin, T.M. Pattern scaling using ClimGen: monthly-resolution future climate scenarios including changes in the variability of precipitation. Clim. Change 134, 353-369, 2016.

590 Peltier, W.: Global isostasy and the surface of the ice-age Earth: The ICE-5G (VM2) model and GRACE, Annu. Rev. Earth Pl. Sc. 32, 111-149, doi:10.1146/annurev.earth.32.082503.144359, 2004.

Rahbek, C. The role of spatial scale and the perception of large-scale species-richness patterns. Ecology Letters, 8, 224–239, 2005.

Rangel, T.F., Edwards, N.R., Holden, P.B., Diniz-Fiho, J.A.F., Gosling, W.D., Coelho, M.T.P., Cassemiro,

595 F.A.S., Rahbek, C. and Colwell, R.K.: Biogeographical cradles, museums, and graves in South America: Modelling the ecology and evolution of diversity, Science, 361, eaar5452, 2018.

Rasmussen, CE: *Gaussian processes in machine learning*, in *Advanced Lectures on Machine Learning*, O. Bousquet, U. von Luxburg, and G. Rätsch (Eds.) Springer, Berlin, Heidelberg, New York. 2004.

Ridgwell, A., Hargreaves, J. C., Edwards, N. R., Annan, J. D., Lenton, T. M., Marsh, R., Yool, A., and Watson, A.: Marine geochemical data assimilation in an efficient Earth System Model of global biogeochemical cycling, Biogeosciences, 4, 87–104, doi:10.5194/bg-4-87-2007, 2007.

Rougier, J.: Probabilistic inference for future climate using an ensemble of climate model evaluations, Climatic Change, 81, 247–264, 2007.

Rougier, J., Sexton, D.M., Murphy, J.M. and D. Stainforth, D.: Analyzing the climate sensitivity of the HadSM3 climate model using ensembles from different but related experiments. Journal of Climate, 22, 3540–3557, 2009.

Roustant, O., Ginsbourger, D., Deville, Y. DiceKriging, DiceOptim: Two R Packages for the Analysis of Computer Experiments by Kriging-Based Metamodeling and Optimization. Journal of Statistical Software, 1-55. URL http://www.jstatsoft.org/v51/i01/, 2012.

Rubino, M., et al.: A revised 1000 year atmospheric δ^{13}C-CO_2 record from Law Dome and South Pole, Antarctica, J. Geophys. Res. Atmos., 118, 8482–8499, doi:10.1002/jgrd.50668, 2013.

Sacks, J. Welch, W.J., Mitchell, T.J. and Wynn, H.P.: Design and analysis of computer experiments. Stat. Sci., 4, 409–423, 1989.

Sansó, B, Forest, C.E., Zantedeschi, D., et al.: Inferring climate system properties using a computer model. Bayesian Analysis, 3, 1–37, 2008.

Santner, T.J., Williams, B.J. and Notz, W.I.: The design and analysis of computer experiments. Springer Verlag, 2003.

Sham Bhat, K., Haran, M., Olson, R. and Keller, K.: Inferring likelihoods and climate system characteristics from climate models and multiple tracers, Environmetrics, 23, 345–362, 2012.

Stap, L.B. van de Wal R.S.W., deBoer, B., Bintanja, R. and Lourens, L.J.: The influence of ice sheets on temperature during the past 38 million years inferred from a one-dimensional ice sheet-climate model, Clim. Past, 13, 1243-1257, doi:10.5194/cp-13-1243-2017, 2017.

Svenning, J.C., Eiserhardt, W.L., Normand, S., Ordonez, A., and Sandel, B.: The Influence of Paleoclimate on Present-Day Patterns in Biodiversity and Ecosystems. Annual Review of Ecology, Evolution, and Systematics, 46, 551–572, 2015.

Tebaldi, C. and Arblaster, J.M., Pattern scaling: Its strengths and limitations, and an update on the latest model simulations, Clim. Change 122, 459–471, 2014.

