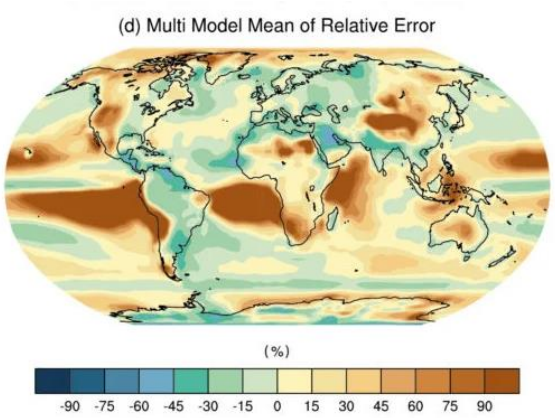
We are grateful to the Reviewers for their helpful comments, which we have accommodated as detailed in the point-by-point response below. We have also clarified the wording throughout the text, and reordered a few sentences, in order to improve the overall readability.

Reviewer 1

I thank the authors for having taken into consideration my concerns on the previous version. To my mind this version is very close to a publishable form have just three minor suggestions: in the introduction, the authors mention the order of magnitude of typical model bias: degrees for temperature, cm for annual precipitation. Since the delta method uses a multiplicative correction for precipitation, I would add also a typical percentage (or range of it) of precipitation bias.

As can be seen from the below Figure 9.4d of the latest IPCC report, present-day relative precipitation biases can range between -100% to over +100%, but vary considerably across space, making it difficult to define a 'typical percentage'.



To accommodate the Reviewer's suggestion, we have changed the sentence to

"these biases can oftentimes be of the order of several degrees of temperature, or tens of percent of precipitation.

In the conclusions, the authors conclude that the Delta method tends to perform better than the other wo. Again, I would mention here again that the results (may) depend on the climate model used. Most readers will just read the conclusions and infer that this conclusion has overall validity, which I think it has not been shown in this study.

We have added the following statement to the Conclusions:

We also reiterate that our results may be different for palaeoclimate simulations other than the ones used here.

Figures 3 and 4 are still difficult to read. I am aware that it is hard to convey the information the authors wish to. I would urge the authors to spend some time thinking about alternatives. One that they could try is to use discreet, instead of continuous, colour coding, but I am not convinced that this will improve the figures

We have changed the colouring to a discrete scale, as suggested, and have increased the resolution and overall quality of Figs. 3 and 4. We feel that the key aim of the figures,

to visualise major spatial clusters where the bias correction performance is particularly good or poor, as well as to highlight the overall spatial heterogeneity in the performance, is achieved well with the current format without requiring excessive manuscript space.

Reviewer 2

First, a clear discussion of the kinds of biases the authors are targeting should appear at the beginning of the paper. The examples of distributions of extreme weather events or climate variability in the introduction seem to be a bit of a red herring, as those "biases" have more to do with higher moments of probability distribution than the the time-mean biases that are ultimately the focus of the paper.

We have replaced the passage pointed out by the Reviewer by the following paragraph:

> In many of these applications, climatological normals at quasi-equilibrium of variables such as temperature and precipitation at different points in time represent the most relevant climatic inputs. [...] Three main methods have been used to bias-correct climatological normals in the palaeocontext: [...] Here we combine a set of high-resolution simulations of the climatological means [...]

We moved part of the replaced passage to the Conclusion.

Second, the absence of a discussion of paleoclimate state estimation and data assimilation is conspicuous given that the goals of those procedures are also to reduce misfits between models and paleo data. A starting point is the literature on offline Kalman filtering (e.g., Tardif et al. 2019, Clim Past, Last Millennium Reanalysis with an expanded proxy database and seasonal proxy modeling), particle filtering (e.g., Goosse 2016, Clim. Dyn., Reconstructed and simulated temperature asymmetry between continents in both hemispheres over the last centuries) and "online" state estimation that changes model forcing to generate new runs (e.g. Kurahashi-Nakamura et al. 2017, Paleoceanography, Dynamical reconstruction of the global ocean state during the Last Glacial Maximum and Amrhein et al. 2018, J. Clim, A Global Glacial Ocean State Estimate Constrained by Upper-Ocean Temperature Proxies). A comparison and discussion of complementarity would strengthen the paper and make it more relevant to CoP readers.

We have added the following paragraph to our discussion of Fig. 5.

> Such an approach would tie in with data assimilation methods, which also use empirical climate proxy records to improve climate simulations. These methods have been used to estimate global climate variables at times at which the quantity and spatial coverage of available empirical records is high enough to allow a robust calibration of the relevant computational methods. As a result, they have focussed either on single points in the past, such as the Mid-Holocene (Mairesse et al. 2013) or Last Glacial Maximum (Kurahashi-Nakamura et al. 2017), or on time intervals across which suitable empirical data are available, namely the last millennium period (Tardif et al. 2019, Gosset 2017). In contrast to the aforementioned approaches, based on Fig. 5 we suggest that it may be possible to use empirical reconstructions even from only a small set of points in time (e.g. the present, Mid-Holocene, LGM and Last Interglacial Period) to parameterise a statistical model of the temporal variation of local biases that could be used to improve simulated data at any time point in the Late Pleistocene-Holocene period.

p1l8-11 “slightly better…methods” It sounds like a more apt description is that the methods are indistinguishable. I would clarify what is meant by “slightly better”

We have added the following statement to the Abstract:

> In most cases, the differences between the bias reductions achieved by the three methods are small.

We would not agree with the conclusion that our findings show that the methods are indistinguishable. As we report in section 3, the Delta Method leads to significantly smaller median absolute biases for a number of variables and points in time than the other two methods, whereas only in 2 out of the 17 total scenarios displayed in Fig. 2 is the median absolute bias not smallest for the Delta Method.
We agree that the relative differences between the bias reductions achieved by the different methods are not always large, and therefore chose the phrasing “overall [...] performs slightly better” in the abstract. We feel that the following caveats “albeit not always to a statistically significant degree” and “however, there is considerable spatial and temporal variation in the performance of each of the three methods” add nuance to the statement with the brevity appropriate for the abstract.

p1l11 Should be a semicolon before however

We have added a semicolon.

p1l11 Please clarify what is meant by reconstructions — data? Potentially confusing because reconstructions often use model output. Please also comment on the utility of using interpolated products (e.g. the MARGO gridded product) to evaluate bias reduction, as those products have their own (likely biased) assumptions of spatiotemporal covariance built in.

We have revised the text and now use the term “reconstruction” exclusively in the context of data derived from empirical records. We agree that empirical climate reconstructions can themselves be subject to biases. This applies to the MARGO dataset, mentioned by the reviewer, as it does to pollen-derived reconstructions, where the transfer functions employed can be biased and subject to uncertainty. We have added the following statement to section 2.1.2 of the Methods accommodate the Reviewer’s comment:

> Empirically derived climate reconstructions can themselves be subject to biases and uncertainties, which arise at the different stages of the reconstruction process, from collecting the data to computationally converting empirical records to climatic variables. Nonetheless, these data represent the best empirically-based estimates of past climatic conditions available, and the most suitable data for our analysis.

p1l12 Please define what is meant by “active calibration” and “bias correction functions”

We have rephrased the sentence as follows:

> Our data also indicate that it could soon be possible to use empirical reconstructions of past climatic conditions not only for the evaluation of bias correction methods, but for fitting statistical relationships between empirical and simulated data through time that can inform more effective bias correction methods.

We have also added a model example to our description of Fig. 5 to illustrate a possible way to use past empirical reconstructions to improve the Delta Method.

p2l12 Please clarify what is meant by medium-scale. More accurate than “millennial-scale averages” might be “quasi-equilibrated climate states” when models are run for millennia. But I would dispute that these issues are not present in paleoclimate studies (e.g., the paleo drought literature).

We removed the sentence containing the phrase “medium-scale” in the course of addressing the Reviewer’s second comment (see above). We now clarify in section 2.1.1. of the Methods that the climate simulations used in our analysis represent climatological normals at quasi-equilibrium, i.e. following a 500-year spin-up period.

p2l27 “Finally…” This sentence needs clarification.

We have rephrased the sentence as follows:

> Quantile Mapping adjusts the cumulative distribution of the simulated data by applying a transformation between the quantiles of present-day simulated and observed climate to the quantiles of past simulated climate.

p2l28 “However…” What about the common practice of comparing paleoclimate anomalies in models and data? Isn’t that a validation of the “Delta method”? e.g., Brady et al. 2013, J. Cli., Sensitivity to Glacial Forcing in the CCSM4.

We agree with the Reviewer that it is not uncommon to compare paleoclimate simulation output to empirically derived reconstructions (cf. Fig. 9.11 from IPCC AR5 WG1); however, this is different from comparing the performance of alternative bias correction methods, which is the aim of our work. To avoid confusion, and because the sentence pointed out by the Reviewer is not essential to the argument, we have removed it.

p3l16 Please provide more detail on the model simulations. Were they run to equilibrium? Biases can emerge when models are run for long periods of time (Amrhein et al. 2018, cited above), but long runs are also necessary to equilibrate climate states to forcings (particularly in the deep ocean, e.g. Jansen et al. 2018 J. Cli., Transient versus Equilibrium Response of the Ocean’s Overturning Circulation to Warming).

We have clarified that the climate simulations used in our analysis represent climatological normals at quasi-equilibrium, i.e. after a 500-year spin-up period.

p3l19 It appears that the reference for the Last Interglacial has not been published. I’m not sure what Clim. Past’s policy is here, but it’s difficult to evaluate that output.

We leave it to the discretion of the Editor to decide whether referencing data that is presented in a preprint is acceptable.

p6l23 Please define “distributional bias” and “quantile.” This introductory paragraph (and the rest of the section) are difficult to to understand. What CDFs are being discussed? Perhaps the following paragraph (p7l3) should come first.

We have rewritten the paragraph as follows, and added an example to clarify the approach:

> Quantile Mapping aims to correct distributional biases in the simulated climate data. The method consists of first computing a transformation that maps the quantiles of the cumulative distribution function of all present-day observed values (i.e. from all land or ocean grid cells) of a climate variable onto the quantiles of the cumulative distribution function of all present-day simulated values. The derived mapping is then applied to the cumulative distribution function of all simulated values at a given point in the past. For example, let the cumulative distribution function of the values of present-day observed terrestrial mean annual temperature (i.e. from all land grid cells) map the value T_1°C onto the value $q \in [0,1]$, and let the analogous cumulative distribution function of present-day simulated terrestrial mean annual temperature map T_2°C onto q. If the value that is mapped onto q by the cumulative distribution function of simulated terrestrial mean annual temperature at a given point in the past is T_3°C, then the bias-corrected mean annual temperature in all grid cells with simulated mean annual temperature T_3°C at that point in time is estimated as $T_3 + (T_1 - T_2)$ °C. Notably, by design of the method, after applying Quantile Mapping, the cumulative distribution function of present-day bias-corrected simulated data is identical to the cumulative distribution functions of present-day observed values.

