We wish to extend our gratitude to the Reviewers for their comments, which we feel have greatly helped to improve the quality of our manuscript. We also wish to apologise for the substantial delay in returning this revised version - the result of a spinal injury of the lead author shortly after our submission of the responses to the Reviewer's comments in October 2019, which has impeded an earlier finalisation of this submission.

Reviewer 1

1) The evaluation criterion is essentially the difference between the corrected and reconstructed climatology. However, the Delta Method has been specifically constructed to eliminate this difference between simulated climatology and present-day climatology. Quantile Mapping pursues a more general correction, namely to correct the whole probability distribution of annual temperature (or precipitation). The GAM method is a statistical model that incorporates (in my understanding) simulated and observed gridpoint climatologies as predictors and predictands , and additionally some other factors like distance to the ocean, etc. The GAM method is therefore also not specifically tailored to eliminate the bias. I wonder if the main result of the manuscript, namely the best performance of the Delta method, is not an artifact. The Delta Method is precisely tailored to maximise the evaluation criterion and thus , it is for me not surprising that it outperforms the other two methods. I am not sure which other, fairer, evaluation criterion could be introduced, but I think that this issue should be addressed or at least thoroughly discussed.

> We have added the following section to clarify that, indeed, all three methods aim at minimising the difference between simulated and real climate, but make different assumptions as to how this aim can best be achieved:
>
> > 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.

Although the Delta Method fully eliminates present-day bias, as pointed out by the Reviewer, a priori, it is not clear whether it would also reduce *past* biases most effectively. Indeed, our analysis demonstrates that this is not the case in several scenarios - which supports the rationale underlying both the GAM Method and Quantile Mapping, i.e. present-day bias is *not* as good an estimate for past bias as the one obtained by using the other two methods.

In our revised version, we begin our results section by providing plots showing, for each climate variable, point in time and bias correction method, the unprocessed, complete set of local biases, thus illustrating the performance of each method across the full spectrum of values of the relevant climate variable. Only after that do we present the statistical summary of these plots, in terms of the median absolute biases.

We would argue that the MAB is the most natural and intuitive way to statistically summarise the set of local biases, providing a simple measure to assess, as we state later on in the text, whether a bias-correction correction method overall improves the raw simulation outputs (namely if the associated MAB is smaller than that of the non-bias-corrected simulations).

In addition, following Reviewer 2's suggestion, we now additionally evaluate the performance of each method in terms of improving the simulated climate change signal, and have summarised these results in the newly added supplementary figure A3.

2) The difference between the corrected simulated climatologies and the reconstructed climatologies does not take into account the presumably large uncertainty in the reconstructions and in the corrected simulated climatologies (the former being presumably much larger?) . This needs to be incorporated in the evaluation of the three methods. If the inter-methodological differences are much smaller than the uncertainties in the estimated paleo-bias , it would be difficult to claim that one particular method is superior to other two. I think that the manuscript should include also these uncertainty estimations, or at least place the inter-methodological differences in the frame of the reconstruction uncertainties.

We have included the Reviewer's suggestion into our analysis. We have added the following paragraph to the Methods:

> We tested whether the median absolute biases associated with two bias-correction methods, a specific 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.

and have added the following paragraph to the results:

> 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 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.

We have also added a caveat in the Abstract and the Conclusions stating that the slightly better overall performance of the Delta Method is not always statistically significant.

3) The readability of the illustrations is poor. it is, for instance, very difficult to discern anything in Figure 2 and Figure 3. The lettering, axis labels, etc, in most figures is too small (e.g Figure 4)

We have provided higher-resolution figures, and have increased the size of the maps by removing unnecessary white spaces. We have increased the font sizes across figures.

4) what is the original spatial resolution of the climate reconstructions ? were they regridded, and how?

We have now specified this in section 2.1.2 as follows:

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°x0.8° 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°x0.8° grid cell containing the sample location. Fig. 3 and Fig. 4 visualise the locations of all reconstructions of terrestrial and marine mean annual temperature, and of annual precipitation.

5) The text refers sometimes to bias , other times to 'error', whereas in my understanding very often both terms carry the same meaning. This can be confusing for some readers. I would recommend to stick to one of those terms when possible.

We have removed the term 'error' (which we had previously used in the sense of 'the bias remaining after bias-correction') as suggested, and now use 'bias' throughout the text.

6) The text also refers to the climate reconstructions as 'the observations', e-g. in equation 5. This can also be confusing. It would be clearer to use 'climate reconstructions' when referring to the proxy-reconstructed climatologies and 'observations' when referring to present-day climatologie.

We now use the terminology suggested by the Reviewer throughout the text.

7) The main conclusion is derived from the analysis of only one model. Perhaps I missed it but I think this a caveat that should be mentioned.

We have added this caveat to section 2.1.1, as suggested.

Reviewer 2

1) Page 1, lines 11-13, The DC method has its name because it is based on adding the simulated difference between two periods to observations. Although this is mathematically equivalent to subtracting the fitting period bias from the simulations (as shown in eqn.1) I think introducing the BC methods using the second definition rather than the one that is directly linked to the name is potentially confusing.

We now introduce the Delta Method as follows:

> The Delta Method is based on adding the difference between past and present-day simulated climate (the 'delta') to present-day observed climate.

We have also changed the order of equations in Eqs. (1) and (2), as suggested by the Reviewer.

2) Page 1, lines 13 -15, Please use clean terminology. GAM is a statistical representation of links between ‘variables’ not between ‘proxies for processes’ and ‘biases’. Define clearly what predictors and predictands are. From the current statement it is impossible to find out which variables are actually linked through GAM.

We have rephrased the statement as follows:

> 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).

We have also rewritten section 2.2.2, in which GAM methods are explained in detail (see our responses to comments further below).

3) Page 1, lines 15-16, QM does not assume the shape of the distribution to be constant in time. If there is climate change the distribution obviously changes. Standard implementations assume that the bias for a given value is constant in time (but there are implementations without this assumption). Please remove wrong statement and include a correct explanation.