Tran, G.T., Oliver, K.I.C., Sobester, A., Toal, D.J.J., Holden, P.B., Marsh, R., Challenor, P. and Edwards, N.R. Building a traceable climate model hierarchy with multi-level emulators Adv. Stat. Clim. Meteorol. Oceanogr., 2, 17–37, doi:10.5194/ascmo-2-17-2016, 2016.

Tran, G.T., Oliver, K.I.C., Holden, P.B., Edwards, N.R., and Challenor, P.: Multi-level emulation of complex climate model responses to boundary forcing data, Climate Dynamics, doi:10.1007/s00382-018-4205-4, 2018.

Warren, R.F., Edwards, N.R., Babonneau, F., Bacon, P.M., Dietrich, J.P., Ford, R.W., Garthwaite, P., Gerten, D., Goswami, S., Haurie, A., Hiscock, K., Holden, P.B., Hyde, M.R., Joshi, S.R., Kanudia, A., Labriet, M., Leimbach, M., Oyebamiji, O.K., Osborn, T., Pizzileo, B., Popp, A., Price, J. Riley, G., Schaphoff, S., Slavin, P., Vielle, M., and Wallace, C.: Producing Policy-relevant Science by Enhancing Robustness and Model Integration for the

Assessment of Global Environmental Change, Environmental Modelling and Software, doi:10.1016/j.envsoft.2018.05.010, 2016.

Wilkinson, R. D.: Bayesian calibration of expensive multivariate computer experiments, in: Large scale inverse problems and the quantification of uncertainty, edited by: Biegler, L. T., Biros, G., Ghattas, O., Heinkenschloss, M., Keyes, D., Mallick, B. K., Tenorio, L., Van Bloemen Waanders, B., and Wilcox, K., Wiley Series in Computational Statistics, 2010.

Williamson, M. S., Lenton, T. M., Shepherd, J. G., and Ed- wards, N. R.: An efficient numerical terrestrial scheme (ENTS) for Earth system modelling, Ecol. Model., 198, 362–374, doi:10.1016/j.ecolmodel.2006.05.027, 2006.

Williamson, D., Goldstein, M., Allison, L., Blaker, A. Challenor, P, Jackson, L and Yamazaki, K.: History matching for exploring and reducing climate model parameter space using observations and a large perturbed physics ensemble. *Clim. Dynam.* **41**, 1703-1729, doi:10.1007/s00382-013-1896-4, 2013.

Zhisheng, A., Kutzbach, J.E., Prell, W.L. and Porter, S.C. Evolution of Asian monsoons and phased uplift of the Himalaya–Tibetan plateau since Late Miocene times, Nature, 411, 62-66, 2001.

i	Output metric	Observations	History matching acceptance range	ML calibration (mean, 1sigma) $\mu_i \pm \sigma_i$	Optimized simulation $g_i(\theta^*)$
1	Global average surface air temperature (°C)	~14 Jones et al (1990)	11 to 17	14 ± 1.5	14.1
2	Global vegetation carbon (GtC)	450 to 650 Bondeau et al (2007)	300 to 800	550 ± 125	696
3	Global soil carbon (GtC)	850 to 2400 Bondeau et al (2007)	750 to 2500	1625 ± 437.5	1170
4	Maximum Atlantic Overturning (Sv)	~19 Kanzow et al (2010)	10 to 30	20 ± 5	17.8
5	Maximum Pacific Overturning (Sv)		<15	0 ± 7.5	2.4
6	Global ocean averaged dissolved O_2 (µmol kg^{-1})	~170 Konkright et al (2002)	130 to 210	170 ± 20	139
7	Global deep ocean $CaCO_3$ flux (GT $CaCO_3$-C yr^{-1})	~0.4 Feely et at (2004)	0.2 to 0.8	0.4 ± 0.15	0.56
8	Atmospheric CO_2 in 1870 (ppm)	288 Rubino et al (2013)	N/A	288 ± 12.5	280
9	Atmospheric CO_2 in 2005 (ppm)	378 Keeling et al (2005)	353 to 403	378 ± 12.5	380
10	(1864-1875) to (1994-2005) warming (°C)	~0.78 IPCC 2013 SPM	0.6 to 1.0	0.78 ± 0.1	0.78
11	Last Glacial Maximum temperature change (°C)	4.0 ± 0.8 Annan and Hargreaves (2013)	N/A	-4.0 ± 1.2	-5.9