In the following paragraph, we have clarified that V represents one of the five climatic variables considered in our analysis. We have also clarified the domain of the cumulative distribution functions.

We feel that knowledge of the analytical details of each of the different bias correction methods developed in sections 2.2.1 - 2.2.3 is required before the methods can be placed into context and discussed in section 2.2.4. We have therefore not changed the order of sections.

p7l21 “By the nature of regression models” -- unclear what is meant here, please clarify

We have rephrase the relevant phase as follows:

> Because regressions generally do not fit the data perfectly, present-day biases modelled by the GAM will not exactly match the observed biases across all grid cells.

p8l9 measures -> measure

We have corrected the typo.

p8l12 I think that writing this out rather than using set notation would be more accessible to the readership of this journal.

We have removed the inaccessible notation, and rephrased the sentence as follows:

> For a climate variable V (representing the relevant temperature and precipitation variables considered here) ...

p10l4 Please define the difference between median bias and median absolute bias.

We have added “(Eq. 7)” and “(Eq. 8)” to refer to the definitions of the two statistics.

Figure 1 Why are error bars only on a subset of the data?

Error bars for temperature data from the Last Interglacial Period are not available to us. The supplementary material of the relevant paper (from 2010) only contains the estimated values, but not uncertainties. We had contacted the authors of the paper at an earlier point in time, but did not receive a response. We have stated in the Methods: “Standard errors of reconstructed values are available for all variables with the exception of terrestrial and marine temperature during the Last Interglacial Period.”

p18l11 “significanty”

We have corrected the typo.

~~A comparative~~ An empirical evaluation of bias correction methods for palaeoclimate simulations

Robert Beyer[1], Mario Krapp[1], and Andrea Manica[1]

[1]Department of Zoology, University of Cambridge, Downing Street, Cambridge CB2 3EJ, United Kingdom

Correspondence: Robert Beyer (rb792@cam.ac.uk)

Abstract.

Even the most sophisticated global climate models are known to have significant biases in the way they ~~reconstruct~~ simulate the climate system. Correcting model biases is therefore an essential step toward realistic palaeoclimatologies, which are ~~crucial for numerous applications~~important for many applications, such as modelling long-term ~~and large-scale~~ ecological dynamics. Here, we evaluate three widely-used bias correction methods – the Delta Method, Generalised Additive Models (GAMs) and Quantile Mapping – against a large global dataset of empirical temperature and precipitation records from the present, the Mid-Holocene (~6,000 years BP), the Last Glacial Maximum (~21,000 years BP) and the Last Interglacial Period (~125,000 years BP). In most cases, the differences between the bias reductions achieved by the three methods are small. Overall, the Delta Method performs slightly better, albeit not always to a statistically significant degree, at minimising the median absolute bias between empirical data and debiased simulations for both temperature and precipitation than GAMs and Quantile Mapping~~,~~ ; however, there is considerable spatial and temporal variation in the performance of each of the three methods. ~~Furthermore, our data indicate that additional~~ Our data also indicate that it could soon be possible to use empirical reconstructions of past climatic conditions ~~might make it possible to soon use past data~~ not only for the ~~validation but for the active calibration of bias correction functions~~evaluation of bias correction methods, but for fitting statistical relationships between empirical and simulated data through time that can inform more effective bias correction methods.

1 Introduction

Realistic reconstructions of global palaeoclimate are a key ~~requirement~~ input for modelling many important long-term and large-scale ecological processes (**??????**). ~~Despite~~ In many of these applications, climatological normals at quasi-equilibrium of variables such as temperature and precipitation at different points in time represent the most relevant climatic inputs. Simulations of these variables remain subject to substantial biases when compared to observational data, despite advancements in how complex physical processes are represented in global climate models ~~, simulated present-day climate remains subject to substantial biases when compared to observational data~~ (**??**). Depending on the region of interest, these biases can be of the order of ~~a few~~ several degrees of temperature, ~~or centimeters of annual~~ and tens of percent of precipitation, which can make the difference between markedly different vegetation types ~~(e.g. the shift from open to closed habitat, or the location of deserts)~~ (**?**).

Bias correction has received a great deal of attention for present-day and near-future simulations (**??**), whereas work on palaeoclimate ~~reconstructions~~ simulations has been much more limited. This is partly due to the different time scale of ~~palaeoecological~~ palaeoclimatological applications, for which computationally intensive bias correction methods that are used for the recent past and near future are not suitable. ~~Several challenges of methods used for bias-correcting future climate simulation data, including the correct representation of distributions of extreme weather events (e.g. precipitation during El Niño events, dry spell lengths), of very small-scale patterns, or of the variability of climatic variables across time scales of a few years or decades (?), are often not present in palaeoclimatological contexts. This is because palaeoclimate simulation data are generally provided at a medium-scale spatial resolution, and oftentimes represent millennial-scale averages. However, in both scenarios it is important to acknowledge that bias-correction methods are unable to substantially correct a fundamentally poor climate model, e.g. with strong circulation biases, which such methods are not capable of removing (?). Seeking to improve the representation of climate dynamics in simulation models therefore remains a priority alongside the development of bias correction methods.~~

~~There are three main methods that have been used so far in the palaeoclimatological context~~Three main methods have been applied thus far to bias-correct climatological normals in the palaeocontext: the Delta Method ~~(http://www.worldclim.org/downscaling)~~(http://www.worldclim.org/downscaling, http://www.paleoclim.org/methods/, **?**), statistical methods based on generalised additive models (GAMs) (**????**) and Quantile Mapping (**?**). ~~All three~~

methods are based on the assumption that the biases between present-day observations and simulated data do not change through time, although each method takes a different approach in the aspect that is assumed to be invariant. The Delta Method assumes bias to be that biases are location-specific (**?**), as it is based on and constant over time; it uses a map of the local differences between observed and simulated values at present-day to bias-correct past simulations (**?**). GAMs attempt to represent statistical relationships between simulated climatic present-day simulated climate variables (as well as other known physical variables, such as elevation and the distance from the coast) and bias-corrected climatic variables (**??**). Finally, Quantile Mapping assumes that biases are specific to their respective quantiles in the present-day observed climate, and apply these relationships to past simulations to reduce biases (**??**). Quantile Mapping adjusts the cumulative distribution of the relevant climatic variable (**?**). However, debiased simulation data have either not been validated against empirical reconstructions at all, or only for a small geographical area and a single point in the past. Here , we use simulated data by applying a transformation between the quantiles of present-day simulated and observed climate to the quantiles of past simulated climate. (**?**).

Here we combine a set of high-resolution climate simulations to evaluate the performance of the Delta Method, a GAM-based approach, and Quantile Mapping, against a global dataset of empirical climatology data from simulations of the climatological means of several temperature and precipitation variables for the present, the Mid-Holocene (~6,000 years BP), the Last Glacial Maximum (~21,000 years BP) and the Last Interglacial Period (~125,000 years BP) . As the first such effort, here, we with a global dataset of empirical climatic reconstructions to evaluate the performance of the Delta Method, a GAM-based approach, and Quantile Mapping in removing simulation biases. We focus on the global performance of the different methods; however, we note that bias-correction is , but point out that bias correction is generally not a one-size-fits-all approach (**?**), and that our results do not remove the need for local re-evaluations of methods in specific continental and subcontinental regions of interest.

Section **??** provides details of the three bias correction methods, the climate simulations, and the empirical palaeoclimatology palaeoclimate reconstructions used in this study. In section **??**, we quantitatively assess the performance of the methods at a global scale , the global scale and with regard to spatial and temporal heterogeneities. Section **??** discusses how empircal palaeoclimate reconstructions could be used not only to evaluate methods, but to help estimate the variation of local model bias biases over time, thus combining the strengths of the Delta Method and statistical bias correction.

2 Material and Methods

2.1 Climate data

2.1.1 Modelled climate data

We used ~~palaeoclimate simulations of monthly temperature and precipitation at a~~ 1.25°×0.83° ~~grid resolution~~ resolution palaeoclimate simulations of monthly mean temperature and monthly precipitation for the present, the Mid-Holocene and the Last Glacial Maximum (LGM) from the HadAM3H atmospheric model (**??**), which is part of the family of HadCM3 climate models (**?**). For the Last Interglacial Period, we do not have simulation data from HadAM3H, but we used the global climate model emulator GCMET (**?**) that is based on the same model and can make predictions at the same spatial resolution. In all cases, simulations represent climatological normals (i.e. 30-year averages) at quasi-equilibrium, following a 500-year spin-up period. Based on the monthly data, we computed the following climate variables, for which suitable empirical reconstructions are available (see section **??**): terrestrial mean temperature, marine mean annual temperature, temperature of the coldest month, temperature of the warmest month, and annual precipitation. We note that the results presented ~~in this article are~~ here may be specific to the particular climate simulations considered~~here~~, and do not claim generalisability to other models.

Empirical ~~data~~ reconstructions (see section **??**) of terrestrial temperature variables were compared against simulated temperature at 1.5 meters height, ~~whereas~~ while simulated air surface temperature was used as a proxy for sea surface temperature, as sea surface temperature is not part of the HadAM3H output. We removed marine data points for which simulated air surface temperature was below the freezing point of saltwater, –1.8°C, as in this case the simulated value corresponds to the temperature of an ice layer rather than that of the top layer of water.

2.1.2 Empirical climate data

All bias correction methods considered ~~in this paper are calibrated using~~ here are calibrated based on present-day observational data. For this, we used monthly terrestrial temperature and precipitation data at a 0.167° grid resolution (**?**), and mean annual sea surface temperature at a 1° grid resolution (**?**), representative of 1960–1990.

These maps were remapped to the $1.25° \times 0.83°$ grid of the palaeoclimate simulations by taking the average of values contained in each target grid cell.

We used global datasets of ~~empirical local palaeoclimate~~ local empirical palaeoclimatic reconstructions of terrestrial mean annual temperature, temperature of the coldest and warmest month, and annual precipitation, for the Mid-Holocene and the LGM from **?**, reconstructions of mean annual sea surface temperature for the Mid-Holocene and the LGM from **?** and **?**, respectively, and reconstructions of mean annual terrestrial and sea surface temperature for the Last Interglacial Period from **?**. Standard errors of reconstructed values are available for all variables with the exception of terrestrial and marine temperature during the Last Interglacial Period.

