A comparative 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 the climate system. Correcting model biases is therefore an essential step toward realistic palaeoclimatologies, which are crucial for numerous applcations, 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). 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 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.

1 Introduction

Realistic reconstructions of global palaeoclimate are a key requirement for modelling many important long-term and large-scale ecological processes (Eriksson et al., 2012; Timmermann and Friedrich, 2016; Leonardi et al., 2018; Zhu et al., 2018; Rangel et al., 2018; Beyer et al., 2020). 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 (Solomon et al., 2007; Ehret et al., 2012). Depending on

the region of interest, these biases can be of the order of a few degrees of temperature, or centimeters of annual 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) (Kottek et al., 2006).

Bias correction has received a great deal of attention for present-day and near-future simulations (Ho et al., 2011; Maraun and Widmann, 2018), whereas work on palaeoclimate reconstructions has been much more limited. This is partly due to the different time scale of palaeoecological 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 (Maraun et al., 2017), 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 (Maraun et al., 2017). 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: the Delta Method (http://www.worldclim.org/downscaling), statistical methods based on generalised additive models (GAMs) (Vrac et al., 2007; Levavasseur et al., 2011; Woillez et al., 2014; Latombe et al., 2018) and Quantile Mapping (Lorenz et al., 2016). 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 location-specific (Maraun and Widmann, 2018), as it is based on a map of differences between observed and simulated values. GAMs attempt to represent statistical relationships between simulated climatic variables (as well as other known physical variables, such as elevation and the distance from the coast) and bias-corrected climatic variables (Vrac et al., 2007; Maraun and Widmann, 2018). Finally, Quantile Mapping assumes that biases are specific to their respective quantiles in the distribution of the relevant climatic variable (Maraun and Widmann, 2018). 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 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 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 focus on the global performance of the different methods; however, we note that bias-correction is not a one-size-fits-all approach (Maraun and Widmann, 2018), and that our results do not remove the need for local re-evaluations of methods in specific continental and subcontinental regions of interest.

Section 2 provides details of the three bias correction methods, the climate simulations, and the empirical palaeoclimatology reconstructions used in this study. In section 3, we quantitatively assess the performance of the methods at a global scale, and with regard to spatial and temporal heterogeneities. Section 4 discusses how palaeoclimate reconstructions could be used not only to evaluate methods, but to help estimate the variation of local model bias 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 $1.25° \times 0.83°$ grid resolution 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 (Valdes et al., 2017). For the Last Interglacial Period, we do not have simulation data from HadAM3H, but we used the global climate model emulator GCMET (Krapp et al., 2019) that is based on the same model and can make predictions at the same spatial resolution. We note that the results presented in this article are specific to the particular climate simulations considered here, and do not claim generalisability to other models.

Empirical data (see section 2.1.2) of terrestrial temperature were compared against simulated temperature at 1.5 meters height, whereas 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 present-day observational data. For this, we used monthly terrestrial temperature and precipitation data at a 0.167° grid resolution (New et al., 2002), and mean annual sea surface temperature at a 1° grid resolution (Reynolds et al., 2002), representative of 1960–1990. These maps were remapped to the 1.25°×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 reconstructions of terrestrial mean annual temperature, temperature of the coldest and warmest month, and annual precipitation, for the Mid-Holocene and the LGM from Bartlein et al. (2011), reconstructions of mean annual sea surface temperature for the Mid-Holocene and the LGM from Hessler et al. (2014) and Waelbroeck et al. (2009), respectively, and reconstructions of mean annual terrestrial and sea surface temperature for the Last Interglacial Period from Turney and Jones (2010). 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 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°×0.8° grid cell of our palaeoclimate simulations (see section 2.1.1) 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 1.25°×0.8° grid cell containing the sample location. Figs. 3 and 4 visualise the locations of all reconstructions of terrestrial and marine mean annual temperature, and annual precipitation.

2.2 Bias correction methods

2.2.1 The Delta Method

The Delta Method is based on adding the difference between past and present-day simulated climate (the 'delta') to present-day observed climate. Thus, the Delta Method assumes that the local (i.e. grid cell-specific) model

bias is constant over time (Maraun and Widmann, 2018). 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, $T_{\text{emp}}(x,0) - T_{\text{sim}}^{raw}(x,0)$. Debiased temperature, $T_{\text{sim}}^{DM}(x,t)$, in location x at some time t is obtained as:

$$\begin{aligned} T_{\text{sim}}^{DM}(x,t) &:= T_{\text{emp}}(x,0) + \big(T_{\text{sim}}^{raw}(x,t) - T_{\text{sim}}^{raw}(x,0)\big) \\ &= T_{\text{sim}}^{raw}(x,t) + \big(T_{\text{emp}}(x,0) - T_{\text{sim}}^{raw}(x,0)\big). \end{aligned} \tag{1}$$

The second expression illustrates that $T_{\text{sim}}^{DM}(x,t)$ is alternatively given by adding the present-day local bias to the simulated temperature, $T_{\text{sim}}^{raw}(x,t)$, at 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 when applying the Delta Method for precipitation, which corresponds to applying the simulated relative change to the observations (Maraun and Widmann, 2018). Analogously to temperature, debiased precipitation is given by

$$\begin{aligned} P_{\text{sim}}^{DM}(x,t) &:= P_{\text{obs}}(x,0) \cdot \frac{P_{\text{sim}}^{raw}(x,t)}{P_{\text{sim}}^{raw}(x,0)} \\ &= P_{\text{sim}}^{raw}(x,t) \cdot \frac{P_{\text{obs}}(x,0)}{P_{\text{sim}}^{raw}(x,0)}. \end{aligned} \tag{2}$$

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) (Vrac et al., 2007; Maraun and Widmann, 2018). Transfer functions representing this relationship are calibrated on the basis of present-day simulated and observed climate, and are then used to derive past climate using the appropriate simulated data. Generalised additive models (GAMs) have gained particular popularity as transfer functions (Vrac et al., 2007; Levavasseur et al., 2011; Woillez et al., 2014; Latombe et al., 2018). They accommodate potential nonlinearities in the response of the individual variables, while – owing to the computational requirements of general high-dimensional nonlinear regressions – assuming that the interactions between predictor variables can be neglected.