We have corrected the statement as follows:

> Quantile mapping assumes that biases are specific to their respective quantiles in the distribution of the relevant climatic variable.

4) Page 1, lines 18-20. Please be more specific about the potential setups in the palaeoclimate context. Some empirical palaeoclimate reconstructions are local or have a high spatial resolution,

which means they are smaller-scale than the climate model output (downscaling), whereas continental-scale empirical reconstructions have a lower resolution than the models (upscaling).

We rewrote section 2.1.2 as as follows, to accommodate the Reviewer's comment:

> We used global datasets of 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 continental 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 Last Interglacial terrestrial and marine temperature.
>
> 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°x0.8° 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°x0.8° grid cell containing the sample location. Fig. 3 and Fig. 4 visualise the locations of all reconstructions of terrestrial and marine mean annual temperature, and of annual precipitation.

5) Page 1, There is a complete lack of critical discussion about the limitations of BC. It is obvious that a fundamentally poor model cannot be improved in a meaningful way by BC (see for instance Maraun and Widmam (2018), Maraun et al., 2017: Towards process informed bias correction of climate change simulations. Nature Climate Change, 7(11), 764-773). These limitations should be discussed in the introduction, in particular in the context of palaeoclimate simulations. Moreover, the validation approach needs to be justified taking into account the potential problems with BC, and clear comments need to be made on whether the validation would identify such problems. It will turn out that it would not (see comments below), which should at least be stated as a limitation of the study.

We have added the following paragraph to the Introduction:

> 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, or 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 paleoclimatological 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 improve a fundamentally poor climate model, e.g. with strong circulation biases that 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.

In addition, we have added the following paragraph to the Introduction:

> [W]e focus on the global performance of the different methods; however, we note that bias-correction is not a one-size-fits-all approach (Maraun et al., 2017), and that our results do not remove the need for local re-evaluations of methods in specific continental and subcontinental regions of interest.

and the following sentence to the Conclusion:

> 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).

We have also added the following caveat to the definition of the MAB (previously MAE):

> We emphasise that the MAB is a summary statistic of the degree 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.

6) Page 4, lines 18-19, The statement about the log-transform is correct, but overcomplicated. It is more helpful to just say that this is a multiplicative delta method, i.e. the simulated relative change is applied to the observations.

We have removed the sentence referring to the log-transformation, as suggested, and have added the following statement:

> [This] corresponds to applying the simulated relative change to the observations.

7) Page 4, lines 21-22, The sentence doesn't work out. The relationship is between climate model output and real-world climate variables, with additional time-invariant predictors such as topography or distance from the coast.

We have rewritten the sentence to clarify dependent and independent variables:

> 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). "

8) Page 5, eqn. 3, Clarify that some x_i are time-dependent (i.e. those that represent climate model output), while others (topography, distance from coast) are not.

We now explicitly state the temporal dependency of the predictor variables in the equations, and have specified in the text:

> [These predictors] 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.

9) Page 5, line 13, does 'wind speed' include the direction?

We have added "(absolute)" to clarify that we mean speed, not velocity.

10) Page 5, line 14-15, It is not clear what the predictor and predictand data are and how the fitting for the f_i works. What are the individual realisations of T_sim and x_i for which the polynomials are fitted? Are these timesteps? But if so, if I understand correctly, there are only three, namely the mean temperatures for the present, Mid-Holocene, and Last Glacial Maximum. What is the spatial resolution? Are the simulated temperatures averaged over the continental areas represented in the proxy-based reconstructions? Or are the realisations in space (if so, is this one value for each continent?), or space and time?

> In our initial submission, we had abused mathematical notation in some instances (e.g. by dropping the dependence of certain variables on time, location, or the climate variable or bias correction method in question), with the aim of facilitating an intuitive understanding of the key concepts. We understand that this may have caused misunderstandings and loss of clarity of our methods. We have therefore completely rewritten the mathematical parts of section 2.2 (bias correction methods) and section 2.3 (method evaluation). We have explicitly added the dependence of variables on time, location, climate variable, and bias-correction method throughout these sections, thus clarifying the details that the Reviewer enquired about.

11) Page 5, line 17-22, It is not clear what the distributions are. Are they annual values of continental means?

> In the course of rewriting the technical details of the methods (see response to previous comment), we have clarified the data that the relevant cumulative distribution are based on. Please also refer to our response to comment 14, regarding the term “continental”.

12) Page 6, The evaluation method needs more justification. For instance it would be a logical first step to validate the three BC methods on instrumental data, using crossvalidation, and focusing on aspect that are important in the palaeoclimatic context, i.e. long-term variability. The argument is probably that the key aspect is the representation of changes on multi-millennial timescales. The evaluation section should start with stating the objectives of the evaluations, followed by a justification of why the chosen evaluation method addresses these objectives. Please keep in mind that BC methods reduce bias by construction, even for completely wrong models (see e.g. Maraun et al, 2017). A reduction in the bias of the mean (DC, GAM), will reduce also the biases for the distribution quantiles, while BC corrects these directly. For strongly biased climate models the

reduction of the biases in the distributions will necessarily lead to a reduction in MAE. Why is the MAE chosen as the evaluation measure? In the paleoclimate context it is also very relevant to compare the climate change signals in the raw and the BC-corrected simulations, and in the proxybased reconstructions. Please add statements and if suitable figures on this.

We have added the following paragraph to the beginning of the section 2.3 (Model evaluation) to clarify the objective of our 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 the differences at specific locations and points in time. Here, we determine these local differences between empirical reconstructions and bias-corrected simulations for each climate variable and bias-correction method, and define a spatially aggregated measure to assess the overall global performance of each method.

After formally defining the local differences between empirically reconstructed and bias-corrected simulated data, we now motivate the use of the MAB as follows:

> We provide complete plots of the distribution of the biases corresponding to each specific climate variable, point in time, and bias correction method. 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 [MAB].