Terrestrial temperature and precipitation reconstructions for the Mid-Holocene and the LGM are ~~provided~~ available on a 2° resolution grid, and LGM marine temperature reconstructions are provided on a 5° grid. We assigned each sample of these datasets to the $1.25° \times 0.8°$ grid cell of our palaeoclimate simulations (see section **??**) that contains the centre of the relevant 2° or 5° cell. Reconstructions for the Last Interglacial Period are not gridded, and were ~~compared to the simulated climate in the~~ assigned to the $1.25° \times 0.8°$ grid cell ~~containing~~ that contains the sample location. Figs. **??** and **??** visualise the locations of all empirical reconstructions of terrestrial and marine mean annual temperature, and annual precipitation.

Empirically derived climate reconstructions can themselves be subject to biases and uncertainties, which arise at the different stages of the reconstruction process, from collecting the data to computationally converting empirical records to climatic variables. Nonetheless, these data represent the best empirically-based estimates of past climatic conditions available, and the most suitable data for our analysis.

2.2 Bias correction methods

2.2.1 The Delta Method

The Delta Method ~~is based on~~ consists of adding the difference between past and present-day simulated climate (the ~~'delta'~~Delta) to present-day observed climate. ~~Thus~~As such, the Delta Method assumes that ~~the~~ local (i.e. grid cell-specific) model ~~bias is~~ biases are constant over time (**?**). For temperature variables (including terrestrial and marine mean annual temperature, and terrestrial temperature of the warmest and coldest month, considered here), the bias in a geographical location x is given by the difference between present-day observed and ~~raw~~ simulated temperature, $T_{\mathrm{emp}}(x,0) - T^{raw}_{\mathrm{sim}}(x,0)$. ~~Debiased temperature , $T^{DM}_{\mathrm{sim}}(x,t)$, in location~~ Bias-corrected

temperature in x at some time t ~~is obtained as :~~ in the past is estimated as

$$T_{\text{sim}}^{DM}(x,t) := T_{\text{emp}}(x,0) + \left(T_{\text{sim}}^{raw}(x,t) - T_{\text{sim}}^{raw}(x,0)\right)$$
$$= T_{\text{sim}}^{raw}(x,t) + \left(T_{\text{emp}}(x,0) - T_{\text{sim}}^{raw}(x,0)\right). \quad (1)$$

The second expression illustrates that $T_{\text{sim}}^{DM}(x,t)$ is alternatively given by adding the ~~present-day local~~ local present-day bias to the ~~simulated temperature , $T_{\text{sim}}^{raw}(x,t)$, at~~ local temperature simulated for time t.

Precipitation is bounded below by zero and covers different orders of magnitude across different regions. A multiplicative rather than additive bias correction is therefore more ~~adequate~~ common when applying the Delta Method for precipitation, which corresponds to applying the simulated relative change to the observations (**?**). Analogously to temperature, debiased precipitation is ~~given by~~ estimated as

$$P_{\text{sim}}^{DM}(x,t) := P_{\text{obs}}(x,0) \cdot \frac{P_{\text{sim}}^{raw}(x,t)}{P_{\text{sim}}^{raw}(x,0)}$$
$$= P_{\text{sim}}^{raw}(x,t) \cdot \frac{P_{\text{obs}}(x,0)}{P_{\text{sim}}^{raw}(x,0)}. \quad (2)$$

2.2.2 Statistical Models / GAMs

Statistical bias correction methods assume the existence of a functional relationship between ~~(i)~~ true climatic conditions (dependent variables), and ~~(ii)~~ climate model outputs as well as additional known forcings such as topography (independent variables) (**??**). Transfer functions representing this relationship are calibrated ~~on the basis of~~ based on present-day simulated and observed climate, and are then ~~used to derive past climate using the appropriate simulated~~ applied to simulations of past climate to derive bias-corrected data. Generalised additive models (GAMs) have gained particular popularity as transfer functions (**????**). They accommodate potential nonlinearities in the response of the individual ~~variables, while~~ predictor variables, but – owing to the computational requirements of general high-dimensional nonlinear regressions – ~~assuming~~ assume that the interactions between ~~predictor variables~~ predictors can be neglected.

For a set of geographical locations $x_1, x_2, \ldots$, we denote by $V_{\text{emp}}(x_i,0)$ the present-day observed value of a climate variable ~~denoted~~ V ~~, at~~ (representing the relevant temperature and precipitation variables considered here) in the location x_i. Here, the $x_1, x_2, \ldots$ represent ~~land and ocean points on~~ the locations of the cells of the $1.25° \times 0.8°$ grid of the climate data (see section **??**) on land and in the ocean in the case of terrestrial and marine climate variables, respectively. In a GAM, ~~these~~ the present-day observed values of V are modelled as

the sum of functions of variables that are available both for the present and the past, such as climate model outputs (typically including the raw simulated data of the ~~climate~~ variable in question~~, V_{sim}^{raw}~~), and/or certain geographical or physical quantities that are known across time. We denote the values of these predictor variables in the location x_i at time t by $X_1^V(x_i,t), X_2^V(x_i,t), \ldots$. ~~In general, the~~ The X_j^V are generally time-dependent~~;~~, not only when they ~~represent~~ are climate model outputs, but also when they represent elevation or the distance to the ~~ocean~~coast, which vary over time as the result of sea level changes. ~~Finally, the GAM is given~~ The GAM is defined by the regression

$$V_{\text{emp}}(\cdot,0) \sim \sum_j f_j\big(X_j^V(\cdot,0)\big), \tag{3}$$

where the ~~$f_1, f_2 \ldots$~~ $f_1, f_2, \ldots$ represent smooth functions that are fitted to minimise the distance between the left and the right hand side in Eq. (**??**). Once the model has been calibrated on the present-day data, it ~~can be~~ is used to estimate the ~~true (i.e.~~ bias-corrected ~~)~~ values of the climate variable ~~of interest~~ V in the location x_i at ~~some point in past t~~ a point t in the past as

$$V_{\text{sim}}^{GAM}(x_i,t) := \sum_j f_j\big(X_j^V(x_i,t)\big). \tag{4}$$

Similar to **?**, here, we used elevation, the shortest distance to the ocean and simulated temperature as predictor variables X_j^V for temperature variables~~; we use elevation~~. Elevation, the shortest distance to the ocean, and simulated annual precipitation, temperature, (absolute) wind speed, air pressure and relative humidity were used as predictors variables for annual precipitation. The functions f_i were estimated as piecewise third order polynomials (using thin plate splines did not change the results) using the *mgcv* package ~~**?** in R~~ in R (**?**).

2.2.3 Quantile Mapping

Quantile Mapping aims to correct distributional biases in the simulated climate data. ~~For a given climate variable , Quantile Mapping applies a correction to present and past simulated climate valuesthat is specific to the quantile associated with the relevant value within the set~~ The method consists of first computing a transformation that maps the quantiles of the cumulative distribution function of all present-day observed values (i.e. from all land or ocean grid cells) of a climate variable onto the quantiles of the cumulative distribution function of all present-day simulated values. The derived mapping is then applied to the cumulative distribution function of all simulated values at ~~the appropriate point in time. This correction is calculated based on the difference~~

~~between present-day simulated and observed quantiles. As a result,~~ a given point in the past. For example, let the cumulative distribution ~~functions~~ function of the values of present-day observed ~~and~~ terrestrial mean annual temperature (i.e. from all land grid cells) map the value T_1°C onto the value $q \in [0,1]$, and let the analogous cumulative distribution function of present-day simulated ~~data that was~~ terrestrial mean annual temperature map T_2°C onto q. If the value that is mapped onto q by the cumulative distribution function of simulated terrestrial mean annual temperature at a given point in the past is T_3°C, then the bias-corrected ~~using Quantile Mappingare identical~~ mean annual temperature in all grid cells with simulated mean annual temperature T_3°C at that point in time is estimated as $T_3 + (T_1 - T_2)^\circ$C. Notably, by design of the method, the cumulative distribution function of present-day bias-corrected simulated data (i.e. after applying Quantile Mapping) is identical to the cumulative distribution functions of present-day observed values.

Formally, denote by $x_1, x_2, \ldots$ the centres of the $1.25^\circ \times 0.8^\circ$ grid cells of the climate data (see section **??**) on land and in the ocean in the case of terrestrial and marine climate variables, respectively. For a climate variable V (representing the relevant temperature and precipitation variables), we denote by $F^V_{\text{emp}}[0]$ the cumulative distribution function of ~~the~~ all present-day empirical observations, $V_{\text{emp}}(x_1, 0), V_{\text{emp}}(x_2, 0), \ldots$ (i.e. $F^V_{\text{emp}}[0]$ is the function that monotonically maps these values onto the interval $[0,1]$). Analogously, we denote by $F^{V,raw}_{\text{sim}}[t]$ the cumulative distribution function of the raw simulated values $V^{raw}_{\text{sim}}(x_1, t), V^{raw}_{\text{emp}}(x_2, t), \ldots$ at time t. We denote by $F^V_{\text{emp}}[0]^{-1}$ and $F^{V,raw}_{\text{sim}}[t]^{-1}$ (both mapping $[0,1]$ to $\mathbb{R}$) the inverse functions of $F^V_{\text{emp}}[0]$ and $F^{V,raw}_{\text{sim}}[t]$, respectively.

With this notation, $F^{V,raw}_{\text{sim}}[t](V^{raw}_{\text{sim}}(x_i, t))$ is the quantile corresponding to the value $V^{raw}_{\text{sim}}(x_i, t)$ in the set of ~~simulated values $V^{raw}_{\text{sim}}(x_1, t), V^{raw}_{\text{emp}}(x_2, t), \ldots$~~ all simulated values of the climate variable V at time t. Under Quantile Mapping, the function ~~$[F^V_{\text{emp}}[0]^{-1} - F^{V,raw}_{\text{sim}}[0]^{-1}]$~~ $F^V_{\text{emp}}[0]^{-1} - F^{V,raw}_{\text{sim}}[0]^{-1}$ maps each such quantile to a quantile-specific correction term, which is then applied to the raw simulation data. Thus, ~~we obtain~~

~~$$V^{QM}_{\text{sim}}(x_i, t) := V^{raw}_{\text{sim}}(x_i, t) + \underbrace{\left[F^V_{\text{emp}}[0]^{-1} - F^{V,raw}_{\text{sim}}[0]^{-1}\right]}_{\text{Correction term specific to the quantile of } V^{raw}_{\text{sim}}(x_i,t)} \left(F^{V,raw}_{\text{sim}}[t](V^{raw}_{\text{sim}}(x_i, t))\right).$$~~

~~for~~ the bias-corrected value of V in the location x_i at time t ~~and location x_i.~~ is estimated as

$$V^{QM}_{\text{sim}}(x_i, t) := V^{raw}_{\text{sim}}(x_i, t) + \underbrace{\left[F^V_{\text{emp}}[0]^{-1} - F^{V,raw}_{\text{sim}}[0]^{-1}\right]}_{\text{Correction term specific to the quantile of the value } V^{raw}_{\text{sim}}(x_i,t)} \left(F^{V,raw}_{\text{sim}}[t](V^{raw}_{\text{sim}}(x_i, t))\right). \quad (5)$$