For a set of 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 the location x_i. Here, the $x_1, x_2, \ldots$ represent land and ocean points on the $1.25° \times 0.8°$ grid of the climate data (see section 2.1) in the case of terrestrial and marine climate variables, respectively. In a GAM, these 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 X_j^V are time-dependent; not only when they represent climate model outputs, but also when they represent elevation or the distance to the ocean, which vary over time as the result of sea level changes. Finally, the GAM is given by the regression

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

where the $f_1, f_2 \ldots$ represent smooth functions that are fitted to minimise the distance between the left and the right hand side in Eq. (3). Once the model has been calibrated on the present-day data, it can be used to estimate the true (i.e. bias-corrected) values of the climate variable of interest in the location x_i at some point in past t as

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

Similar to Latombe et al. (2018), 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, the shortest distance to the ocean, and simulated precipitation, temperature, (absolute) wind speed, air pressure and relative humidity 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 Wood (2004) 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 values that is specific to the quantile associated with the relevant value within the set 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, the cumulative distribution functions of present-day observed and present-day simulated data that was bias-corrected using Quantile Mapping are identical.

For a climate variable V, we denote by $F^{V}_{\text{emp}}[0]$ the cumulative distribution function of the present-day empirical observations $V_{\text{emp}}(x_1,0), V_{\text{emp}}(x_2,0),\dots$ (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 $V^{raw}_{\text{sim}}(x_1,t), V^{raw}_{\text{emp}}(x_2,t),\dots$ 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.

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

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

for the bias-corrected value at time t and location x_i.

2.2.4 Method discussion

All three bias correction methods considered here aim at minimising biases in past simulated data, but they make different assumptions as to how this aim can best be achieved. The Delta Method assumes that the (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 present-day model biases across grid cells via the predictor variables. As a result (and unlike in the case of the Delta Method), GAM-corrected present-day simulations are not identical to the present-day observed climate. This drawback is accepted under the assumption that the derived statistical model captures the *mechanisms* underlying local model biases better than the time-constant local correction term used in the Delta Method, and indeed to an extent that allows better 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.

2.3 Method evaluation

In ecological applications, the objective of applying a bias-correction method to past simulated climate data is generally to reduce the difference between the simulated and the (generally unknown) true past climate. Empirical palaeoclimatic reconstructions allow us to assess these differences at specific locations and points in time. Here, we determine local differences between empirical reconstructions and bias-corrected simulations for each climate variable and bias-correction method, and define a spatially aggregated measures 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 marine mean annual temperature, temperature of the coldest or warmest month, or annual precipitation, respectively. For $M \in \{\text{raw}, \text{DM}, \text{GAM}, \text{QM}\}$, we denote by $V_{\text{sim}}^{M}(x,t)$ the simulated value of the climate variable V at time t in 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. The local bias between the empirically reconstructed and the simulation data at time t in the location x is then given by

$$B_V^M(x,t) = \begin{cases} V_{\text{sim}}^{M}(x,t) - V_{\text{emp}}(x,t) & \text{if } V \in \{T^{\text{ter.mean}}, T^{\text{Tmar.mean}}, T^{\text{cold}}, T^{\text{warm}}\} \\ \dfrac{V_{\text{sim}}^{M}(x,t) - V_{\text{emp}}(x,t)}{V_{\text{emp}}(x,t)} & \text{if } V = P^{\text{ann}} \end{cases} \tag{6}$$

We use 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 at time t for climate variable V. In section 3, we provide complete plots of the distribution of the biases corresponding to each specific climate variable, point in time, and bias correction method. Furthermore, as a summary statistic of these distributions, and an aggregated measure for evaluating and comparing the performance of the three bias correction methods, we use 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 by $\{\omega_{\text{emp}}(x_i^{(t,V)}, t)\}_{i=1,2,\ldots}$,

rescaled such that $\sum_i \omega_{\text{emp}}(x_i^{(t,V)}, t) = 1$. The median absolute bias for variable V and bias-correction method M at time t is then formally given by

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

Thus, the median absolute bias is a measure of the average difference between empirical and the bias-corrected simulated data. We consider 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 *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 climatic signals that are not captured by the empirical data used here.

We tested whether the median absolute biases associated with any two 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 simulated values obtained after applying the different bias-correction methods. For each of these 10^4 sets of 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 method was significantly smaller than that associated with a different 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 method was found to perform significantly better than another one in more than half of the realisations, we report this result in section 3.

Debiased simulated data should ideally not contain any systematic bias, in that the median bias,

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

(where the weighted median is calculate analogously as in Eq. (7)) should not differ substantially from zero. In addition to considering the median absolute bias, here, we also examine how different methods affect the associated median bias.

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 of a climate variable V bias-corrected using method M at a location x and between the present and time t in the past is given by

$$CCB_V^M(x_i^{(t,V)}, t) = \begin{cases} \big(V_{\text{sim}}^M(x,t) - V_{\text{sim}}^M(x,0)\big) - \big(V_{\text{emp}}(x,t) - V_{\text{emp}}(x,0)\big) \\ \qquad\qquad \text{if } V \in \{T^{\text{ter.mean}}, T^{\text{Tmar.mean}}, T^{\text{cold}}, T^{\text{warm}}\} \\ \dfrac{V_{\text{sim}}^M(x,t) - V_{\text{sim}}^M(x,0)}{V_{\text{sim}}^M(x,0)} - \dfrac{V_{\text{emp}}^M(x,t) - V_{\text{emp}}(x,0)}{V_{\text{emp}}(x,0)} \\ \qquad\qquad \text{if } V = P^{\text{ann}}, \end{cases} \tag{9}$$

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

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

where the weighted median is calculated analogously as in Eq. (6). Here, we also compare the performance of the different bias correction methods for this quantity. We did not determine the median absolute bias for the climate change signal between points in the past, due to the much smaller number of empirical reconstructions that are available from the same location across time, 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.