We would argue that the MAB is the most natural and intuitive way to statistically summarise the set of local biases, providing a simple measure to assess, as we state later on in the text, whether a bias-correction correction method overall improves the raw simulation outputs (namely if the associated MAB is smaller than that of the non-bias-corrected simulations).

However, we have added Figs.1a-e, showing for each climate variable, point in time and bias correction method, the unprocessed complete set of local biases, thus illustrating the performance of each method across the full spectrum of values of the relevant climate variable. Only then do we show the statistical summary of these plots, in terms of the MAB, in Fig. 2.

As suggested by the Reviewer, we have also included an evaluation of the performance of each bias-correction method in terms of reducing the average bias between the empirically reconstructed and the simulated climate change signal, which may be relevant in certain applications. We have added the formal details of this evaluation to section 2.3 (Model evaluation). The newly added supplementary figure A3 shows that the differences between the methods in terms of bias-correcting the climate change signal are extremely small.

We agree with the Reviewer that bias-correction methods reduce the overall bias in present-day simulations, and we now explicitly state this in the text. However, we would argue that it is not clear, a priori, whether any of the three bias-correction methods considered also reduces biases in past simulations. Indeed, our analysis shows that this is not always the case: Some bias-corrected simulations have a higher MAB than the raw simulation data.

13) Page 5, line 7, 'standard errors' of what? It is said later that it is the error of the reconstructions, but it needs to be said the first time this is mentioned.

We have added information on the standard errors of the empirical data to the description of the empirical reconstructions in section 2.1.2.

14) Page 6, eqn. 6, The notation is very unclear. It is also not clear what 'grid cell' refers to. Earlier it was mentioned that continental means are used. This problem is related to the lack of clarity about predictand and predictor data mentioned in previous comments.

As mentioned in our response to a previous comment by the Reviewer, we have completely rewritten and clarified the mathematical parts of our methods. This includes the section referred to by the Reviewer.

We feel that the term "continental", which we have used in the sense of "terrestrial" (following Bartlein et al. (2011), our source of Mid-Holocene and LGM empirical reconstructions), may have led to confusion about the spatial scale of the empirical reconstructions used in our analysis. These are always local/gridded, never spatially aggregated across continents. (Thus, "Continental mean annual temperature" referred to the (locally specific) mean annual temperature of terrestrial data points.) We have clarified

this in our methods by emphasising the locality and spatial dependence of variables. In addition, we now use the term “terrestrial” instead of “continental” throughout the text.

15) Page 7, figure 1. It seems not plausible that QM leads to substantially larger MAEs than for the raw simulations, with values up to 10 K. Surprisingly this is not even discussed. There might be a problem with the implementation. If the implementation is correct, please give a detailed explanation how this is possible. If I understand correctly the BCcorrected distribution of the present simulation is identical to the distribution of the instrumental observations. This means that the instrumental observations have also a very high MAE for the present. How can this be the case? If suitable, please add information about how the instrumental data, which are the training data for all BC methods, perform in this evaluation framework. When addressing this please state explicitly what is compared with what for calculating the MAE; the information that is currently given is incomplete.

There was indeed an error in the implementation of Quantile Mapping, and we thank the Reviewer very much for bringing this to our attention. We have corrected the error, and find that Quantile Mapping also slightly reduces model biases, as expected by the Reviewer. We have updated the figures and text accordingly.

The Reviewer is correct in that the cumulative distribution function of present-day simulated climatic values obtained after applying Quantile Mapping is identical to the cumulative distribution function of present-day observed values. However, this does not imply that the underlying climate maps must be identical (in which case the MAB would be 0). Indeed, any spatial permutation of present-day observed climate values would have the same cumulative distribution function as present-day observed climate, but the MAB would not necessarily be 0. Only in the case of the Delta Method are present-day observed climate and bias-correct simulated data identical (as are, by extension, their cumulative distribution functions).

16) Page 11, figure 4. If I understand correctly, this figure shows that the simulated climate change signal is different from the reconstructed climate change signal. If this is correct, please include this straightforward interpretation.

This is not the case. Letting $V_{sim}(x,t)$ and $V_{emp}(x,t)$ denote the simulated and empirical values of a climate variable V at time t and location x. The figures suggest a relationship between "Past minus present model bias" - i.e. $(V_{emp}(x,t)-V_{sim}(x,t)) - (V_{emp}(x,0)-V_{sim}(x,0))$ - on the one hand, and the "Simulated climate change" - i.e. $V_{sim}(x,t) - V_{sim}(x,0)$ - on the other hand. This is different from the Reviewer's suggestion that "the simulated climate change signal ", $V_{sim}(x,t) - V_{sim}(x,0)$, "is different from the reconstructed climate change signal", $V_{emp}(x,t) – V_{emp}(x,0)$. We have clarified the relevance of the observed relationships in the figure caption.

A ~~systematic comparison~~ comparative evaluation of bias correction methods for ~~paleoclimate~~ 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 ~~paleoclimatologies~~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 and quantile mapping~~ 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~~ Mid-Holocene (~6,000 years BP), the ~~last glacial maximum~~ Last Glacial Maximum (~21,000 years BP) and the ~~last interglacial period~~ Last Interglacial Period (~125,000 years BP). Overall, the ~~delta method performs best~~ Delta Method performs slightly better, albeit not always to a statistically significant degree, at minimising the median absolute ~~error~~ bias between empirical data and debiased simulations for both temperature and precipitation ~~, although~~ than GAMs and Quantile Mapping, however, there is considerable spatial and temporal variation in the performance of each of the three methods. ~~We~~ 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 ~~paleoclimate~~ palaeoclimate are a key requirement for modelling many important long-term and large-scale ecological processes ~~(?????)~~(??????). 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 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) (**?**).