2.2.4 Method discussion

All three bias correction methods considered here aim at minimising biases in past simulated data, but they ~~make~~ are based on different assumptions as to how this aim can best be achieved. The Delta Method assumes that the ~~(known)~~ known present-day model bias is also a good estimate for past model bias. GAM methods and Quantile Mapping operate on the premise that this assumption of that Delta Method – local biases remaining constant over time – is too strong. Instead, GAM methods assume that a better estimate of past model biases can be obtained by deriving a statistical relationship between present-day bias and present-day simulations, and then applying this relationship to past simulations in order to estimate past bias. ~~By the nature of regression models, GAM methods do not perfectly explain~~ Because regressions generally do not fit the data perfectly, present-day ~~model biases across grid cells via the predictor variables. As a result (and unlike~~ biases modelled by the GAM will not exactly match the observed biases across all grid cells. Unlike in the case of the Delta Method~~)~~, GAM-corrected present-day simulations are therefore not identical to the present-day observed climate. This drawback is accepted under the assumption that the derived statistical model captures the ~~*mechanisms* underlying~~ mechanisms that underlie local model biases better than the ~~time-constant~~ time-invariant local correction term used in the Delta Method, and indeed to an extent that ~~allows better~~ results in more accurate estimates of past model biases. Similarly, Quantile Mapping assumes that the distributional correction of climate quantiles – whilst, again, not perfectly eliminating biases in present-day simulations – ultimately represents a better strategy for minimising past bias than the rigid local correction of the Delta Method.

Another important commonality between the methods is that they are calibrated only using present-day simulated and observed data. All three are based on the concept of establishing a relationship between present-day simulated and observed data, and then extrapolating that relationship in order to estimate past biases. The specific aspect that is assumed to be invariant over time is the present-day local bias in the case of the Delta Method, the regression model linking present-day simulated and observed data in the case of GAMs, and the present-day distributional correction in the case of Quantile Mapping.

2.3 Method evaluation

~~In ecological applications, the objective of applying a bias-correction method to past simulated climate datais generally to reduce the difference between the simulated and the (generally unknown) true past climate. Empirical palaeoclimatic reconstructions~~

Empirical palaeoclimate reconstructions of climatological normals allow us to assess ~~these differences at specific locations and points in time. Here, we determine~~ the performance of different bias correction methods in removing biases in past simulated data. In the following, we define the local differences between empirical reconstructions and bias-corrected simulations for ~~each climate variable and bias-correction method , and define~~ the different climate variables and bias correction method considered, and develop a spatially aggregated ~~measures~~ measure to assess the ~~overall~~ global performance of each method.

We denote by $V_{\text{emp}}(x,t)$ the empirically reconstructed value ~~at time t in a location x of the climate variable $V \in \{T^{\text{ter.mean}}, T^{\text{Tmar.mean}}, T^{\text{cold}}, T^{\text{warm}}, P^{\text{ann}}\}$, representing terrestrial or~~ of a climate variable V (representing terrestrial mean temperature, marine mean annual temperature, temperature of the coldest ~~or~~ month, temperature of the warmest month, or annual precipitation~~, respectively. For $M \in \{\text{raw, DM, GAM, QM}\}$~~) at a time t in a location x. For a bias correction method M (representing the Delta Method, GAM-based statistical bias correction, Quantile Mapping), we denote by $V_{\text{sim}}^{M}(x,t)$ the simulated value of the climate variable V at the time t in the location x ~~, where the underlying simulation data was not debiased or was bias-corrected using the Delta Method, the GAM method or Quantile Mapping, respectively~~that was processed using the method M. The ~~local bias~~ remaining local bias, B, between the empirically reconstructed and the bias-corrected simulation data at time t in the location x is then given by

$$B_V^M(x,t) = \begin{cases} V_{\text{sim}}^M(x,t) - V_{\text{emp}}(x,t) & \text{if } V \text{ is a temperature variable} \\ \dfrac{V_{\text{sim}}^M(x,t) - V_{\text{emp}}(x,t)}{V_{\text{emp}}(x,t)} & \text{if } V \text{ is annual precipitation} \end{cases} \tag{6}$$

~~We use~~ Thus, we used the absolute difference between empirical and simulated data for temperature variables, and the relative difference in the case of precipitation. We denote by $x_1^{(t,V)}, x_2^{(t,V)}, \ldots$ the geographical locations of the available ~~empirically reconstructed samples~~ empirical records at time t for the climate variable V. In section **??**, we provide ~~complete plots of the distribution of the biases corresponding to each specific~~ the complete distributions of the local biases that were derived for each climate variable, point in time, and bias correction method. ~~Furthermore~~In addition, as a summary statistic of these distributions ~~, and an~~ and a spatially aggregated measure for evaluating ~~and comparing~~ the performance of ~~the three~~ each bias correction

methods, we ~~use~~ used the median of the available local absolute biases $\{|B_V^M(x_i^{(t,V)},t)|\}_{i=1,2,\ldots}$. The median is weighted by grid cell area for the present, and by the local inverse standard errors of the empirical data for the past. We ~~denote the latter~~ rescaled the latter proportionally so that their sum equals 1, and denote the result by $\{\omega_{\text{emp}}(x_i^{(t,V)},t)\}_{i=1,2,\ldots}$ ~~, rescaled such that~~ (i.e. $\sum_i \omega_{\text{emp}}(x_i^{(t,V)},t) = 1$~~. The~~). Formally, the median absolute bias~~for variable V and bias-correction~~ , *MAB*, for the variable V and the bias correction method M at time t is ~~then formally~~ given by

$$
\begin{aligned}
MAB_V^M(t) &= \text{weighted median}\left(\{|B_V^M(x_i^{(t,V)},t)|\}_{i=1,2,\ldots}\right) \\
&= |B_V^M(x_k^{(t,V)},t)|, \text{ where the median index } k \text{ satisfies} \\
&\sum_{\substack{|B_V^M(x_i^{(t,V)},t)|<\ldots \\ |B_V^M(x_k^{(t,V)},t)|}} \omega_{\text{emp}}(x_i^{(t,V)},t) \leq \frac{1}{2} \text{ and } \sum_{\substack{|B_V^M(x_i^{(t,V)},t)|>\ldots \\ |B_V^M(x_k^{(t,V)},t)|}} \omega_{\text{emp}}(x_i^{(t,V)},t) \leq \frac{1}{2}. \qquad (7)
\end{aligned}
$$

Thus, the median absolute bias is a measure of the average difference between empirical and the bias-corrected simulated data. We ~~consider~~ considered a bias correction method to overall improve the raw simulation outputs if the associated median absolute bias is smaller than the median absolute difference between raw simulations and empirical data. ~~We emphasise that~~ The local biases B in Eq. (**??**) and the *MAB* ~~is a summary statistic of the extent to which a given bias-correction method reduces the difference between simulated and empirical climatic data, i.e. it does not allow inference of the goodness of the climate model, or of the performance of the different methods in improving~~ only permit to assess the performace of the three methods in bias-corrrecting quasi-quilibrated climatological normals of the variables considered here, not of climatic signals that are not captured by the ~~empirical dataused here~~underlying data, such as short-term climatic variability.

We tested whether the median absolute biases associated with any two ~~bias-correction~~ bias correction methods, a certain climate variable and point in time, were statistically significantly different ~~,~~ under the given uncertainty in the empirical reconstructions, using the following approach. For each climate variable and point in time, we generated 10^4 Monte Carlo realisations of empirical past climatic values in the locations where reconstructions are available by applying a normally-distributed noise term, with mean zero and standard deviation equal to the error of the local empirical reconstruction, to the value provided by the empirical reconstruction. Next, we calculated the local absolute biases between these empirical past climatic values, and the ~~appropriate~~ relevant simulated values obtained after applying the different ~~bias-correction~~ bias correction methods. For each of these 10^4 sets of local absolute biases between empirical and simulated data, we used a one-sided Wilcoxon

rank sum test to assess whether the median of the absolute biases associated with one ~~bias-correction~~ bias correction method was significantly smaller than that associated with a different ~~bias-correction~~ bias correction method (at a 5% significance level). We then determined the number of iterations, out of the total 10^4 Monte Carlo realisations, in which this was the case. If, for a given climate variable and point in time, a ~~bias-correction~~ bias correction method was found to perform significantly better than another one in more than half of the realisations, we report this result in section **??**.

Debiased simulated data should ideally not contain any systematic bias, in that the median bias, *MB*, given by

$$MB_V^M(t) = \text{weighted median}\left(\{B_V^M(x_i^{(t,V)}, t)\}_{i=1,2,\ldots}\right), \tag{8}$$

~~(~~where the weighted median is ~~calculate~~ calculated analogously as in Eq. (**??**)~~)~~ , should not differ substantially from zero. In addition to ~~considering~~ the median absolute bias ~~, here~~(Eq. (**??**)), we also ~~examine how~~ examined how the different methods affect the associated median bias (Eq. (**??**)).

In some applications, the climate change signal, i.e. the difference between past and present climatic states, may be more relevant than the climate at a fixed point in time. The difference between the empirical and the simulated climate change signal, *CCB*, of a climate variable V that was bias-corrected using method M at a location x and between the present and time t in the past is ~~given by~~ calculated as

$$CCB_V^M(x_i^{(t,V)}, t) = \begin{cases} \left(V_{\text{sim}}^M(x,t) - V_{\text{sim}}^M(x,0)\right) - \left(V_{\text{emp}}(x,t) - V_{\text{emp}}(x,0)\right) \\ \qquad\qquad \text{if } V \text{ is a temperature variable} \\ \dfrac{V_{\text{sim}}^M(x,t) - V_{\text{sim}}^M(x,0)}{V_{\text{sim}}^M(x,0)} - \dfrac{V_{\text{emp}}^M(x,t) - V_{\text{emp}}(x,0)}{V_{\text{emp}}(x,0)} \\ \qquad\qquad \text{if } V \text{ is annual precipitation,} \end{cases} \tag{9}$$

and the median absolute bias associated with the climate change signal, *CCMAB*, is given by

$$CCMAB_V^M(t) = \text{weighted median}\left(\{|CCB_V^M(x_i^{t,V)}, t)|\}_{i=1,2,\ldots}\right), \tag{10}$$

where the weighted median is calculated analogously as in Eq. (**??**). ~~Here, we also compare~~ We also compared the performance of the ~~different~~ three bias correction methods ~~for~~ in terms of this quantity. We did not determine the median absolute bias for the climate change signal between different points in the past, due to the much

smaller number of empirical ~~reconstructions~~ records that are available from the same location across ~~time~~the past, and due to the increased uncertainty of the local empirical climate change signals, which are given by the sum of the uncertainties of the local reconstructions of the relevant points in time.