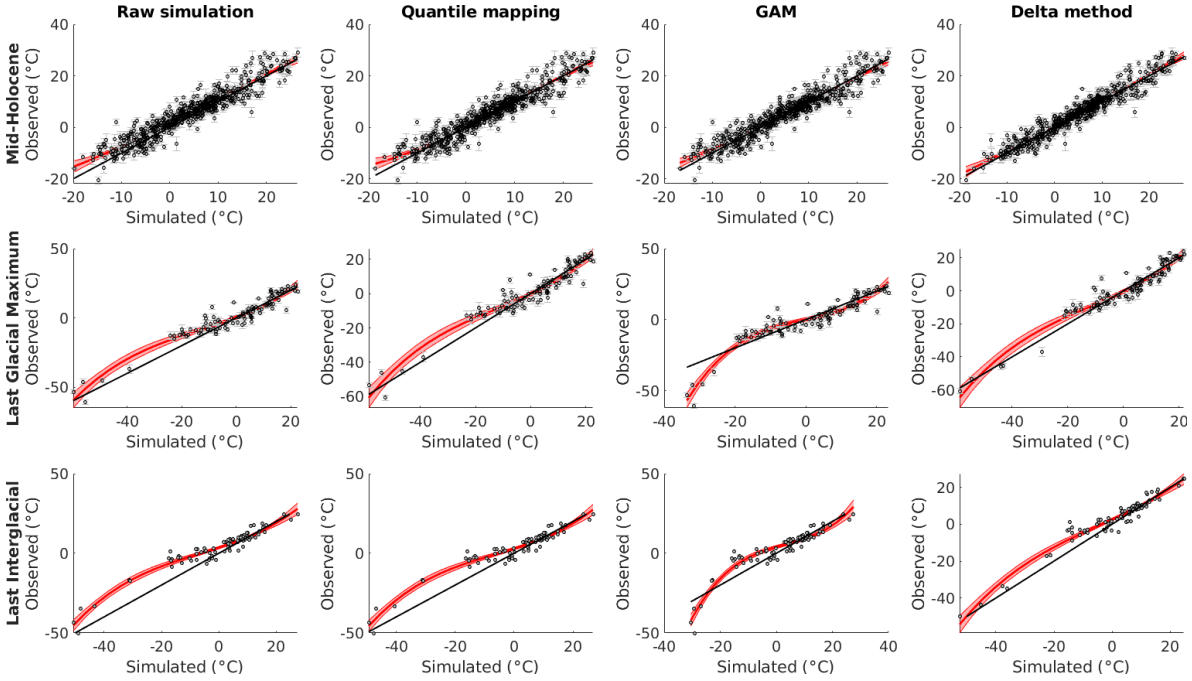
3 Results

Fig. 1a–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 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 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 methods, an analogous patterns can be observed in the case of precipitation.

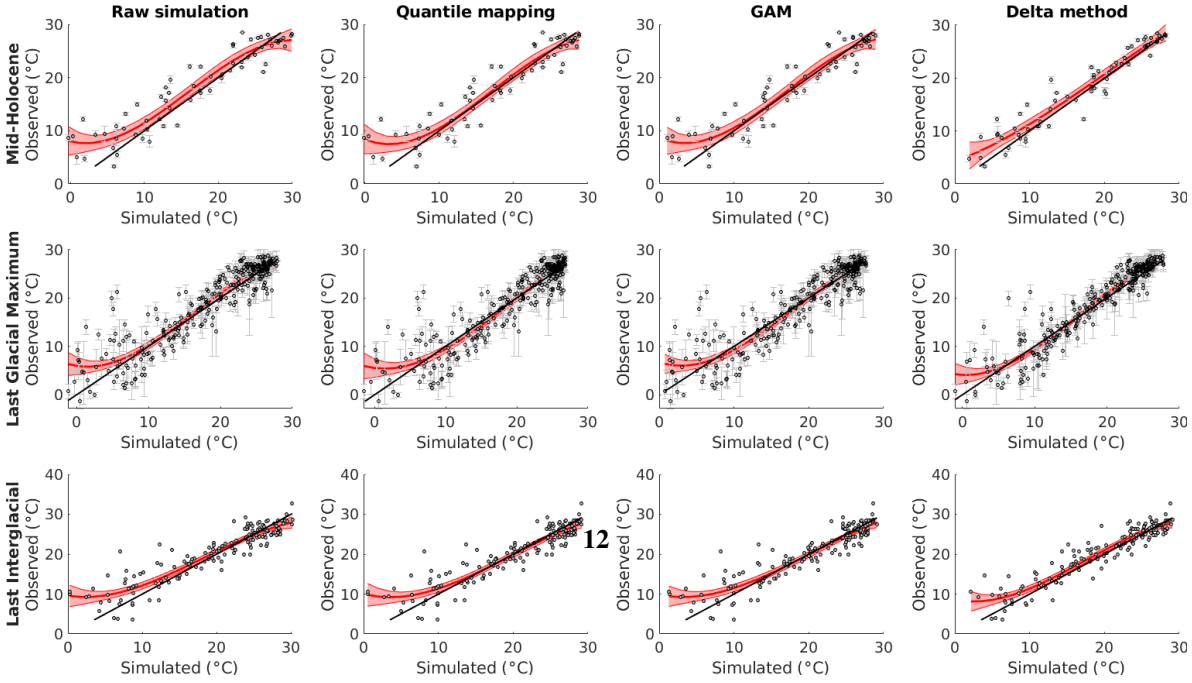
All bias-correction methods reduce the median absolute bias (*MAB* in Eq. (6)) of present-day simulated data for all climate variables – as would be expected (Maraun and Widmann, 2018) – although, by construction, only the Delta Method completely eliminates all differences between simulated and observed data (Fig. 2). The Delta Method also provides the strongest reduction in median absolute bias (*MAB* in Eq. (6)) 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. 2). The comparatively good performance of the Delta Method is also reflected in the correlations between present-day and past model biases, which the Delta Method assumes to be similar (Fig. A1). The GAM method and Quantile Mapping generally lead to a reduction in bias, even though overall not as effectively as the Delta Method. In a few cases, the original bias is actually increased after applying a correction method (Fig. 2).

These trends in the performances of the different 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 2.3) when compared against Quantile Mapping and the GAM method, respectively), marine mean 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 for LGM marine mean annual temperature (65% of realisations). The GAM method performed significantly better than Quantile