~~Whilst bias~~ Bias correction has received a great deal of attention for present-day and near-future simulations (**??**), ~~work on paleoclimate~~ whereas work on palaeoclimate reconstructions has been much more limited. This is partly due to the different time scale of ~~paleoecological~~ 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 (**?**), 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 ~~: the delta method~~ in the palaeoclimatological context: the Delta Method (http://www.worldclim.org/downscaling), statistical methods based on generalised additive models (GAMs) (**????**) and ~~quantile mapping~~ 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, ~~even though~~ although each method takes a different approach in the aspect that is assumed to be invariant. The ~~delta method~~ Delta Method assumes bias to be location-specific (**?**), as it is based on a map of differences between observed and simulated values. GAMs attempt to represent ~~local processes by finding statistical association between proxies for those processes (e.g. altitude,~~ statistical relationships between simulated climatic variables (as well as other known physical variables, such as elevation and the distance from the coast) and ~~biases for present-day observation~~ bias-corrected climatic variables (**??**). Finally, ~~quantile mapping assumes the shape of~~ Quantile Mapping assumes that biases are specific to their respective quantiles in the distribution of ~~a certain variable to be constant through time~~ the relevant climatic variable (**?**). However, debiased simulation data have

either not been validated against ~~observational data~~ empirical reconstructions at all, or only for a small geographical area and a single point in the past. ~~Furthermore, because of the limited spatial resolution of many paleoclimate reconstructions, bias correction is conflated with downscaling, thus confounding any estimate of the actual effect of debiasing the data.~~

Here, we use a set of high-resolution climate simulations to ~~systematically~~ evaluate the performance of the ~~delta method~~Delta Method, a GAM-based approach, and ~~quantile mapping~~Quantile Mapping, against a global dataset of empirical climatology data from the present, the ~~mid-holocene~~ Mid-Holocene (~6,000 years BP), the ~~last glacial maximum (~21~~Last Glacial Maximum (~21,000 years BP) and the ~~last interglacial~~ Last Interglacial Period (~125,000 years BP). ~~Thanks to the high resolution of our simulation data, we can isolate the effect of debiasing from downscaling~~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 (**?**), and that our results do not remove the need for local re-evaluations of methods in specific continental and subcontinental regions of interest.

Section ~~3~~ **??** provides details of the three bias correction methods, the climate simulations, and the empirical ~~paleoclimatology~~ palaeoclimatology reconstructions used in this study. In section ~~4~~**??**, we quantitatively assess the performance of the methods at a global scale, and with regard to spatial and temporal heterogeneities. Section ~~5 discusses how paleoclimate~~ **??** 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~~ Delta Method and statistical bias correction.

2 Material and Methods

2.1 Climate data

2.1.1 Modelled climate data

We used ~~paleoclimate~~ 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~~ 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~~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. 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 ~~below) of continental~~ section **??**) 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, ~~since~~ 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~~ –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 ~~on~~ using present-day observational data. ~~We used monthly continental~~ 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 ~~aggregated~~ remapped to the 1.25°×0.83° grid of the ~~paleoclimate~~ palaeoclimate simulations by taking the average of values contained in each target grid cell.

We used ~~paleoclimate reconstructions of continental~~ 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~~ , for the Mid-Holocene and the LGM from **?**, reconstructions of mean annual sea surface temperature for the ~~mid-Holocene~~ Mid-Holocene and the LGM from **?** and **?**, respectively, and reconstructions of mean annual ~~continental~~ terrestrial and sea surface temperature for the ~~last interglacial period~~ 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 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 **??**) 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.

?? and ?? 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~~Delta Method

The ~~delta method~~ 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 (**?**). For temperature ~~, the local bias~~ 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 ~~simulated and observed temperature, $T_{\text{obs}}(0) - T_{\text{sim}}(0)$~~observed and raw simulated temperature, $T_{\text{emp}}(x,0) - T^{raw}_{\text{sim}}(x,0)$. Debiased temperature, ~~$\widehat{T}_{\text{sim}}(t)$, $T^{DM}_{\text{sim}}(x,t)$~~, in location x at some time t is obtained ~~by adding the bias term to the simulated temperature, $T_{\text{sim}}(t)$, at time t:~~ as:

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

The second expression illustrates that ~~$\widehat{T}_{\text{sim}}(t)$ $T^{DM}_{\text{sim}}(x,t)$~~ is alternatively given by adding the ~~simulated climate change signal to the~~ present-day ~~observed temperature~~local bias to the simulated temperature, $T^{raw}_{\text{sim}}(x,t)$, 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~~Delta Method for precipitation, which corresponds to applying the simulated relative change to the observations (**?**). Analogously to temperature, debiased precipitation is given by

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

~~This is equivalent to log-transforming simulated and observed precipitation values, applying the additive delta method, and back-transforming the result.~~

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 ~~potentially additional~~ additional known forcings such as topography ~~, and real climate~~ (independent variables) (**??**). 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 ~~based on the appropriate simulations~~using the appropriate simulated data. Generalised additive models (GAMs) have gained particular popularity as transfer functions (**????**). They accommodate potential nonlinearities in the response of the individual variables, ~~but assume that the interactions between predictor variables can be neglected (~~while – owing to the computational requirements of general high-dimensional nonlinear regressions ~~). GAMs compute the expected value of debiased local temperature (and, analogous, precipitation) as~~

$$E[\widehat{T}_{\text{sim}}(t)|x_1,\ldots,x_n] = \sum_{i=1}^{n} f_i(x_i),$$

~~where $x_1,\ldots,x_n$ are predictors provided by the climate model~~ – 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 **??**) 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 ~~simulated temperature and precipitation, or represent other known~~ climate model outputs (typically including the raw simulated data of the climate variable in question, V_{sim}^{raw}), and/or certain geographical or physical ~~variables such as local~~ 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 ~~shortest~~ distance to the ocean~~. $f_1,\ldots,f_n$ are generally nonlinear~~ , 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 ~~using~~ to minimise the distance between the left and the right hand side in Eq. (**??**). Once the model has been calibrated on the present-day ~~observed and simulated variables. Debiased past variables are calculated by applying the fitted functions to past simulated variables.~~

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

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

Similar to **?**, here we used elevation, the shortest distance to the ocean and simulated temperature as ~~predictors for debiased temperature , and the~~ 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 ~~for debiased~~ 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 (?)~~ *mgcv* package **?** in R.