650

Table 1: Simulation output metrics for history matching and maximum likelihood calibration

Module	Parameter	Description	Units	Min	Max	Prior	Optimised θ^*
PLASIM	TDISSD	Horizontal diffusivity of divergence	days	0.01	10	LOG	0.01245
	TDISSZ	Horizontal diffusivity of vorticity	days	0.01	10	LOG	0.04627
	TDISST	Horizontal diffusivity of temperature	days	0.01	10	LOG	1.03202
	TDISSQ	Horizontal diffusivity of moisture	days	0.01	10	LOG	0.06188
	VDIFF	Vertical diffusivity	m	10	1000	LOG	12.9576
	TWSR1	Short wave clouds (visible)		0.01	0.5	LOG	0.32403
	TWSR2	Short wave clouds (infrared)		0.01	0.5	LOG	0.03297
	ACLLWR	Long wave clouds	$m^{-2}g^{-1}$	0.01	5	LOG	0.50152
	TH2OC	Long wave water vapour		0.01	0.1	LOG	0.02357
	RCRITMIN	Minimum relative critical humidity		0.7	1.0	LIN	0.94867
	GAMMA	Evaporation of precipitation		0.001	0.05	LOG	0.00799
	ALBSM	Equator-pole ocean albedo difference		0.2	0.6	LIN	0.44992
	ALBIS[1]	Ice sheet albedo		0.8	0.9	LIN	0.8
	APM[2]	Atlantic-Pacific moisture flux adjustment	Sv	0.0	0.32	LIN	0.0
GOLDSTEIN	OHD	Isopycnal diffusivity	m^2s^{-1}	500	5000	LOG	2005.24
	OVD	Reference diapycnal diffusivity	m^2s^{-1}	2e-5	2e-4	LOG	1.35386e-4
	ODC	Inverse ocean drag	days	1	3	LIN	2.55463
	SCF	Wind stress scaling		2	4	LIN	2.44654
	OP1	Power law for diapycnal diffusivity profile		0.5	1.5	LIN	1.07740
BIOGEM	PMX	Maximum PO_4 uptake	$mol\ kg^{-1}\ yr^{-1}$	5e-7	5e-5	LOG	2.27102e-5
	PHS	PO_4 half-saturation concentration	$mol\ kg^{-1}$	5e-8	5e-6	LOG	1.21364e-6
	PRP	Initial proportion POC export as recalcitrant fraction		0.01	0.1	LIN	0.031471
	PRD	e-folding remineralisation depth of non-recalcitrant POC	m	100	1000	LOG	802.258
	PRC	Initial proportion $CaCO_3$ export as recalcitrant fraction		0.1	1.0	LIN	0.22708
	CRD	e-folding remineralisation depth of non-recalcitrant $CaCO_3$	m	300	3000	LIN	1315.25
	RRS	Rain ratio scalar		0.01	0.1	LIN	0.076452
	TCP	Thermodynamic calcification rate power		0.2	2.0	LIN	0.510763
	ASG	Air-sea gas exchange parameter		0.3	0.5	LIN	0.46006
ENTS	VFC	Fractional vegetation dependence on carbon density	$m^2\ kgC^{-1}$	0.1	1.0	LIN	0.84249
	VBP	Base rate of photosynthesis	$kgC\ m^{-2}\ s^{-1}$	9.5e-8	2.2e-7	LIN	1.2040e-7
	LLR	Leaf litter rate	s^{-1}	2.4e-9	8.2e-9	LIN	2.4197e-9
	SRT	Soil respiration temperature dependence	K	197	241	LIN	218.356
	VPC[3]	CO2 fertilization Michaelis-Menton half-saturation	ppm	29	725	LOG	215.368