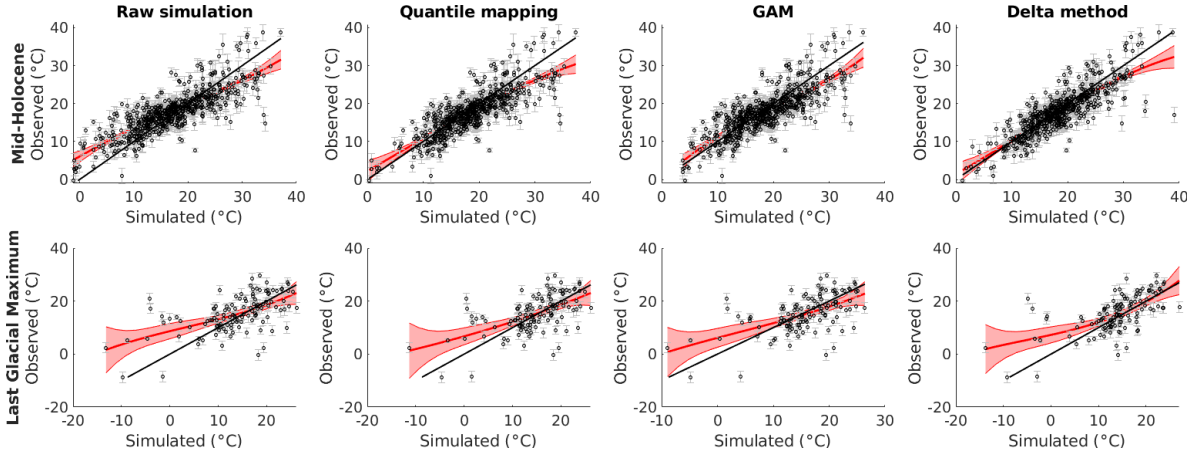
(a) Terrestrial mean annual temperature



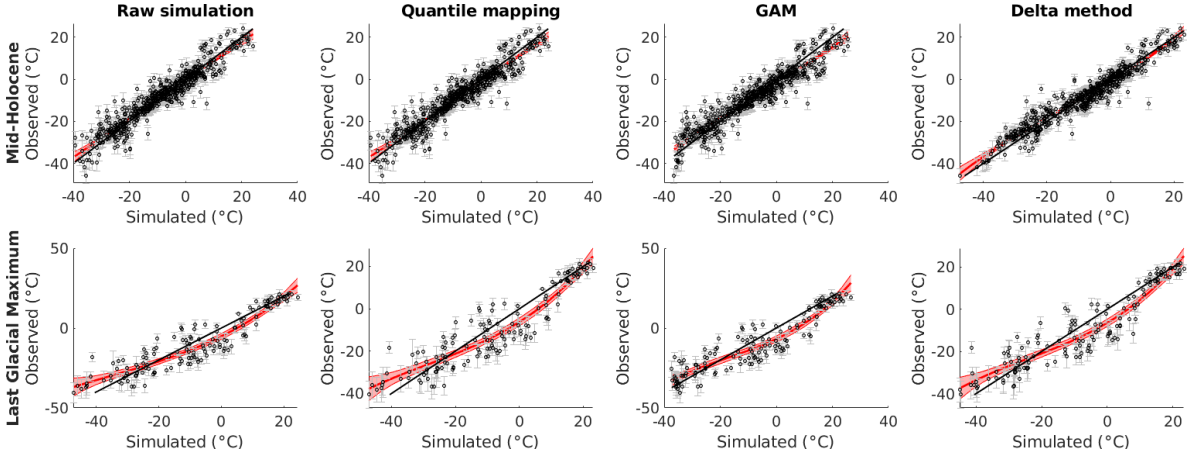
(b) Marine mean annual temperature



(c) Mean temperature of the warmest month



(d) Mean temperature of the coldest month



(e) Annual precipitation

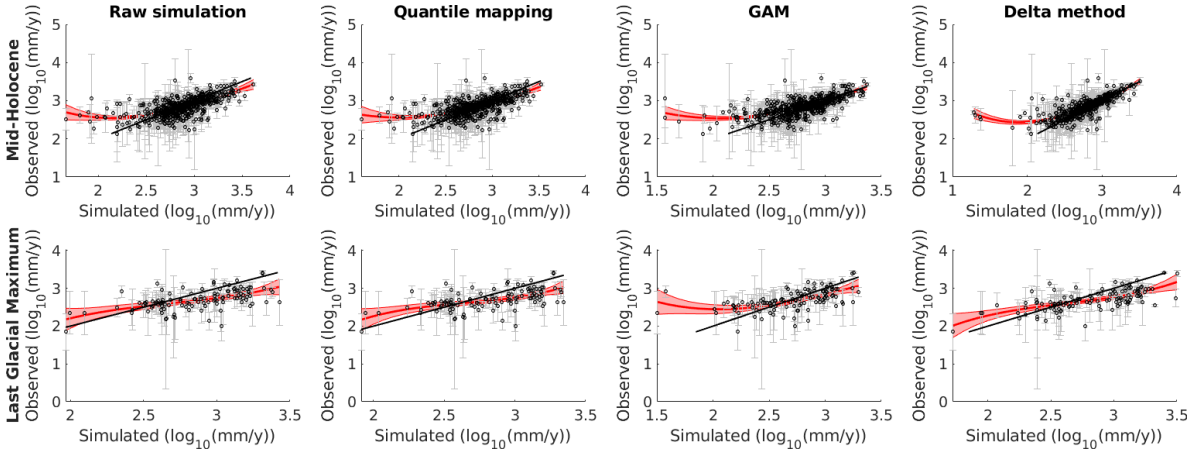


Figure 1. Comparison of bias-corrected simulated and empirically reconstructed climate variables. Black lines show 1:1 relationships. Red lines and shades show 5th degree polynomial regression and 95% confidence intervals, respectively.

mapping for LGM precipitation (100% of realisations). By construction, the Delta Method has a significantly lower median absolute bias (namely zero) than both other methods for all variables at present day.

On average, raw simulations underestimated terrestrial and marine mean annual temperature and terrestrial temperature of the warmest month, and overestimated annual precipitation across time periods (Fig. A2). These trends are as present in the bias-corrected data. Indeed, methods consistently reduced the absolute value of the median bias (*MB* in Eq. (8)) of the raw simulations, except in the case of terrestrial temperature of the coldest month.

The differences in how raw and bias-corrected simulation outputs improve the representation of the climate change signal (*CCMAB* in Eq. (10)) 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. A3).

The performance of the different methods is not uniform across space nor time. Fig. 3 illustrates this heterogeneity for the Delta Method. For example, the Delta Method significantly reduced the original bias of modelled

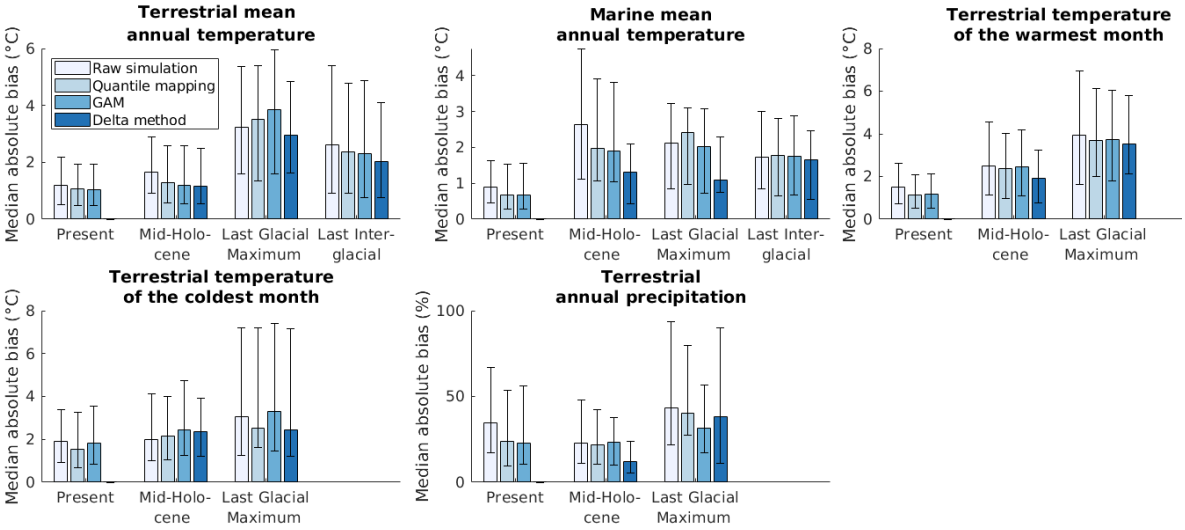


Figure 2. Median absolute biases (*MAB*, Eq. (7)) of the raw and bias-corrected climate simulation data. Error bars represent 25% and 75% weighted quantiles of the local absolute biases available for the given climatic variable and point in time.