2.2.3 Quantile ~~mapping~~Mapping

~~This method involves mapping quantiles of the simulated distribution of~~ 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 ~~onto the appropriate observed~~ V, we denote by $F_{\text{emp}}^V[0]$ the cumulative distribution function of the present-day ~~quantiles, so as to remove systematic distributional biases in the simulation data (?). Debiased temperature (~~empirical observations $V_{\text{emp}}(x_1, 0), V_{\text{emp}}(x_2, 0), \ldots$ (i.e. $F_{\text{emp}}^V[0]$ is the function that monotonically maps these values onto the interval $[0, 1]$). Analogously, we denote by $F_{\text{sim}}^{V,raw}[t]$ the cumulative distribution function of the raw simulated values $V_{\text{sim}}^{raw}(x_1, t), V_{\text{emp}}^{raw}(x_2, t), \ldots$ at time t. We denote by $F_{\text{emp}}^V[0]^{-1}$ and $F_{\text{sim}}^{V,raw}[t]^{-1}$ (both mapping $[0, 1]$ to $\mathbb{R}$) the inverse functions of $F_{\text{emp}}^V[0]$ and $F_{\text{sim}}^{V,raw}[t]$, respectively.

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

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

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), ~~analogous, precipitation)is calculated as~~

$\widehat{T}_{sim}(t) = F^{-1}_{obs}\left(F_{sim}[t](T_{sim}(t))\right)$

~~where F_{obs} and $F_{sim}[t]$ denote the cumulative probability functions of the global set of observed values at present day and of simulated values~~ 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.

2.4 ~~Method evaluation~~

~~We assessed the local error between empirical and debiased simulated data at a given~~ 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 ~~and location as~~ in the location x is then given by

$measure for evaluating and comparing the performance of the three bias correction methods consider e d is, we use the median of the a l$

$\{|\text{B}_V^M(x_i^{(t,V)},t)|\}_{i=1,2,\ldots}$. $The median is weighted by grid cell area for the present, and by the available local inverse standard errors of th$ the local errors (Eq. $) from all locations for which empirical data is available. Then the median absolute error is given by to$

~~and where the weights w_i are given by~~ We denote the ~~inverse standard error associated with the appropriate empirical reconstructions, rescaled such that $\sum_{i=1}^{n} w_i = 1$~~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

$$MAB_V^M(t) = \text{weighted median}\big(\{|B_V^M(x_i^{(t,V)},t)|\}_{i=1,2,\ldots}\big) \tag{6}$$

$= |B_V^M(x_k^{(t,V)},t)|$, where the median index k satisfies

$$\sum_{\substack{|B_V^M(x_i^{(t,V)},t)|<\ldots \\ |B_V^M(x_k^{(t,V)},t)|}} \omega_{\text{emp}}(x_i^{(t,V)},t) \leq \frac{1}{2} \text{ and } \sum_{|B_V^M(x_i^{(t,V}$$

$|B_V^M(x_k^{(t,$

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 ~~error~~ bias is smaller than the median absolute difference between raw simulations and empirical data~~, which is calculated analogously~~. 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 **??**.

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

$$AME = |\text{weighted median}(E)| = |E_k|, \text{ where } k \text{ satisfies } \sum_{E_i<E_k} w_i \leq \frac{1}{2} \text{ and } \sum_{E_i>E_k} w_i \leq \frac{1}{2},$$

median bias,

$$MB_V^M(t) = \text{weighted median}\left(\{B_V^M(x_i^{(t,V)}, t)\}_{i=1,2,\ldots}\right) \quad (7)$$

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

3 ~~Results~~

~~Overall, the delta method provided the strongest reduction in~~ 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

$$CCB_V^M(x_i^{(t,V)},t) = \begin{cases} \left(V_{\text{sim}}^M(x,t) - V_{\text{sim}}^M(x,0)\right) - \left(V_{\text{emp}}(x,t) - V_{\text{emp}}(x,0)\right) \\ \qquad\qquad \text{if } V \in \{T^{\text{ter.mean}}, T^{\text{Tmar.mean}}, T^{\text{cold}}, T^{\text{warm}}\} \\ \dfrac{V_{\text{sim}}^M(x,t) - V_{\text{sim}}^M(x,0)}{V_{\text{sim}}^M(x,0)} - \dfrac{V_{\text{emp}}^M(x,t) - V_{\text{emp}}(x,0)}{V_{\text{emp}}(x,0)} \\ \qquad\qquad \text{if } V = P^{\text{ann}}, \end{cases} \tag{8}$$

and the median absolute bias ~~(*MAE*, Eq.) for all variables and past time points~~ associated with the climate change signal is given by

$$CCMAB_V^M(t) = \text{weighted median}\left(\{|CCB_V^M(x_i^{t,V)},t)|\}_{i=1,2,\ldots}\right), \tag{9}$$

where the weighted median is calculated analogously as in Eq. (**??**). 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.