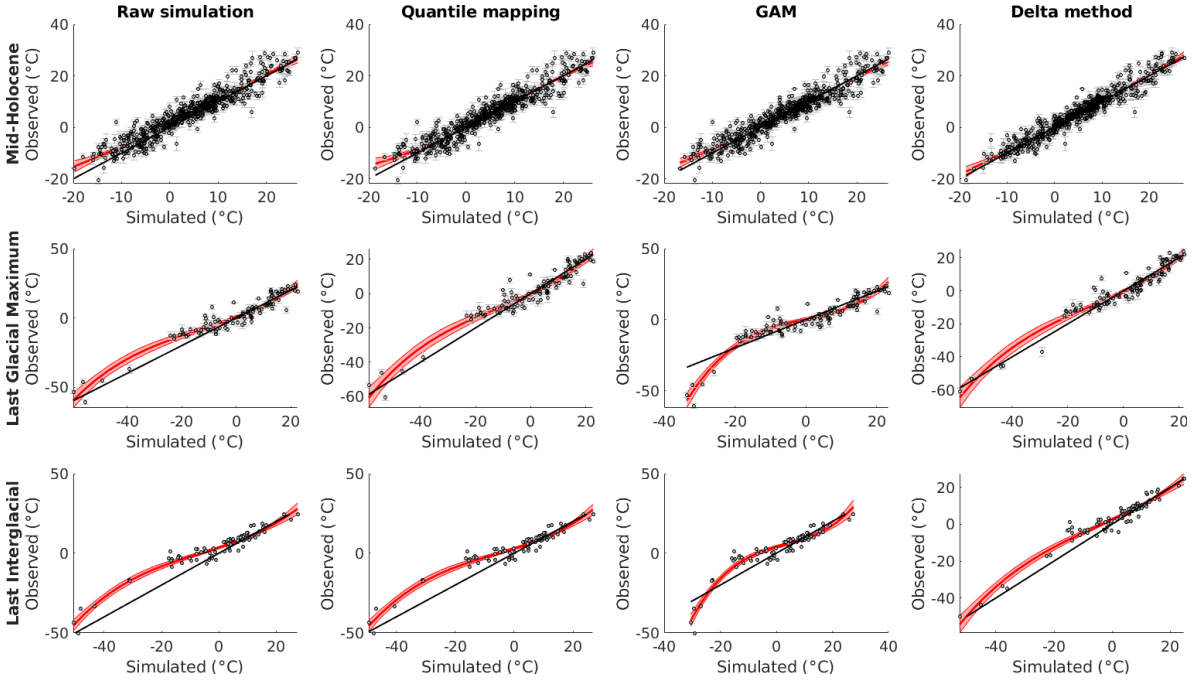
3 Results

Fig. **??**a–e compare empirically reconstructed and bias-corrected simulated ~~climate~~ data for the five climate variables considered. They show that ~~biases remaining after applying the different bias-correction methods~~ the biases that remain after applying a bias correction method are not uniformly distributed across the range of simulated values. In a number of cases, very low temperatures in several bias-corrected simulations tend to be lower than empirically reconstructed values, while very high temperatures in the simulated data tend to be higher than what empirical reconstructions suggest (e.g. Mid-Holocene and Last Interglacial mean annual marine temperature, and Mid-Holocene and LGM temperature of the warmest month). For some ~~bias-correction~~ bias correction methods, an analogous patterns can be observed in the case of precipitation.

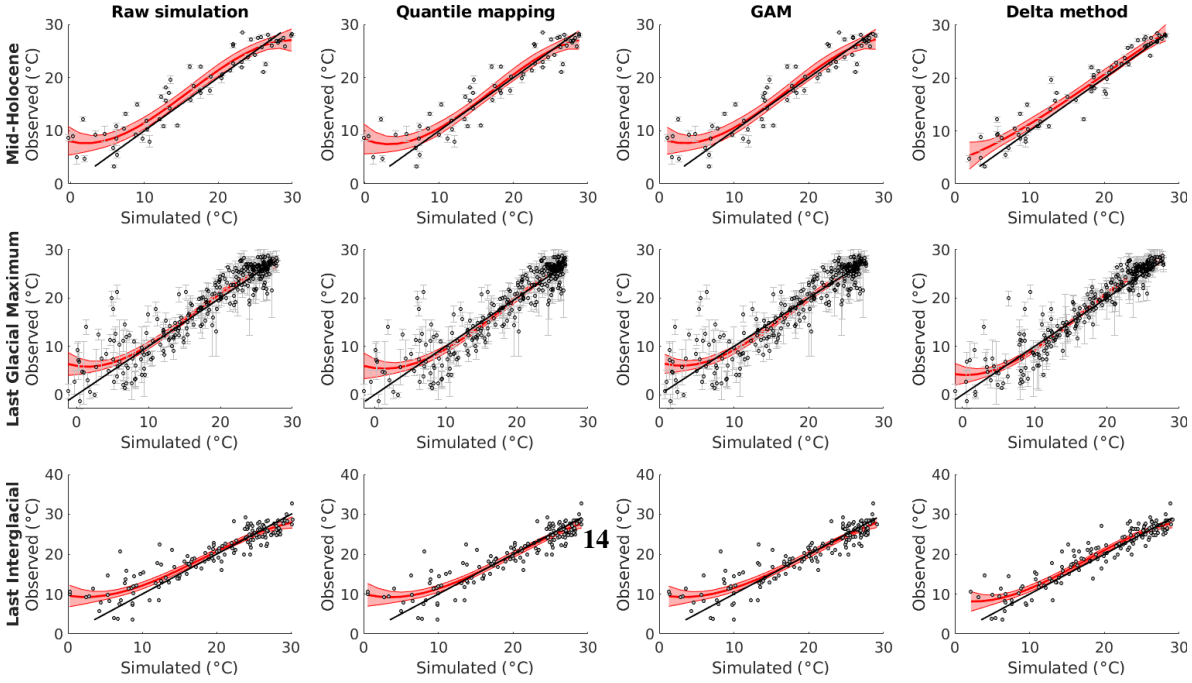
All ~~bias-correction~~ bias correction methods reduce the median absolute bias (*MAB* in Eq. (**??**)) of present-day simulated data for all climate variables~~–~~, as would be expected (**?**) ~~– although, by construction, only~~ (Fig. **??**). By construction, the Delta Method completely eliminates all differences between present-day simulated and observed data~~(Fig. **??**)~~. The Delta Method also provides the strongest reduction in the median absolute bias (*MAB* in Eq. (**??**)) for all variables and points in time, with the expection of temperature of the coldest month at the Mid-Holocene ~~,~~and precipitation at the LGM (Fig. **??**). The comparatively good performance of the Delta Method is ~~also reflected in the~~ reflected in strong correlations between present-day and past model biases, which the Delta Method assumes to be similar (Fig. **??**). The GAM method and Quantile Mapping also generally lead to a reduction in bias, ~~even~~ though overall not ~~as effectively~~ quite as strongly as the Delta Method. In a few cases, the original bias is actually increased after applying a correction method (Fig. **??**).

These above trends in the performances of the different ~~bias-correction~~ bias correction methods in terms of the median absolute bias are not always statistically significant. The median absolute bias associated with the Delta Method was significantly smaller ($p < 0.05$) than that associated with Quantile Mapping and the GAM method for Mid-Holocene terrestrial mean annual temperature (in 96% and 83% of Monte Carlo realisations (see section **??**) when compared against Quantile Mapping and the GAM method, respectively), marine mean