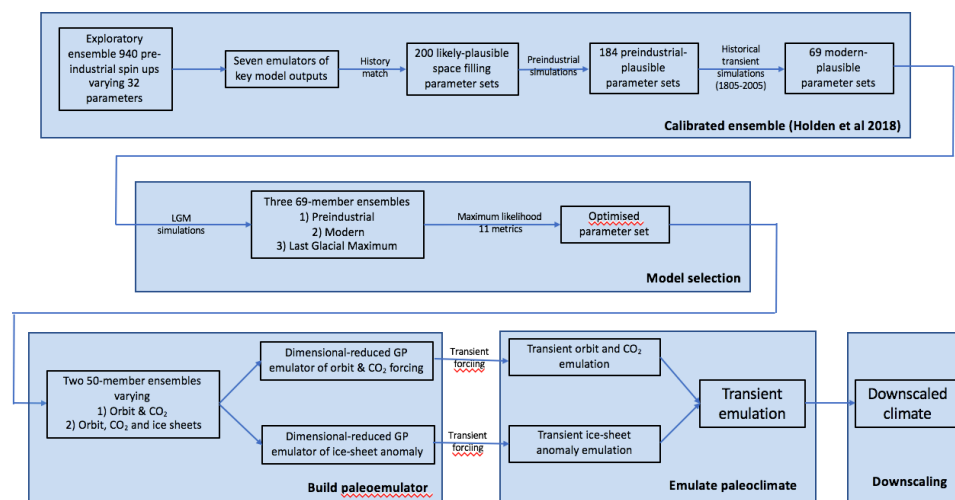
655 **Table 2: Prior distributions for PLASIM-GENIE varied parameters (uniform between ranges in log/linear space as stated). Notes. 1) ALBIS ice sheet albedo was fixed at 0.8 in the final ensemble. 2) APM was fixed at zero in the final ensemble (no flux correction). 3) VPC was not constrained by the emulator filtering as this parameter has no effect in the preindustrial spin up state. The final calibration step, selecting 69 simulations that satisfy present-day plausibility after the historical transient was primarily an exercise to calibrate the VPC parameter. Prior distributions are**
660 **discussed and derived from Holden et al (2010, 2013a, 2013b, 2014 and 2016). The final column tabulates the optimised parameter set.**

	Matern 3/2	Matern 5/2	Gaussian	Exponential	**Power exponential**
Orbit and CO_2 emulator					
Max precipitation	81.7%	80.2%	76.9%	81.0%	**82.7%**
Min precipitation	81.9%	80.9%	78.3%	81.8%	**82.7%**
Max SAT	97.7%	97.3%	96.8%	97.8%	**98.1%**
Min SAT	95.1%	95.2%	95.3%	95.2%	**95.0%**
Ice-sheet emulator					
Max precipitation	74.7%	72.5%	67.0%	71.9%	**75.4%**
Min precipitation	72.1%	69.3%	60.7%	69.4%	**73.3%**
Max SAT	94.2%	93.6%	92.3%	95.1%	**95.2%**
Min SAT	79.3%	77.5%	74.6%	80.8%	**80.9%**

665

Table 3. Optimization of the Gaussian process covariance function. The percentage of variance explained is quantified by the metric P (Eq. 3, including ten components) for each of the eight emulators, considering various tested covariance functions. A power exponential is favored for the final emulator, having similar average performance to exponential covariance function, but outperforming it for the more difficult precipitation variables.