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 across both space and time. For example, while the Delta Method has a slight overall edge over the GAM approach, the comparison of the two methods in Fig. 4 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 point in time generally does not guarantee 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.

4 Discussion

Whilst, overall, the Delta Method performs slightly better at debiasing temperature and precipitation compared to GAMs and Quantile mapping for the empirical data considered here, we note that this method is only ap-

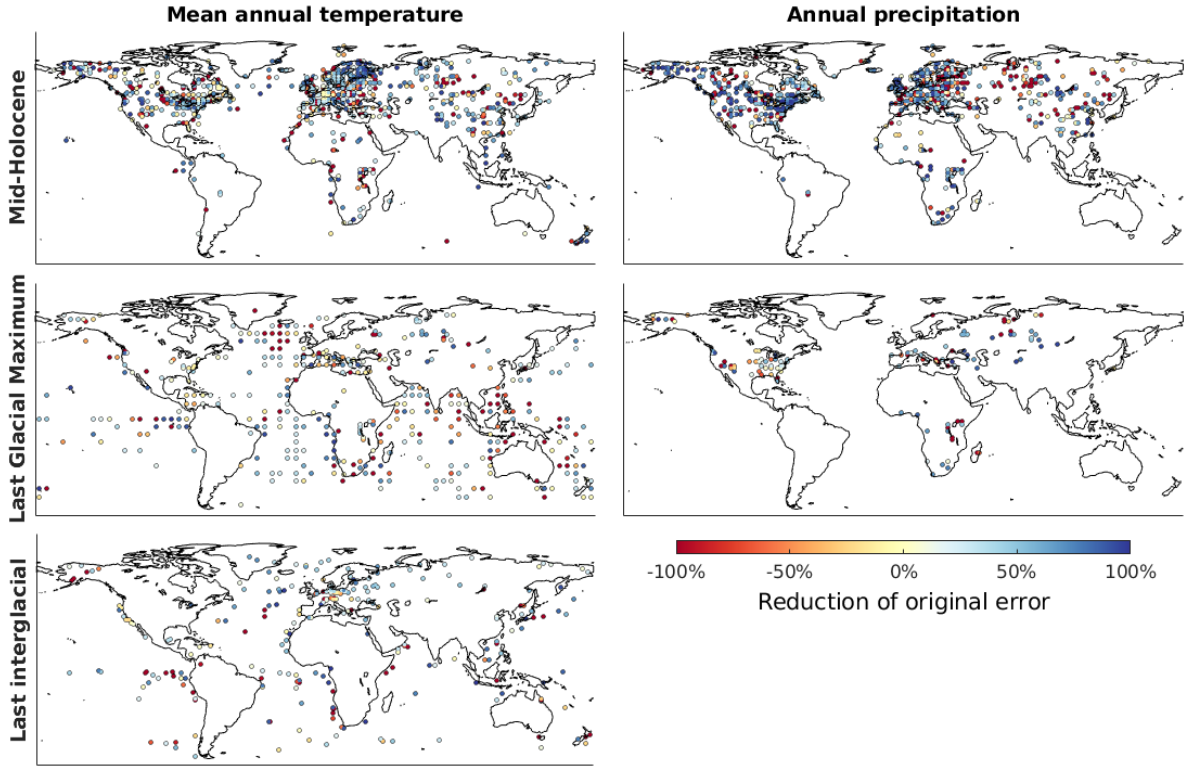


Figure 3. Reduction of the original model bias by the Delta Method for terrestrial and marine mean annual temperature and terrestrial annual precipitation. The lower end of the colour scale was capped at -100% (i.e. a doubling of the original bias).

propriate for a given land conformation. Thus, it is only appropriate for the Late Quaternary, and even for this period, changes in sea levels are problematic as they expose areas for which we have no bias information as well as changing the areas affected by maritime climate. GAMs should, in theory, obviate these problems by quantifying local processes as statistical relationships with appropriate proxies. Whilst this approach might be the only option for the deeper past, our results point to the fact that reconstructing such local processes in such a way is challenging, as demonstrated by its 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. By fitting interactions, it would