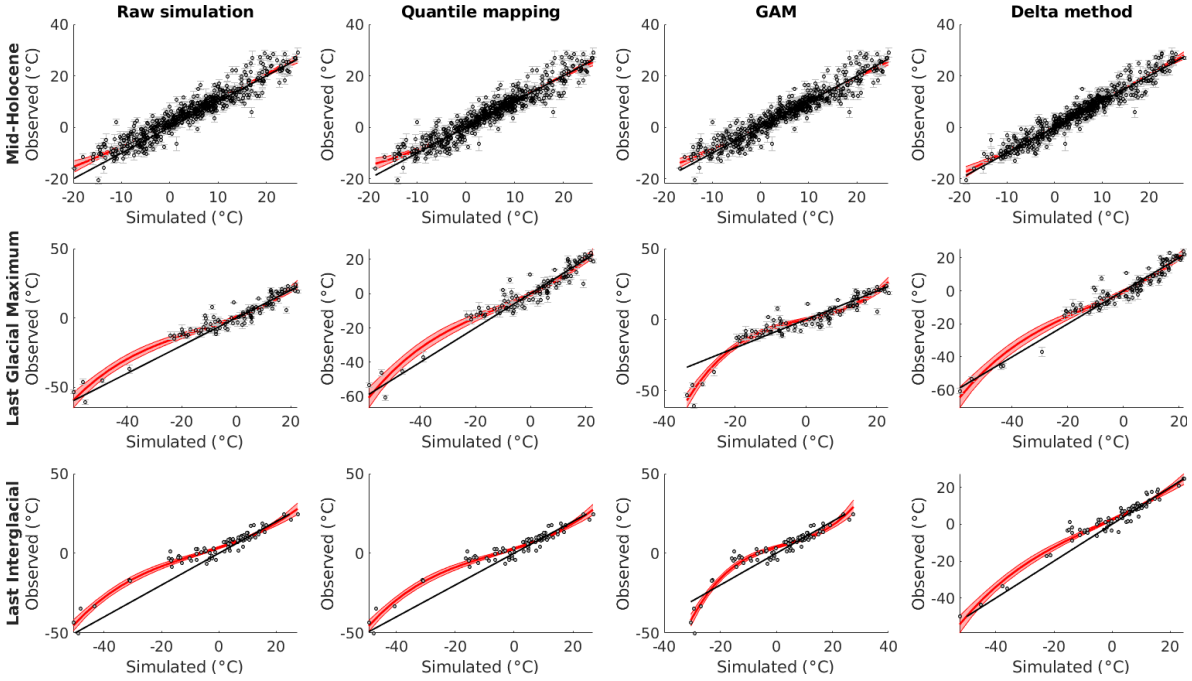
3 Results

Fig. ??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 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.

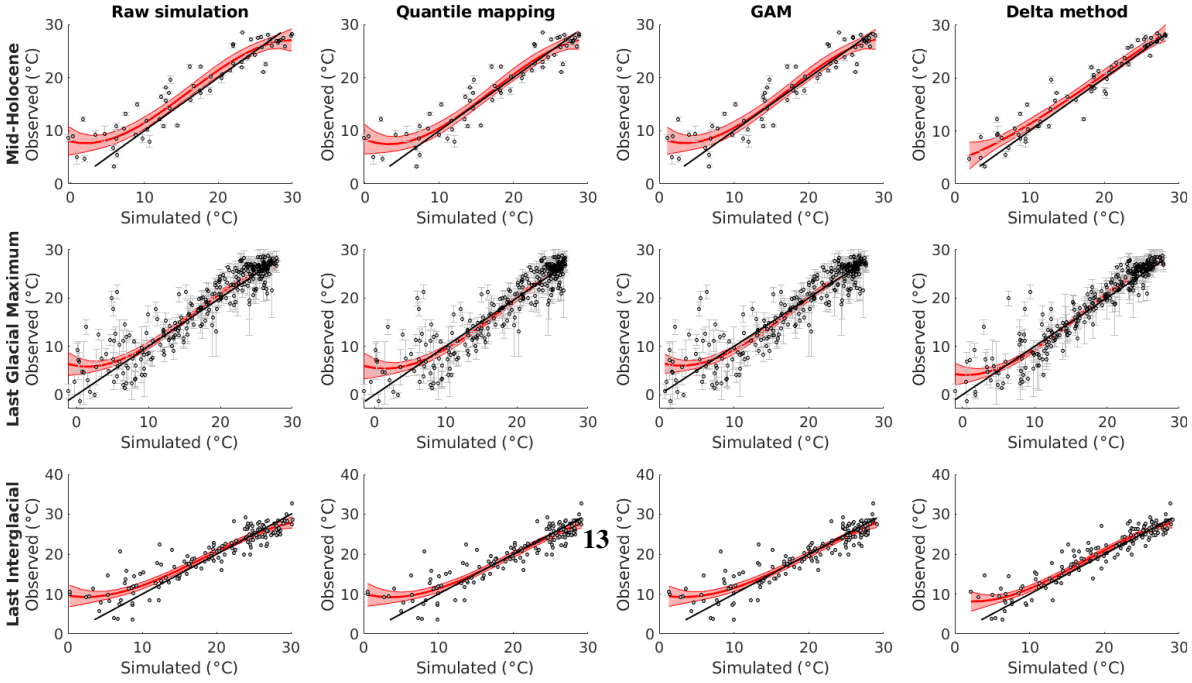
All bias-correction methods reduce the median absolute bias (*MAB* in Eq. (??)) of present-day simulated data for all climate variables – as would be expected (?) – although, by construction, only the Delta Method completely eliminates all differences between simulated and observed data (Fig. ~~??),~~ ??). The Delta Method also provides the strongest reduction in 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 ~~, where quantile mapping performs better. The relatively~~ (Fig. ??). The comparatively good performance of the ~~delta method~~ Delta Method is also reflected in the correlations between present-day and past model ~~bias~~ biases, which the Delta Method assumes to be similar (Fig. ??). ~~Quantile mapping significantly increased original biases in continental mean annual temperature throughout the time periods considered, as well as in present-day and mid-Holocene precipitation. GAMs generally led~~ The GAM method and Quantile Mapping generally lead to a reduction in bias, even though overall not as effectively as the ~~delta method ; however, for LGM continental~~ Delta Method. In a few cases, the original bias is actually increased after applying a correction method (Fig. ??).

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 ??) 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).