670

Figures

675 **Figure 1: Schematic of experimental design**

680

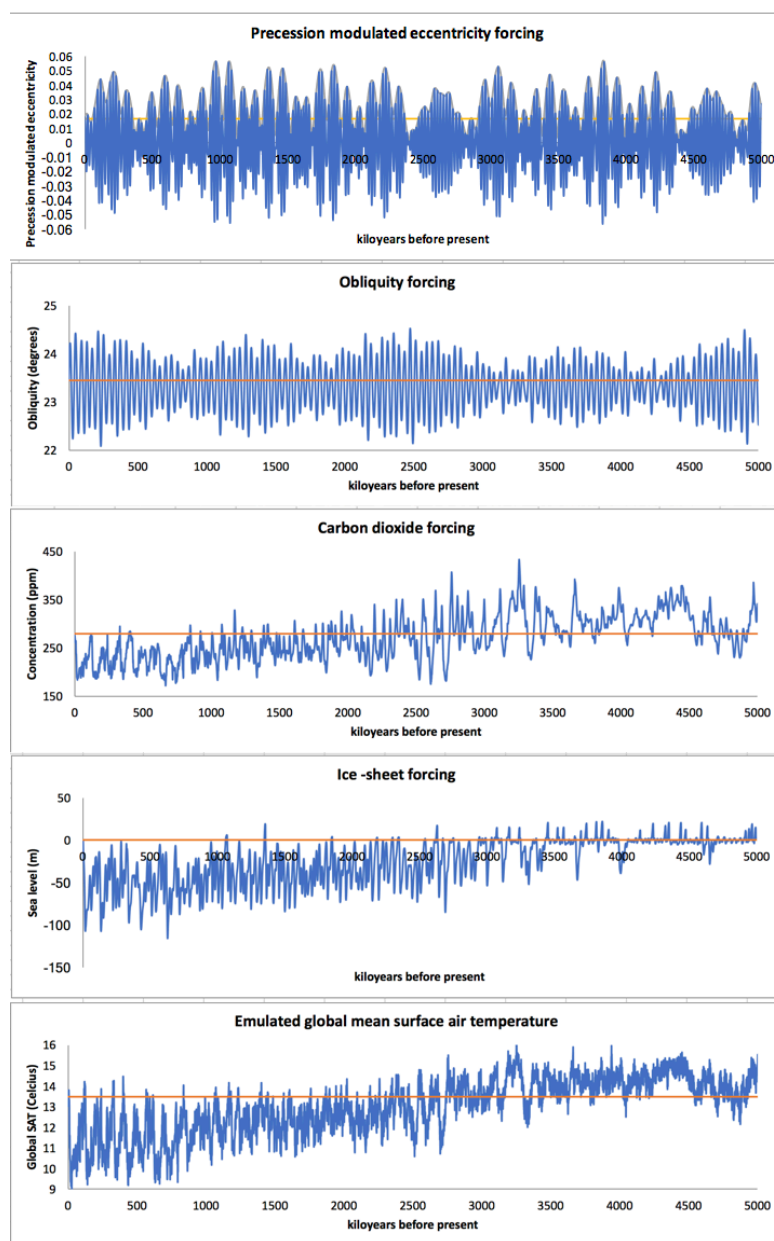



Figure 2: Emulator time-series forcing and reconstructed global surface air temperature. Orbital forcing is Berger and Loutre (1999). Ice-sheet forcing is the sea-level reconstruction of Stap et al (2017). Carbon dioxide forcing after 800,000 years BP is ice-core data (Luethi et al 2008), using the Stap et al (2017) reconstruction in the earlier period.