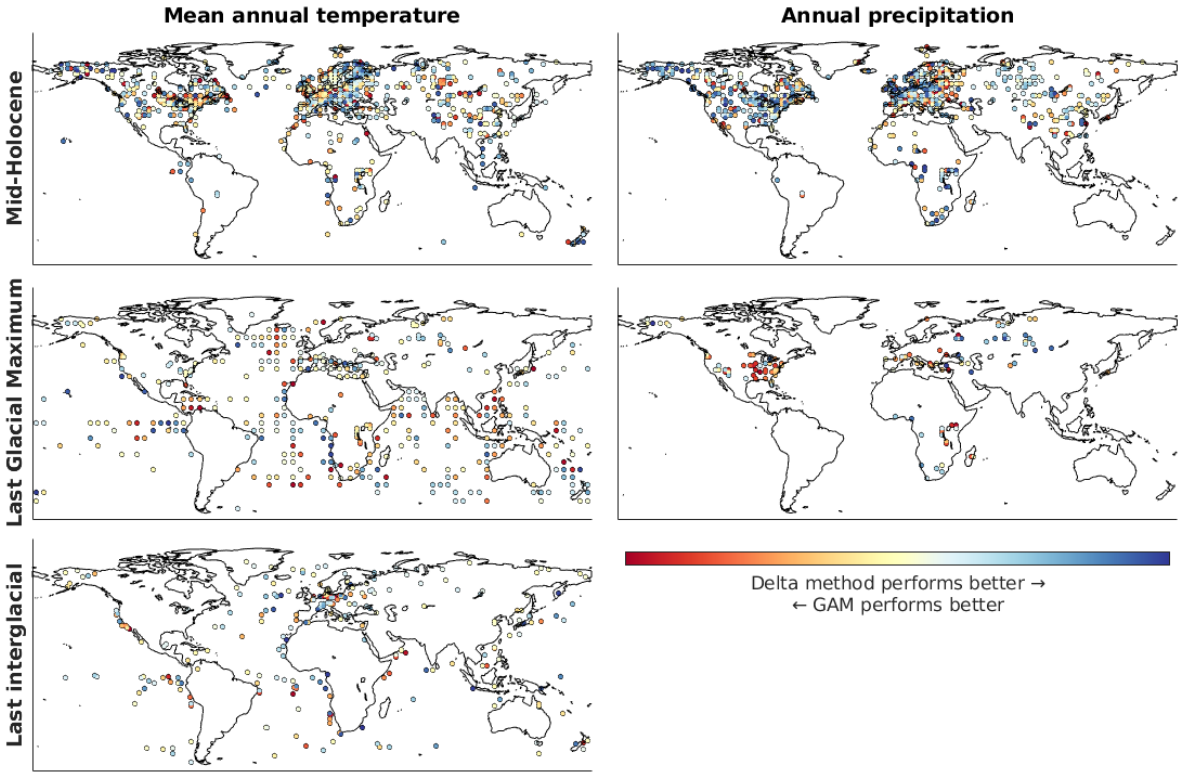


Figure 4. Relative performances of the Delta Method and the GAM approach in terms of debiasing simulated mean annual temperature (left column) and annual precipitation (right column). The colour spectrum represents the interval [0,1], and marker colours are calculated as the ratio of the absolute value of the local bias (Eq. (6)) of the GAM-based approach divided by the sum of the absolute local biases of both methods.

be possible to allow for these effects to differ depending on the local climatic conditions, but the computational complexitiy of interactions with such large datasets is non-trivial.

A major limitation of current approaches to debias climate model data is that they all assume biases in present-day climate to be fully representative of the past. With the progressive increase in the number of empirical reconstructions of past climatic conditions, it might be possible to soon move from a situation where past data are use to verify correction schemes (as we did in this manuscript) to using those data to actively calibrate the

bias correction function. Fig. 5 suggests an intriguing relationship between the temporal variation of the local model bias and the simulated climate change signal of the variable of interest. 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. A robust statistical model will require not only additional data from currently underrepresented geographical areas (specifically the southern hemisphere), but also curating empirical reconstructions, as successfully done for the last millenium (Hakim et al., 2016; Tardif et al., 2018).

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 the Delta Method is good starting point for bias removal of simulated Late Quaternary climate data at a global scale. However, given the considerable variability in the effectiveness of the different methods in different locations, we echo earlier propositions that studies focussing on specific regions require case-by-case assessments of which bias-correction method is most suitable for improving palaeoclimate model outputs (Maraun et al., 2017).

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.

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.

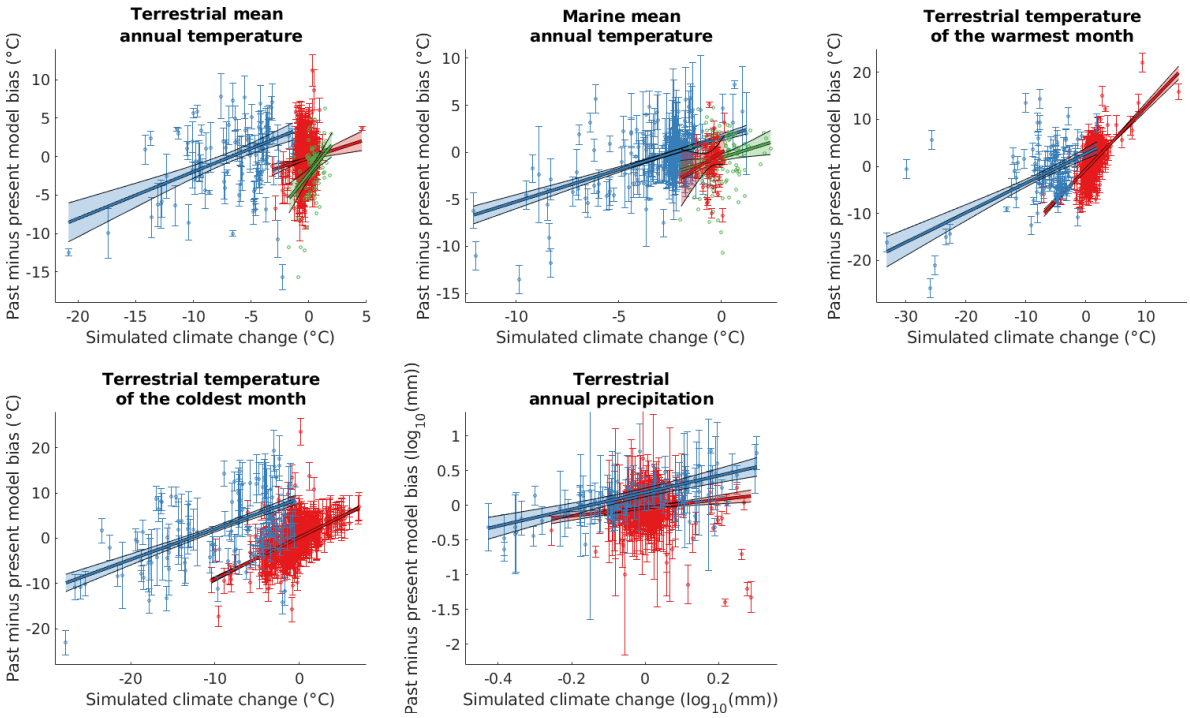


Figure 5. Differences between local past and present model bias (at locations for which reconstructions are available) against the local simulated climate change signal (i.e. the difference between past and present simulated value) of the variable of interest. Red, blue and green markers represent data 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.

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. R.B., M.K. and A.M. were supported by the ERC Consolidator Grant 647787 (”LocalAdaptation”).