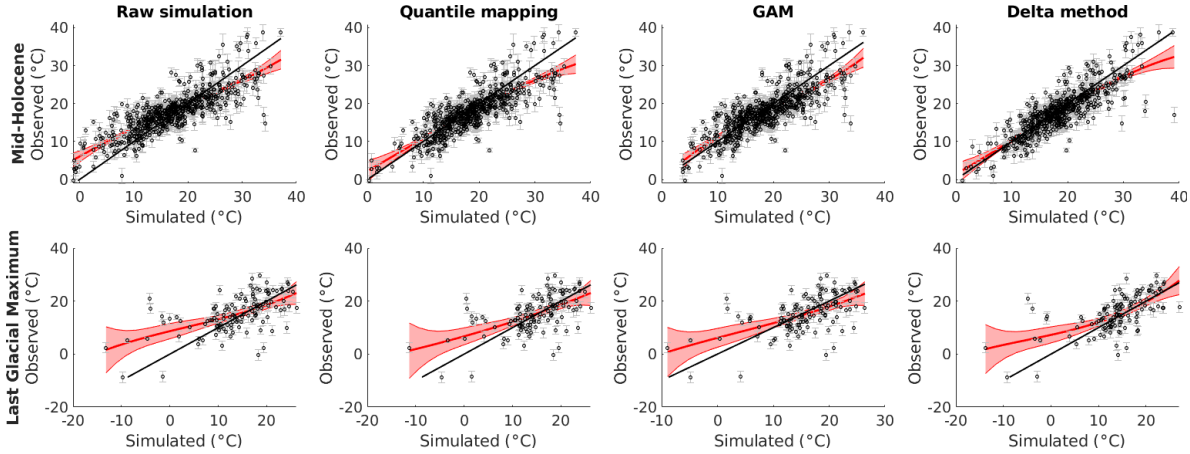
(a) Terrestrial mean annual temperature



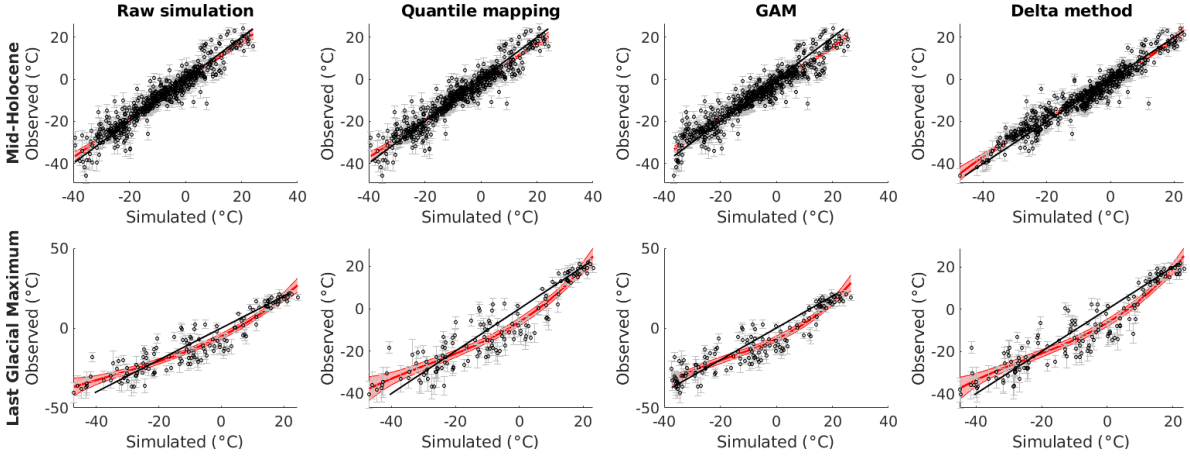
(b) Marine mean annual temperature



(c) Mean temperature of the warmest month



(d) Mean temperature of the coldest month



(e) Annual precipitation

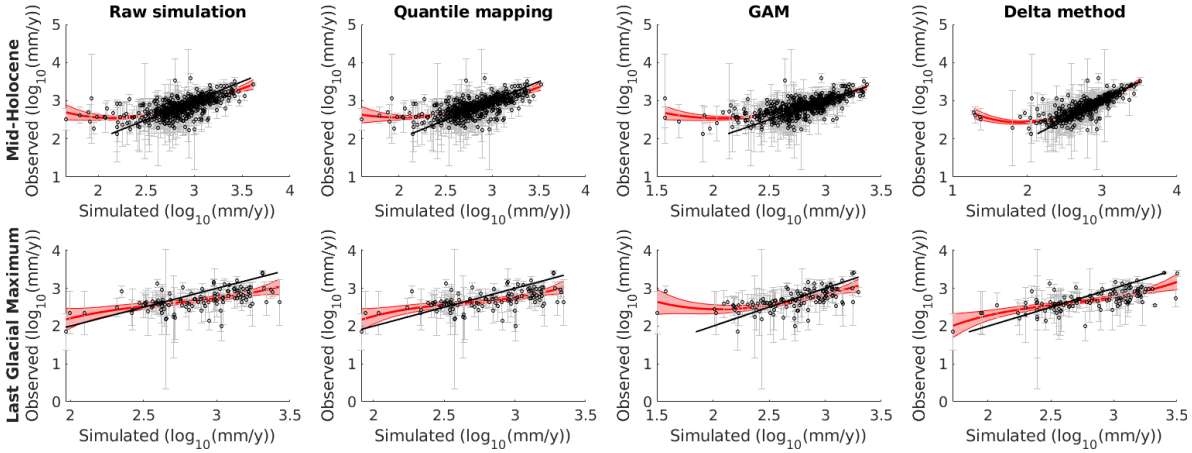


Figure 1. Comparison of bias-corrected simulated and empirically reconstructed climate variables. Black lines show 1:1 relationships. Red lines and shades show 5th degree polynomial regression and 95% confidence intervals, respectively.

annual temperature (in 93% and 89% of realisations, respectively), terrestrial mean temperature of the warmest month (in 92% and 100% of realisations, respectively), and precipitation (in 100% and 100% of realisations, respectively). The Delta Method also performed significantly better than the GAM method for Mid-Holocene terrestrial mean temperature of the coldest month (86% of realisations), and significantly better than Quantile ~~mapping~~ Mapping for LGM marine mean annual temperature (65% of realisations). The GAM method performed significantly better than Quantile ~~mapping~~ Mapping for LGM precipitation (100% of realisations). By ~~construction~~design, the Delta Method has a significantly lower median absolute bias (namely zero) than both other methods for all variables at present day.

~~On average~~Across time periods, raw simulations ~~underestimated~~ tended to underestimate terrestrial and marine mean annual temperature and terrestrial temperature of the warmest month, and overestimated annual precipitation ~~across time periods~~ (Fig. **??**). These trends are ~~as~~ less present in the bias-corrected data~~. Indeed,~~ : methods consistently reduced the absolute value of the median bias (*MB* in Eq. (**??**)) of the raw simulations, except in the case of terrestrial temperature of the coldest month.

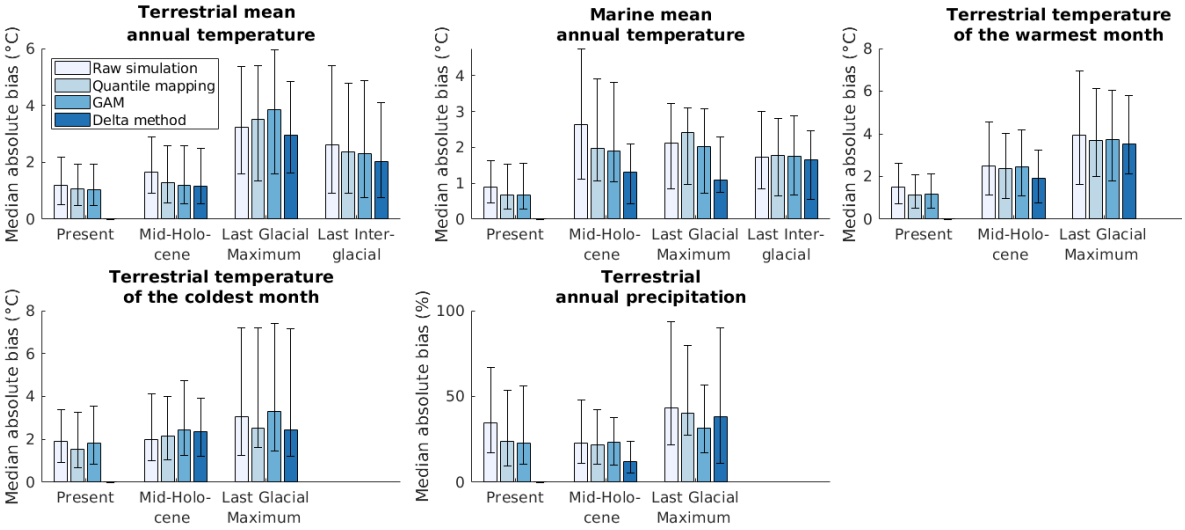


Figure 2. Median absolute biases (*MAB*, Eq. (**??**)) of the raw and bias-corrected climate simulation data. Error bars represent 25% and 75% weighted quantiles of the local absolute biases available for the given climatic variable and point in time.

The differences ~~in how raw and bias-corrected simulation outputs improve the representation of~~ between bias correction methods in terms of improving the climate change signal (*CCMAB* in Eq. (**??**)) are negligible in all scenarios except for marine mean annual temperature during the Last Glacial Maximum, where the GAM method performs slightly better than other methods (Fig. **??**).

The performance of the different methods is not uniform across space nor time. Fig. **??** illustrates this heterogeneity for the Delta Method. For example, the Delta Method significantly ~~reduced~~ reduces the original bias of modelled precipitation in Eastern North America in the Mid-Holocene, but hardly improved the raw simulations in the Sahara, whereas the opposite pattern can be observed at the LGM.

The performances of the methods relative to each other also ~~varied significantly~~ vary substantially across both space and time. For example, ~~while~~ whilst globally the Delta Method has a slight overall edge over the GAM approach (Fig. **??**), the comparison of the two methods in Fig. **??** shows that even within small geographical regions neither method performs consistently better than the other. Moreover, a better performance of one method in a ~~specific location at a certain~~ certain location at some point in time generally does not guarantee

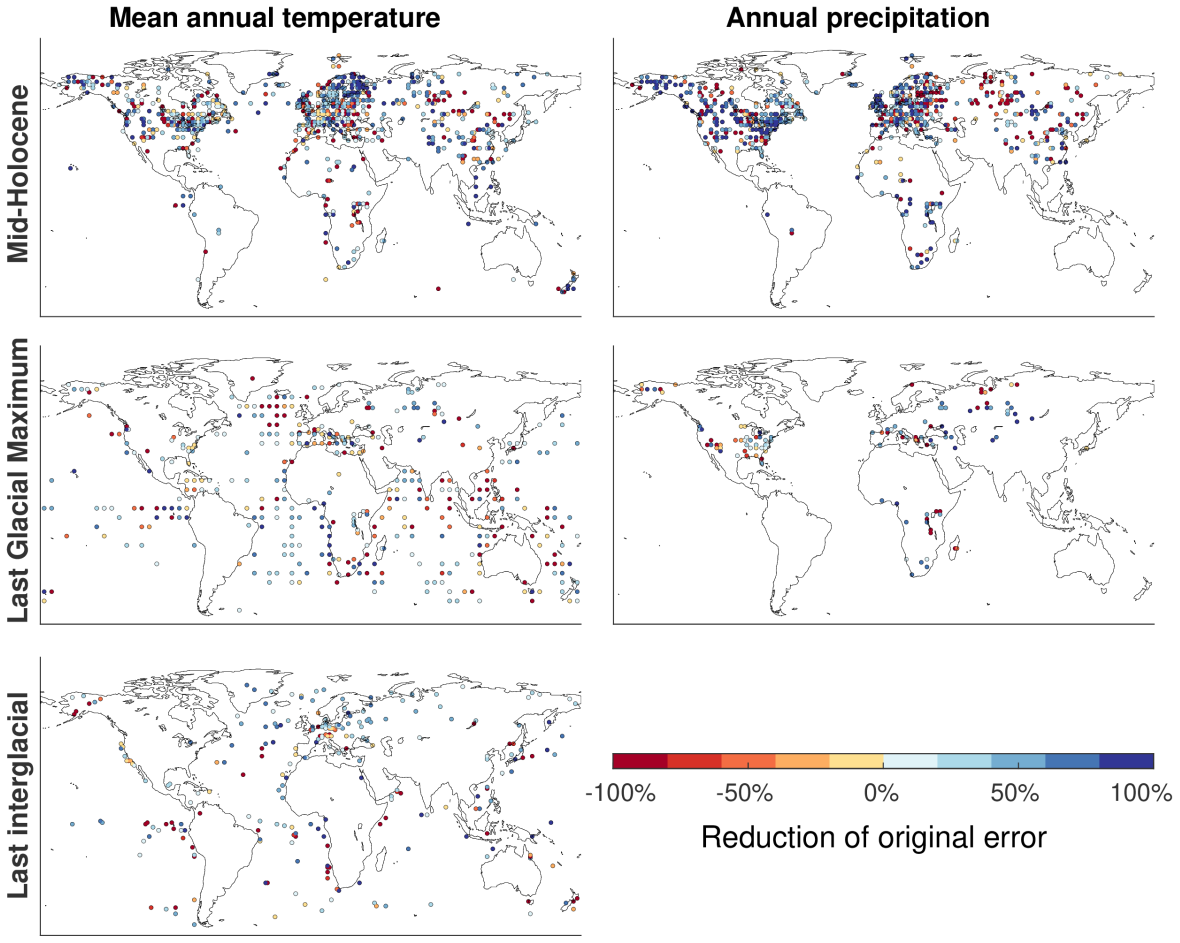


Figure 3. Reduction of the original model bias by the Delta Method for terrestrial and marine mean annual temperature and terrestrial annual precipitation. The lower end of the colour scale was capped at -100% (i.e. a doubling of the original bias).

the same result at a different time. For instance, the Delta Method overall reduced the original bias of modelled precipitation more than the GAM approach in Eastern North America during the Mid-Holocene, but less during the LGM (Fig. ??).