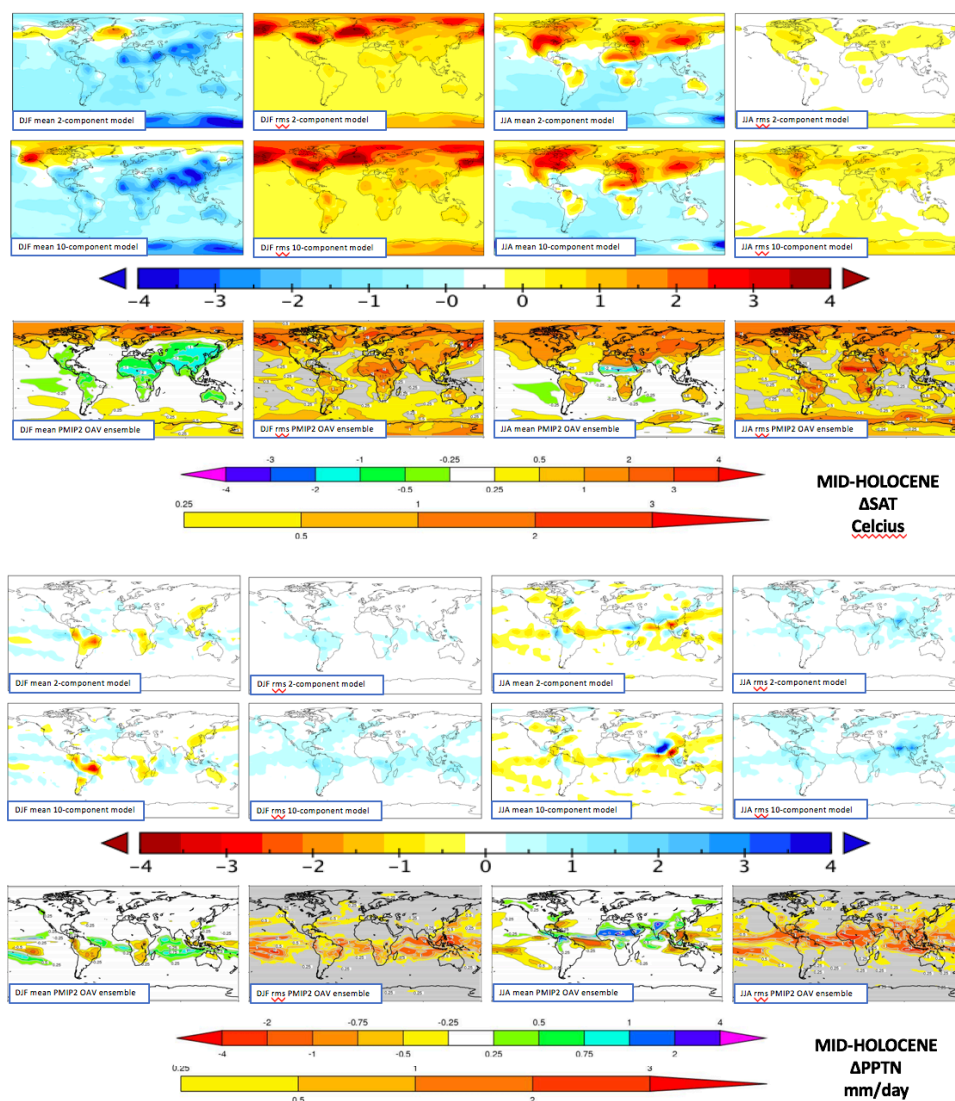