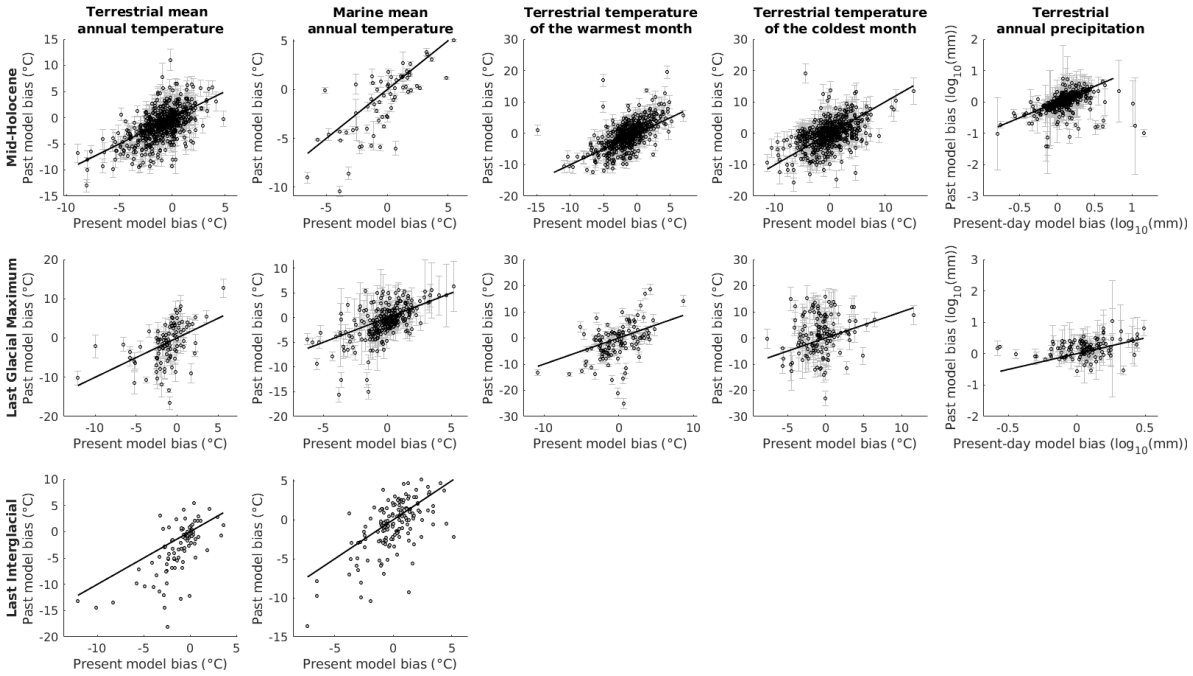


Figure A1. Comparison of present-day and past model biases (which the Delta Method assumes to be similar) from locations where reconstructions are available. Lines represent 1:1 relationships.

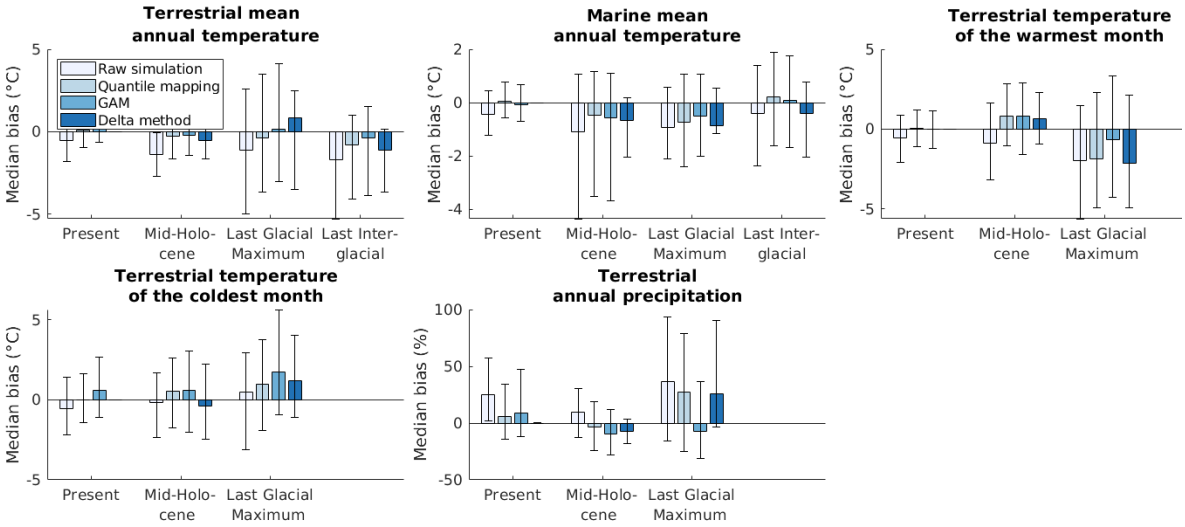


Figure A2. Median biases of the raw and bias-corrected climate simulation data. Error bars represent 25% and 75% weighted quantiles of the local biases available for the given climatic variable and point in time.

References

Bartlein, P., Harrison, S., Brewer, S., Connor, S., Davis, B., Gajewski, K., Guiot, J., Harrison-Prentice, T., Henderson, A., Peyron, O., et al.: Pollen-based continental climate reconstructions at 6 and 21 ka: a global synthesis, Climate Dynamics, 37, 775–802, 2011.

Beyer, R., Krapp, M., Eriksson, A., and Manica, A.: Windows out of Africa: A 300,000-year chronology of climatically plausible human contact with Eurasia, bioRxiv, 2020.

Ehret, U., Zehe, E., Wulfmeyer, V., Warrach-Sagi, K., and Liebert, J.: "Should we apply bias correction to global and regional climate model data?", Hydrology and Earth System Sciences, 16, 3391–3404, https://doi.org/https://doi.org/10.5194/hess-16-3391-2012, 2012.

Eriksson, A., Betti, L., Friend, A. D., Lycett, S. J., Singarayer, J. S., von Cramon-Taubadel, N., Valdes, P. J., Balloux, F., and Manica, A.: Late Pleistocene climate change and the global expansion of anatomically modern humans, Proceedings of the National Academy of Sciences, 109, 16 089–16 094, 2012.

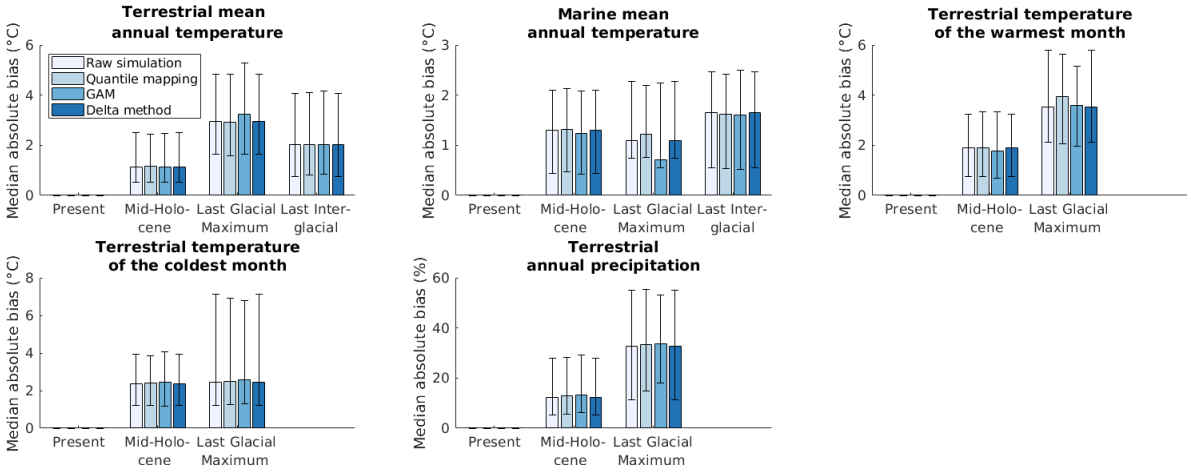


Figure A3. Median absolute biases of the climate change signal (*CCMAB*, Eq. (10)). Error bars represent 25% and 75% weighted quantiles of the local absolute climate change biases available for the given climatic variable and point in time.