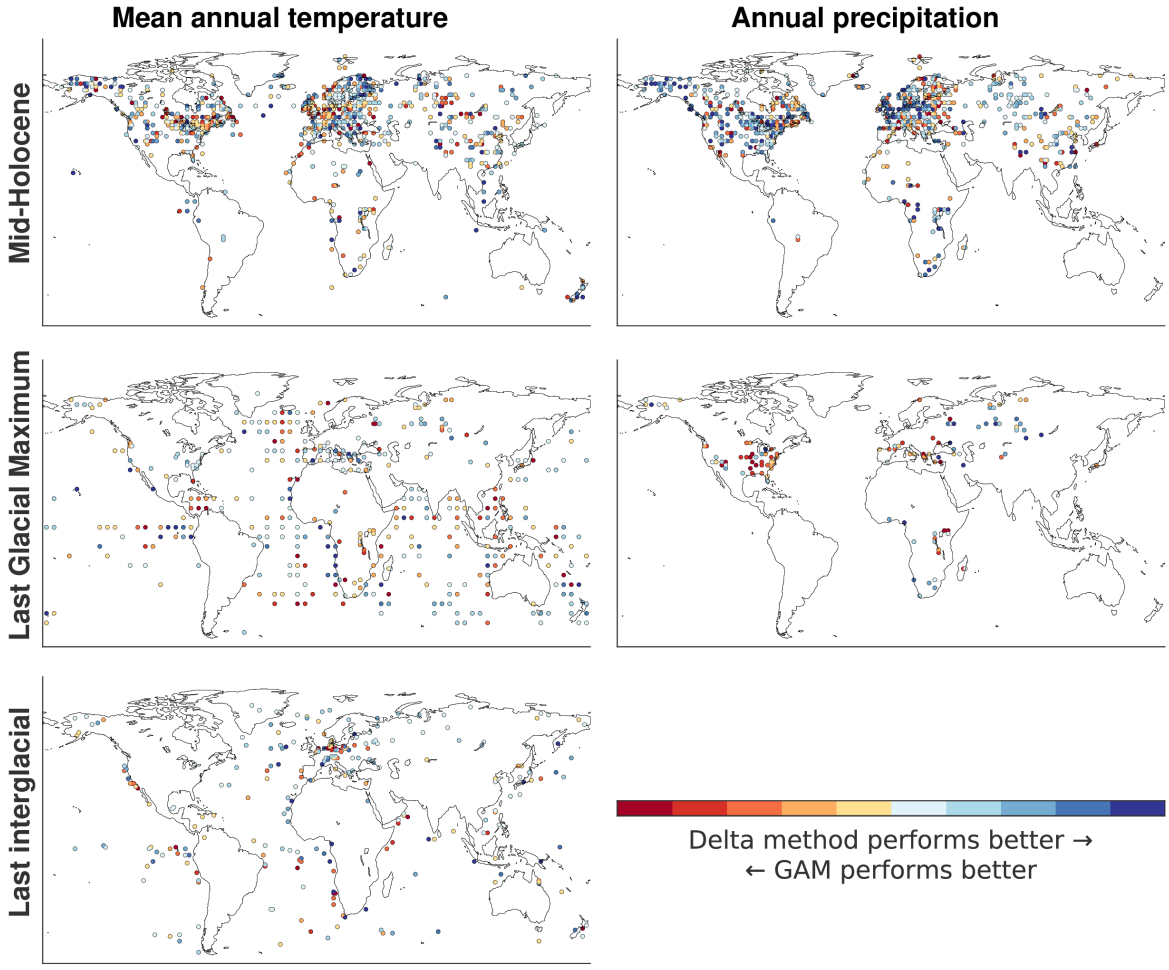


Figure 4. Relative performances of the Delta Method and the GAM approach in terms of debiasing simulated mean annual temperature (left column) and annual precipitation (right column). The colour spectrum represents the interval [0,1], and marker colours are calculated as the ratio of the absolute value of the local bias (Eq. (**??**)) of the GAM-based approach divided by the sum of the absolute local biases of both methods.

4 Discussion

Whilst, overall, the Delta Method performs slightly better at debiasing temperature and precipitation compared to ~~GAMs and Quantile mapping~~ the GAM-based method and Quantile Mapping for the empirical data considered here, we note that this method is only appropriate for a given land conformation. Thus, it is only ~~appropriate~~ suitable for the Late Quaternary, and even for this period, changes in sea levels are problematic as they expose areas for which ~~we have~~ there is no bias information~~as well as changing the areas affected by maritime climate~~. GAMs should, in theory, obviate ~~these problems by quantifying local~~ this problem by quantifying bias-related processes as statistical relationships~~with appropriate proxies. Whilst~~ ; however, whilst this approach might be the only option for the deeper past, our results point to the fact that ~~reconstructing such local~~ estimating such processes in such a way is challenging, as demonstrated by its overall inferior performance to the Delta Method. A possible limitation of GAMs as currently applied ~~to bias correction and downscaling~~ is that they assume additivity ~~, thus estimating the effect of given proxies for the prevaling climate state observed at present day~~between predictor variables. By fitting interactions, it would be possible to allow for ~~these effects to differ depending on the local climatic conditions~~more complex processes, but the computational complexitiy of interactions with such large datasets is non-trivial.

A major limitation of current approaches ~~to debias~~ for bias-correcting climate model data is that they all assume ~~biases~~ bias patterns in present-day climate to be fully representative of the past (see section **??**). With the progressive increase in the number of empirical ~~reconstructions~~ records of past climatic conditions, it ~~might~~ may be possible to soon move from a situation where past ~~data~~ reconstructions are use to verify ~~correction schemes~~ bias correction methods (as we did in this manuscript) to using those data to actively calibrate ~~the bias correction function.~~ bias correction methods. Despite large uncertainties, and patterns that are not fully consistent across time, Fig. **??** suggests an intriguing relationship between the ~~temporal variation of the local model bias and the~~ variation of local model biases across time on the one hand, and simulated climate change ~~signal of the variable of interest~~signals on the other hand. Such a ~~statistical~~ relationship could, in principle, be used to refine the Delta Method by accounting for the change ~~in local model bias with time. However, uncertainties are large, patterns do not seem fully consistent across time, and available data points do not represent the world uniformly~~of local

model biases over time. For example, instead of Eq. (**??**), we would have

$$T_{\mathrm{sim}}^{DM+}(x,t) := T_{\mathrm{sim}}^{raw}(x,t) + \underbrace{\left(T_{\mathrm{emp}}(x,0) - T_{\mathrm{sim}}^{raw}(x,0)\right)}_{\substack{\text{standard time-invariant Delta}\\ \text{Method bias correction term}}} + \underbrace{f\Big(\underbrace{T_{\mathrm{sim}}^{raw}(x,t) - T_{\mathrm{sim}}^{raw}(x,0)}_{\substack{\text{simulated climate}\\ \text{change signal}}}, \underbrace{\ldots}_{\substack{\text{additional}\\ \text{predictor}\\ \text{variables}}}\Big)}_{\text{time-variable correction term}}, \tag{11}$$

where f represents a non-linear regression model satisfying $f|_{t=0} = 0$. A robust statistical model f will require not only additional data from currently underrepresented geographical areas (specifically the southern hemisphere), but also ~~curating~~ the curation of empirical reconstructions, as successfully done for the last millenium (**??**).

Such an approach would tie in with data assimilation methods, which also use empirical climate proxy records to improve climate simulations. These methods have been used to estimate global climate variables at times at which the quantity and spatial coverage of available empirical records is high enough to allow a robust calibration of the relevant computational methods. As a result, they have focussed either on single points in the past, such as the Mid-Holocene (**?**) or Last Glacial Maximum (**?**), or on time intervals across which suitable empirical data are available, namely the last millennium period (**??**). In contrast to the aforementioned approaches, based on Fig. **??** we suggest that it may be possible to use empirical reconstructions even from only a small set of points in time (e.g. the present, Mid-Holocene, LGM and Last Interglacial Period) to parameterise a statistical model of the temporal variation of local biases that could be used to improve simulated data at any time point in the Late Pleistocene-Holocene period.

5 Conclusions

Our comparison of global debiased palaeosimulation data and empirical reconstructions suggests that, ~~overall, the Delta Method provides slightly better performance at debiasing compared to GAMs and Quantile Mapping – though not in all cases to at a statistically significanty extent. Given our results, we suggest that~~ despite its conceptual simplicity, the Delta Method is good starting point for bias ~~removal~~ correction of simulated Late Quaternary climate data at a global scale~~.~~, providing slightly stronger bias reductions compared to GAMs and Quantile Mapping. However, given the lack of statistical significance of the superior performance in some cases, and the considerable variability in the effectiveness of the ~~different methods in different locations~~ three methods