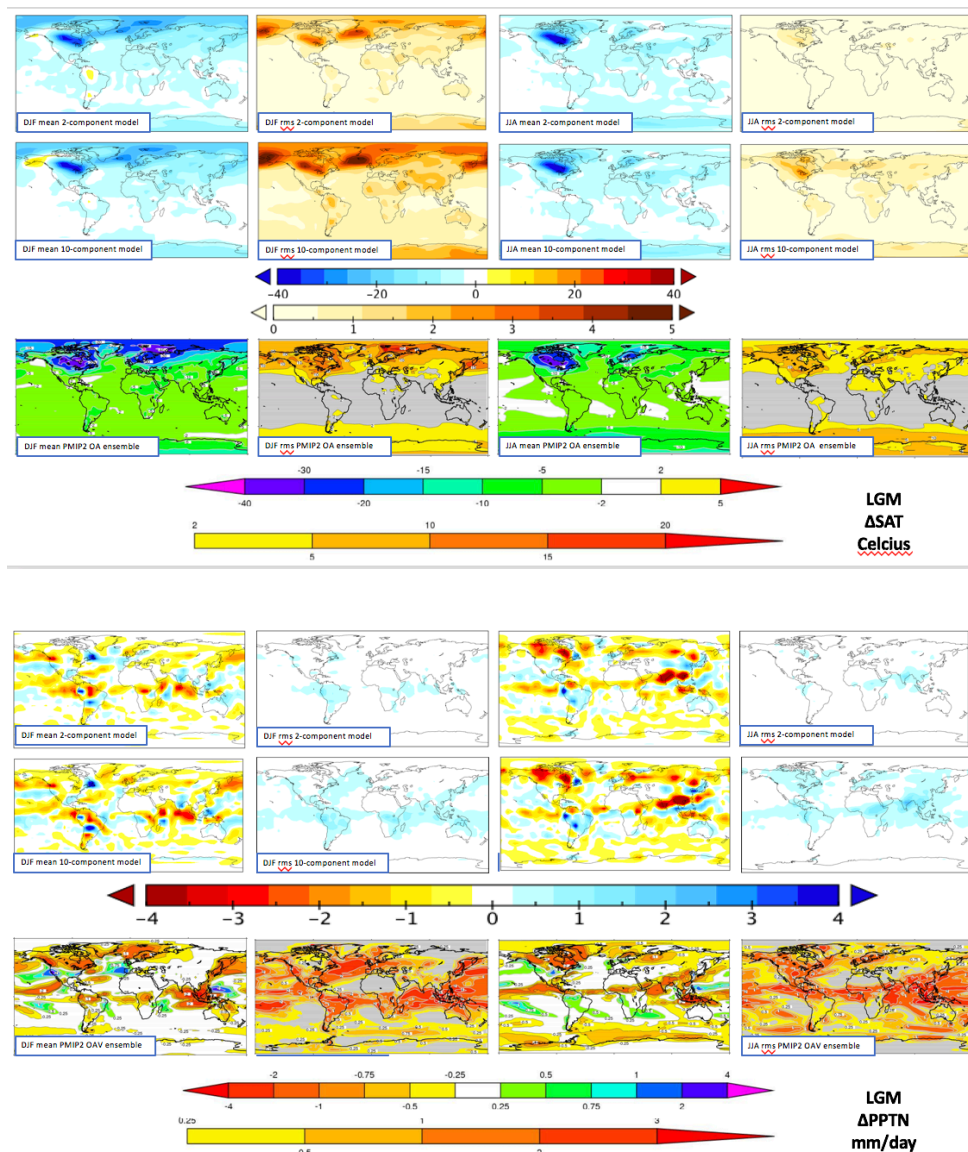
685