Hakim, G. J., Emile-Geay, J., Steig, E. J., Noone, D., Anderson, D. M., Tardif, R., Steiger, N., and Perkins, W. A.: The last millennium climate reanalysis project: Framework and first results, Journal of Geophysical Research: Atmospheres, 121, 6745–6764, 2016.

Hessler, I., Harrison, S., Kucera, M., Waelbroeck, C., Chen, M.-T., Anderson, C., De Vernal, A., Fréchette, B., Cloke-Hayes, A., Leduc, G., et al.: Implication of methodological uncertainties for mid-Holocene sea surface temperature reconstructions, Climate of the Past, 10, 2237–2252, 2014.

Ho, C. K., Stephenson, D. B., Collins, M., Ferro, C. A. T., and Brown, S. J.: Calibration Strategies: A Source of Additional Uncertainty in Climate Change Projections, Bulletin of the American Meteorological Society, 93, 21–26, https://doi.org/10.1175/2011BAMS3110.1, https://journals.ametsoc.org/doi/abs/10.1175/2011BAMS3110.1, 2011.

Kottek, M., Grieser, J., Beck, C., Rudolf, B., and Rubel, F.: World map of the Köppen-Geiger climate classification updated, Meteorologische Zeitschrift, 15, 259–263, 2006.

Krapp, M., Beyer, R., Edmundsson, S. L., Valdes, P. J., and Manica, A.: A comprehensive climate history of the last 800 thousand years, EarthArXiv, https://doi.org/10.31223/osf.io/d5hfx, 2019.

Latombe, G., Burke, A., Vrac, M., Levavasseur, G., Dumas, C., Kageyama, M., and Ramstein, G.: Comparison of spatial downscaling methods of general circulation model results to study climate variability during the Last Glacial Maximum, Geoscientific Model Development, 11, 2563–2579, 2018.

Leonardi, M., Boschin, F., Giampoudakis, K., Beyer, R. M., Krapp, M., Bendrey, R., Sommer, R., Boscato, P., Manica, A., Nogues-Bravo, D., et al.: Late Quaternary horses in Eurasia in the face of climate and vegetation change, Science advances, 4, eaar5589, 2018.

Levavasseur, G., Vrac, M., Roche, D., Paillard, D., Martin, A., and Vandenberghe, J.: Present and LGM permafrost from climate simulations: contribution of statistical downscaling, Climate of the Past, 7, 1225–1246, 2011.

Lorenz, D. J., Nieto-Lugilde, D., Blois, J. L., Fitzpatrick, M. C., and Williams, J. W.: Downscaled and debiased climate simulations for North America from 21,000 years ago to 2100AD, Scientific data, 3, 160 048, 2016.

Maraun, D. and Widmann, M.: Statistical downscaling and bias correction for climate research, Cambridge University Press, 2018.

Maraun, D., Shepherd, T. G., Widmann, M., Zappa, G., Walton, D., Gutiérrez, J. M., Hagemann, S., Richter, I., Soares, P. M., Hall, A., et al.: Towards process-informed bias correction of climate change simulations, Nature Climate Change, 7, 764, 2017.

New, M., Lister, D., Hulme, M., and Makin, I.: A high-resolution data set of surface climate over global land areas, Climate research, 21, 1–25, 2002.

Rangel, T. F., Edwards, N. R., Holden, P. B., Diniz-Filho, J. A. F., Gosling, W. D., Coelho, M. T. P., Cassemiro, F. A., Rahbek, C., and Colwell, R. K.: Modeling the ecology and evolution of biodiversity: Biogeographical cradles, museums, and graves, Science, 361, eaar5452, 2018.

Reynolds, R. W., Rayner, N. A., Smith, T. M., Stokes, D. C., and Wang, W.: An improved in situ and satellite SST analysis for climate, Journal of climate, 15, 1609–1625, 2002.

Solomon, S., Qin, D., Manning, M., Chen, Z., Marquis, M., Avery, K., Tignor, M., and Miller, H., eds.: IPCC: Climate Change 2007: The Physical Science Basis. Contribution of Working Group I to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change, Cambridge University Press, 2007.

Tardif, R., Hakim, G. J., Perkins, W. A., Horlick, K. A., Erb, M. P., Emile-Geay, J., Anderson, D. M., Steig, E. J., and Noone, D.: Last Millennium Reanalysis with an expanded proxy database and seasonal proxy modeling, Climate of the Past Discussions, pp. 1–37, 2018.

Timmermann, A. and Friedrich, T.: Late Pleistocene climate drivers of early human migration, Nature, 538, 92, 2016.

Turney, C. S. and Jones, R. T.: Does the Agulhas Current amplify global temperatures during super-interglacials?, Journal of Quaternary Science, 25, 839–843, 2010.

Valdes, P. J., Armstrong, E., Badger, M. P., Bradshaw, C. D., Bragg, F., Davies-Barnard, T., Day, J. J., Farnsworth, A., Hopcroft, P. O., Kennedy, A. T., et al.: The BRIDGE HadCM3 family of climate models: HadCM3@ Bristol v1. 0, Geoscientific Model Development, 10, 3715–3743, 2017.

Vrac, M., Marbaix, P., Paillard, D., and Naveau, P.: Non-linear statistical downscaling of present and LGM precipitation and temperatures over Europe, Climate of the Past, 3, 669–682, 2007.

Waelbroeck, C., Paul, A., Kucera, M., Rosell-Melé, A., Weinelt, M., Schneider, R., Mix, A. C., Abelmann, A., Armand, L., Bard, E., et al.: Constraints on the magnitude and patterns of ocean cooling at the Last Glacial Maximum, Nature Geoscience, 2, 127, 2009.

Woillez, M.-N., Levavasseur, G., Daniau, A.-L., Kageyama, M., Urrego, D., Sánchez-Goñi, M.-F., and Hanquiez, V.: Impact of precession on the climate, vegetation and fire activity in southern Africa during MIS4, Climate of the Past, 10, 1165–1182, 2014.

Wood, S. N.: Stable and efficient multiple smoothing parameter estimation for generalized additive models, Journal of the American Statistical Association, 99, 673–686, 2004.

Zhu, D., Ciais, P., Chang, J., Krinner, G., Peng, S., Viovy, N., Peñuelas, J., and Zimov, S.: The large mean body size of mammalian herbivores explains the productivity paradox during the Last Glacial Maximum, Nature ecology & evolution, 2, 640, 2018.