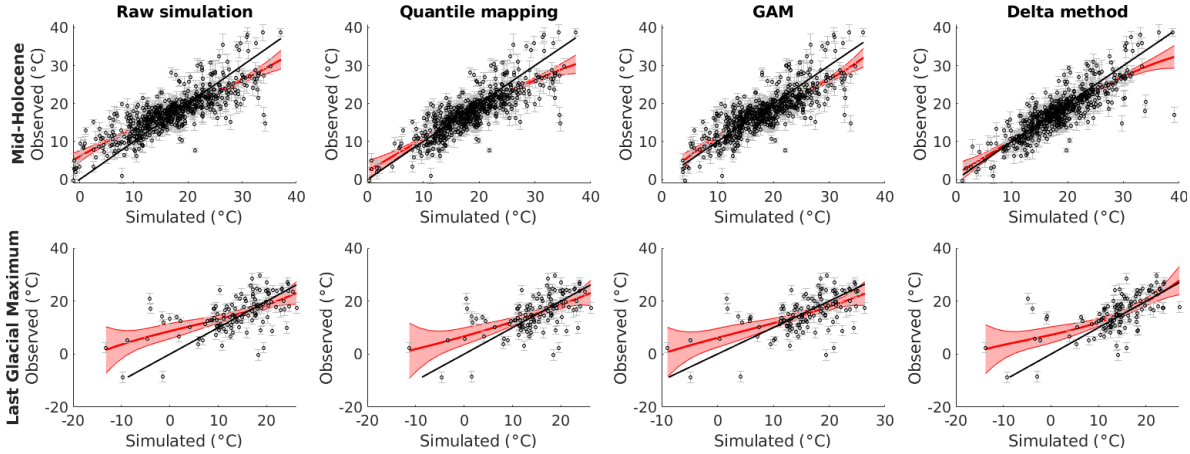
(a) Terrestrial mean annual temperature



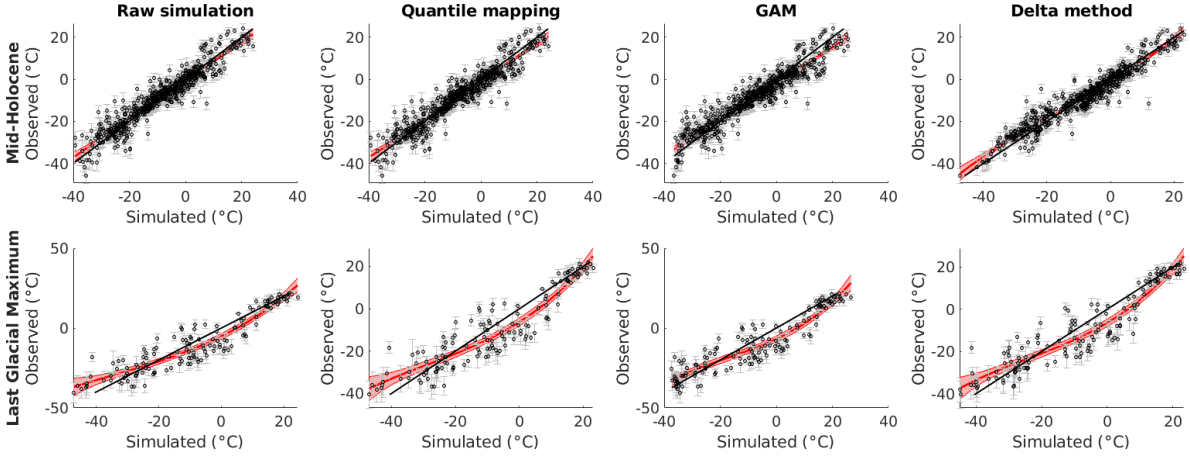
(b) Marine mean annual temperature



(c) Mean temperature of the warmest month



(d) Mean temperature of the coldest month



(e) Annual precipitation

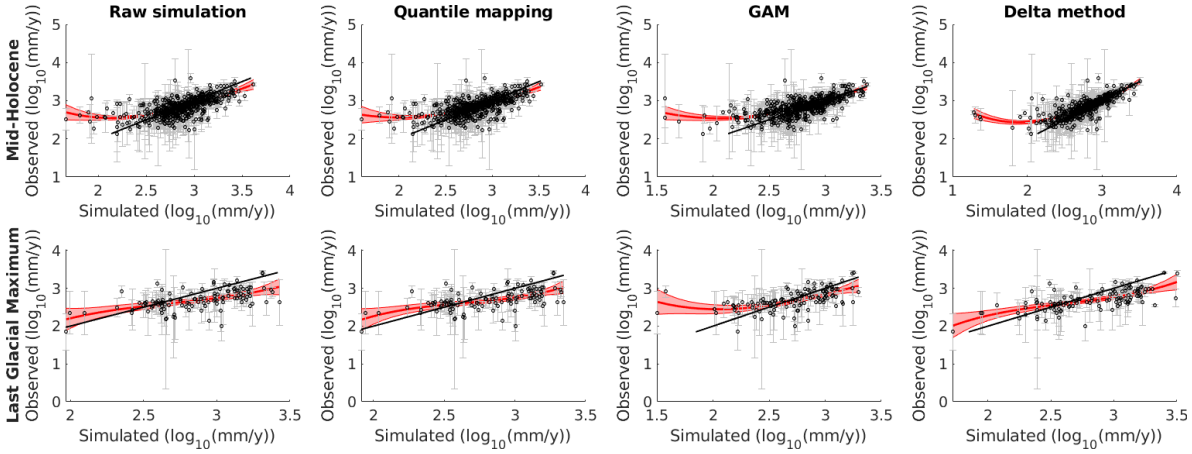


Figure 1. Comparison of bias-corrected simulated and empirically reconstructed climate variables. Black lines show 1:1 relationships. Red lines and shades show 5th degree polynomial regression and 95% confidence intervals, respectively.