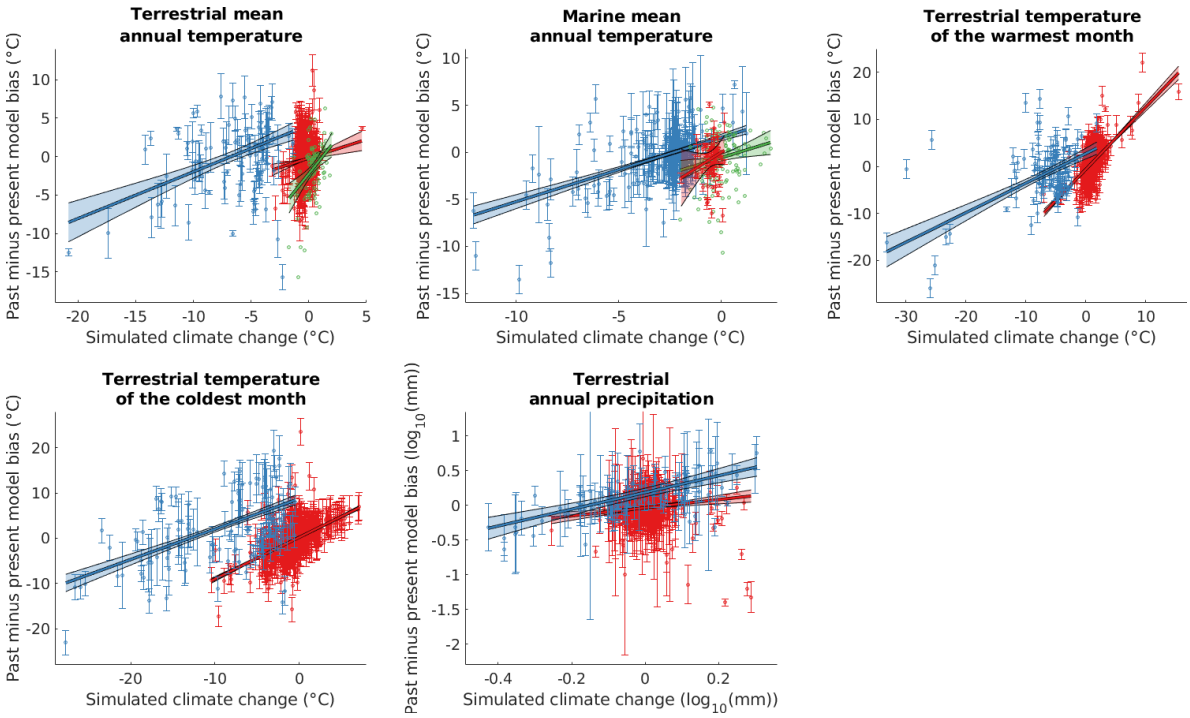


Figure 5. Differences between local past and present model bias (at locations for which empirical reconstructions are available) against the local simulated climate change signal (i.e. the difference between past and present simulated value) of the variable of interest. Red, blue and green markers represent data from the Mid-Holocene, the LGM and the Last Interglacial Period, respectively. Error bars represent standard errors of the empirical reconstructions. Lines and shades show robust linear regressions and 95% confidence intervals, respectively. Whilst weak, the relationships suggest that it may be possible to model some of the variability of local model biases over time, using only available simulation data. Such an approach could potentially significantly enhance the Delta Method, which currently operates on the simplifying assumption that this variability is negligible.

across different locations and points in time, we echo earlier propositions that studies focussing on specific
410 regions require case-by-case assessments of which ~~bias-correction~~ bias correction method is most suitable for improving palaeoclimate ~~model outputs (?).~~ simulations (?). We also reiterate that our results may be different for palaeoclimate simulations other than the ones used here. Finally, it is important to bear in mind that bias

correction methods are unable to substantially correct a fundamentally poor climate model, e.g. with strong circulation biases, which such methods are not capable of removing (**?**). Seeking to improve the representation of climate dynamics in simulation models therefore remains a priority alongside the development of bias correction methods.

~~Whilst the datasets used in this paper are a step in the right direction, they are still too sparse and diverse for the purpose of actively parameterising debiasing functions. Such a resource would arguably allow bias removal methods to greatly improve in their effectivness~~A key limitation of all three methods considered here is their assumption that present-day patterns between simulated and observed climate can be extrapolated to estimate model biases in the past. High uncertainties, and the spatial and temporal sparseness associated with currently available empirical palaeoclimate datasets will likely impede a robust assimilation of these data into bias correction methods at this stage; however, our data indicate the increasing quantity and quality of global proxy records could soon make it possible to use empirical reconstructions in the development of improved methods that effectively account for the variation of local model biases through time.

Code and data availability. Code and datasets used in this analysis will be made publicly available on the Open Science Framework repository upon acceptance of the manuscript.

Author contributions. All authors conceived the study. R.B. conducted the analysis and wrote the manuscript. All authors interpreted the results and revised the manuscript.

Competing interests. The authors declare no competing interests.

Acknowledgements. The authors are grateful to Paul J. Valdes and Joy S. Singarayer for providing the climate simulation data used in this study, and to three anonymous reviewers for their helpful comments. R.B., M.K. and A.M. were supported by the ERC Consolidator Grant 647787 (”LocalAdaptation”).

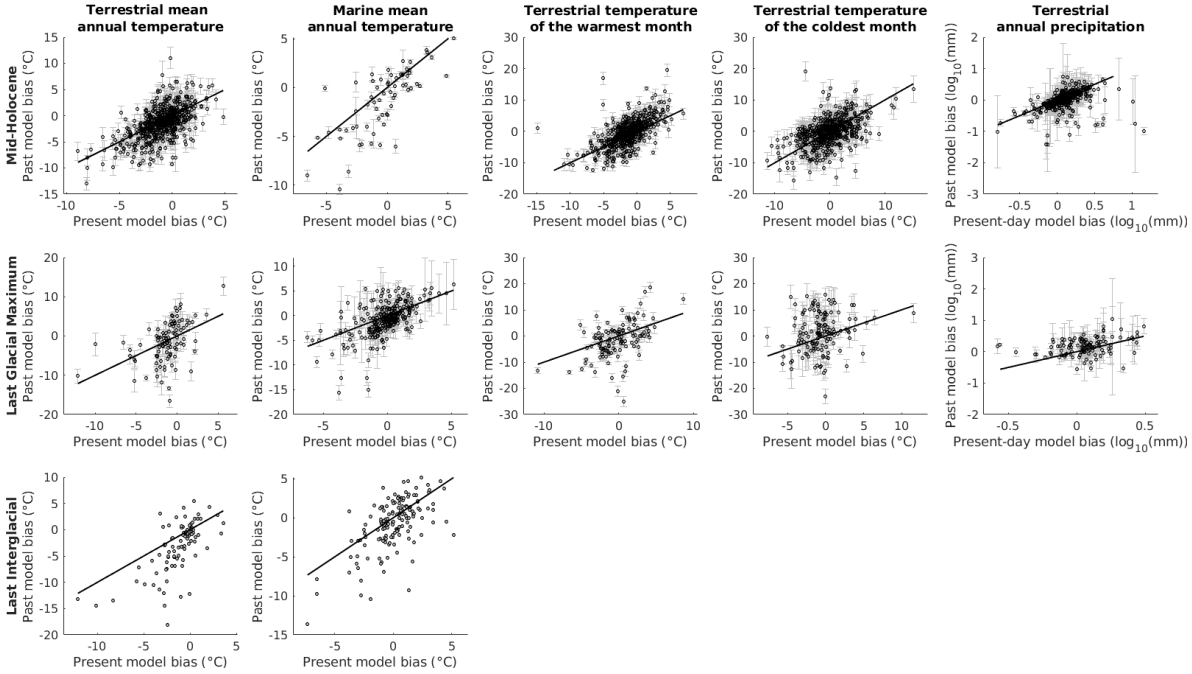


Figure A1. Comparison of present-day and past model biases (which the Delta Method assumes to be similar) from locations where empirical reconstructions are available. Lines represent 1:1 relationships.

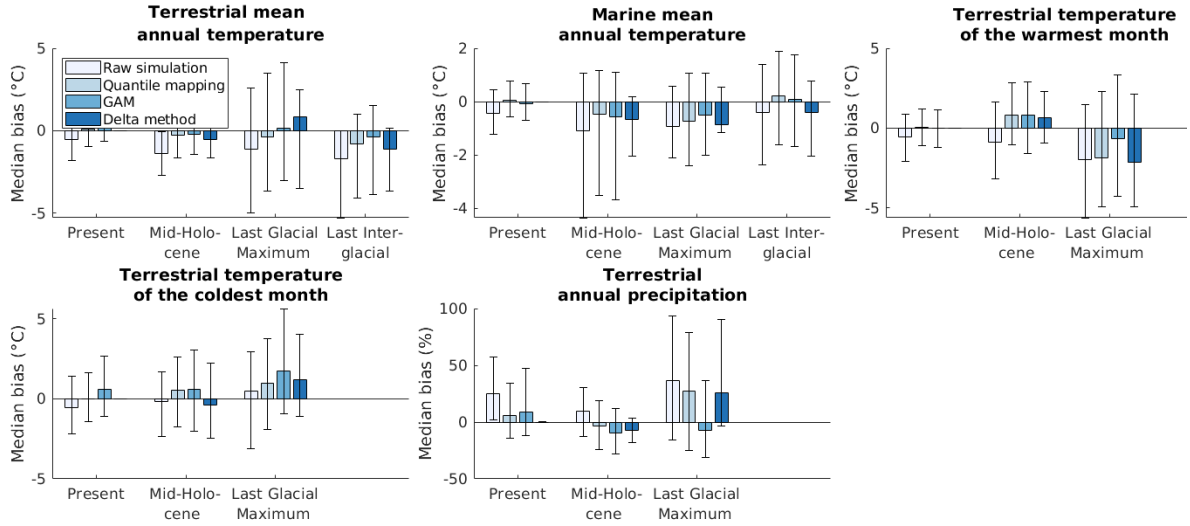


Figure A2. Median biases of the raw and bias-corrected climate simulation data. Error bars represent 25% and 75% weighted quantiles of the local biases available for the given climatic variable and point in time.

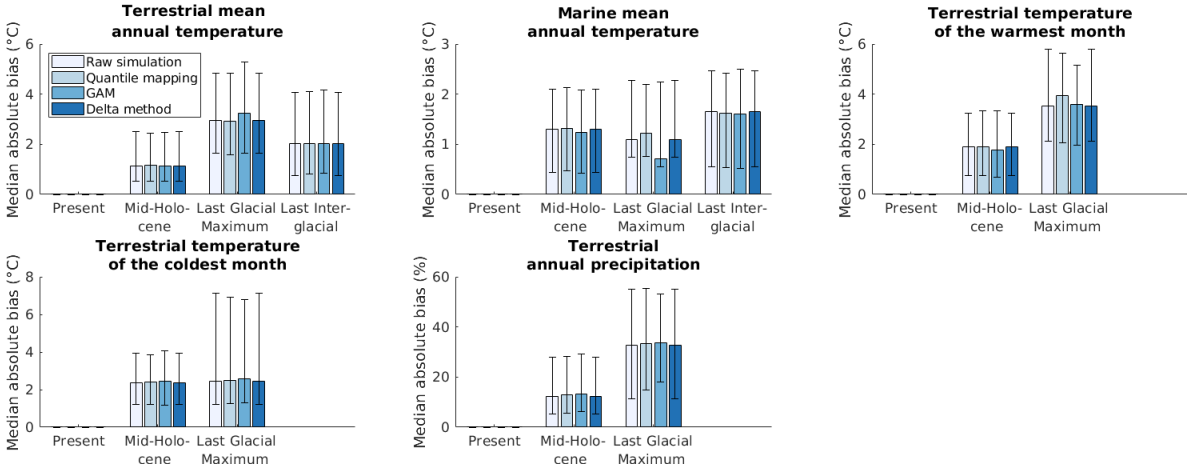


Figure A3. Median absolute biases of the climate change signal (*CCMAB*, Eq. (**??**)). Error bars represent 25% and 75% weighted quantiles of the local absolute climate change biases available for the given climatic variable and point in time.