690

Figure 3: PALEO-PGEM emulated ensemble comparison with PMIP2 Ocean-Atmosphere-Vegetation Ensemble (Braconnot et al 2007) for the mid Holocene.

695

700

705

710 **Figure 4: PALEO-PGEM emulated ensemble comparison with PMIP2 Ocean-Atmosphere Ensemble (Braconnot et al 2007) for the Last Glacial Maximum.**

715

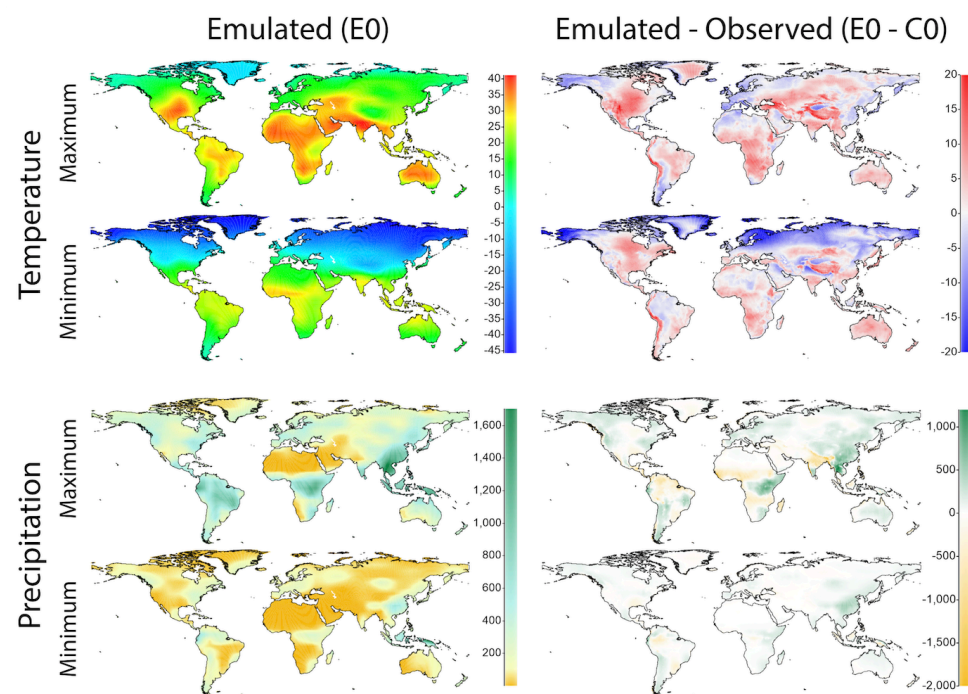



Figure 5: Downscaling the emulated climate. Left hand panels are the preindustrial emulations of the seasonal bioclimatic variables at native (T21) model resolution, interpolated onto the high-resolution grid. Right hand panels illustrate the differences with respect to high resolution climatology (Hijmans et al 2005).

720

725

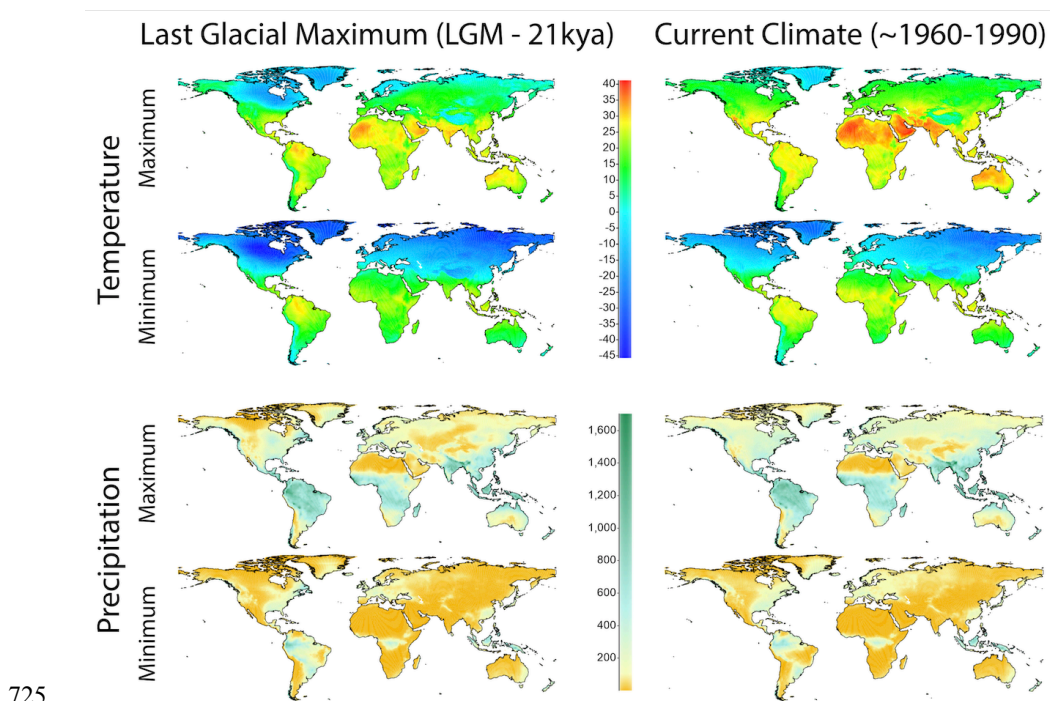



Figure 6: Downscaled emulated climate. Left hand panels are the downscaled emulated bioclimatic variables at the Last Glacial Maximum. Right hand panels are the present-day climatology (Hijmans et al 2005). Note that downscaled climates are derived by applying emulated anomalies to this present-day climatology. An animation of the complete 5 Ma reconstruction is provided as supplementary material.

730