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 ~~, the bias was actually increased~~(65% of realisations). The GAM method performed significantly better than Quantile 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. **????**). These trends ~~were not completely removed by any bias correction method, although both the delta method and GAMs~~ are as present in the bias-corrected data. Indeed, methods consistently reduced the ~~original absolute median error. GAMs, in particular, minimised the absolute median errors for past~~ absolute value of the median bias (*MB* in Eq. (**??**)) 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. (**??**)) are negligible in all scenarios except for marine mean annual temperature

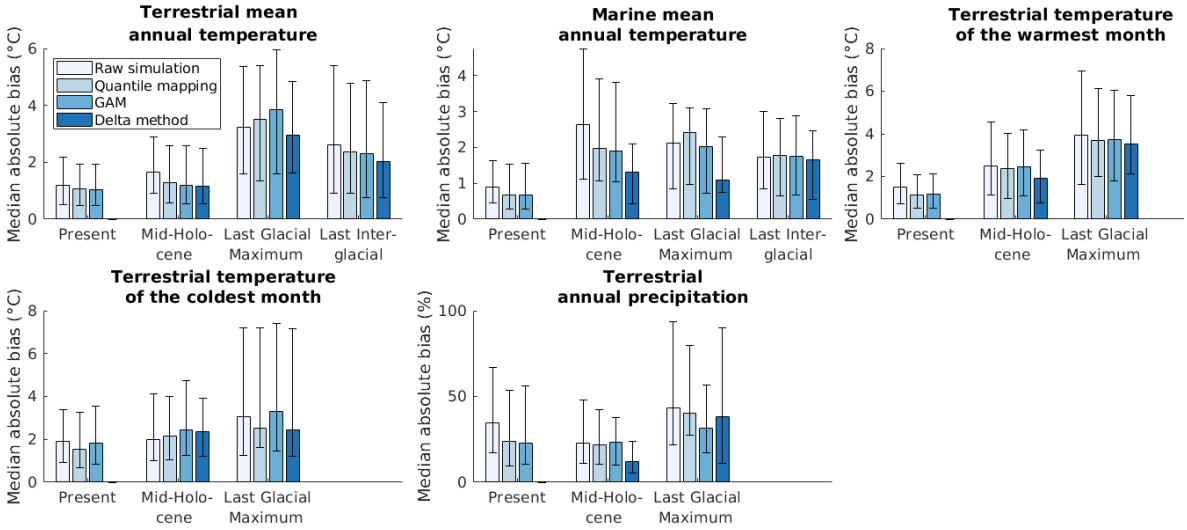


Figure 2. Median absolute ~~errors~~ biases (~~*MAE*~~*MAB*, Eq. (**??**)) of the ~~simulated~~ raw and bias-corrected climate simulation data. Error bars represent 25% and 75% weighted quantiles ~~. Figs. A1 – A5 show the complete scatter plots~~ of ~~empirical against raw~~ the local absolute biases available for the given climatic variable and ~~debiased simulated data~~point in time.

~~and mid-Holocene and LGM precipitation~~ 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~~Delta Method. For example, the ~~delta method~~ Delta Method significantly reduced the original bias of modelled precipitation in Eastern North America in the ~~mid-Holocene~~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. ~~While the delta method~~ For example, while the Delta Method has a slight overall edge over the GAM approach, 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 point in time generally does not guarantee the same result at a different time. For ~~example, the delta~~

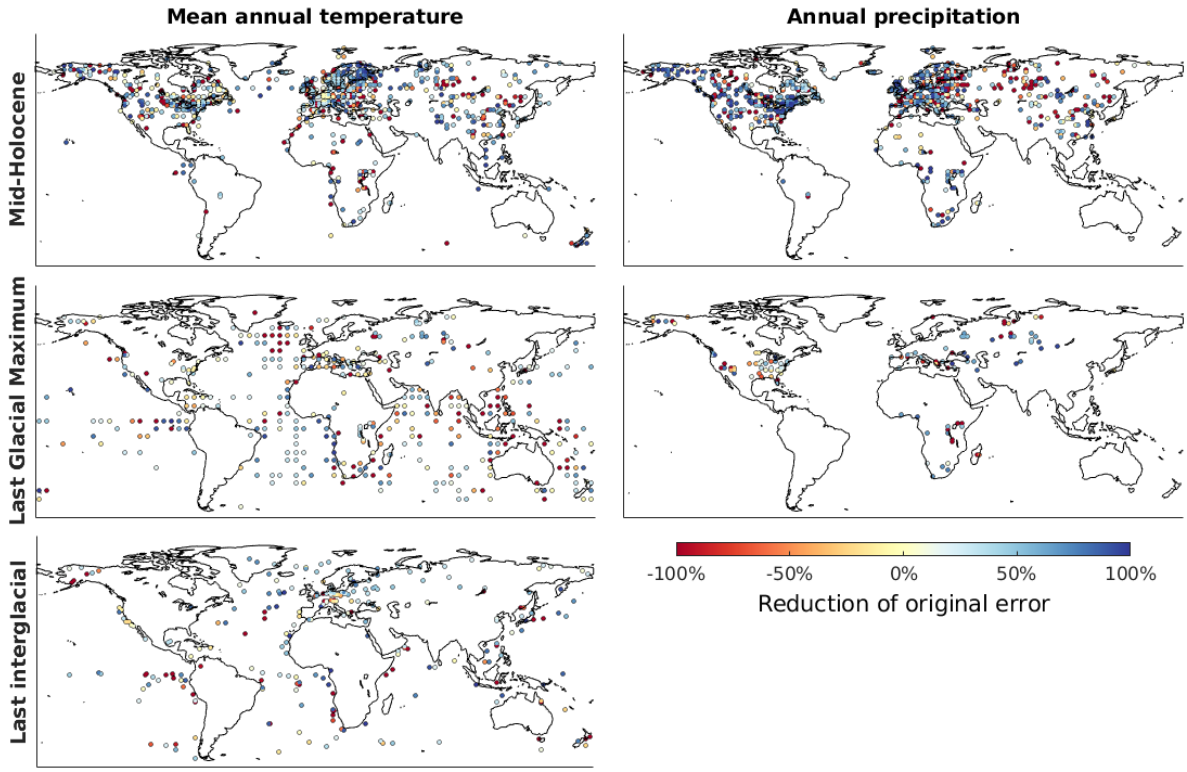


Figure 3. Reduction of the original model bias by the ~~delta method~~ Delta Method for ~~continental~~ terrestrial and marine mean annual temperature and ~~continental~~ terrestrial annual precipitation. The lower end of the colour scale was capped at -100% (i.e. a doubling of the original ~~error~~bias).

~~method~~ instance, the Delta Method overall reduced the original ~~error~~ 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~~the delta method~~ , 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

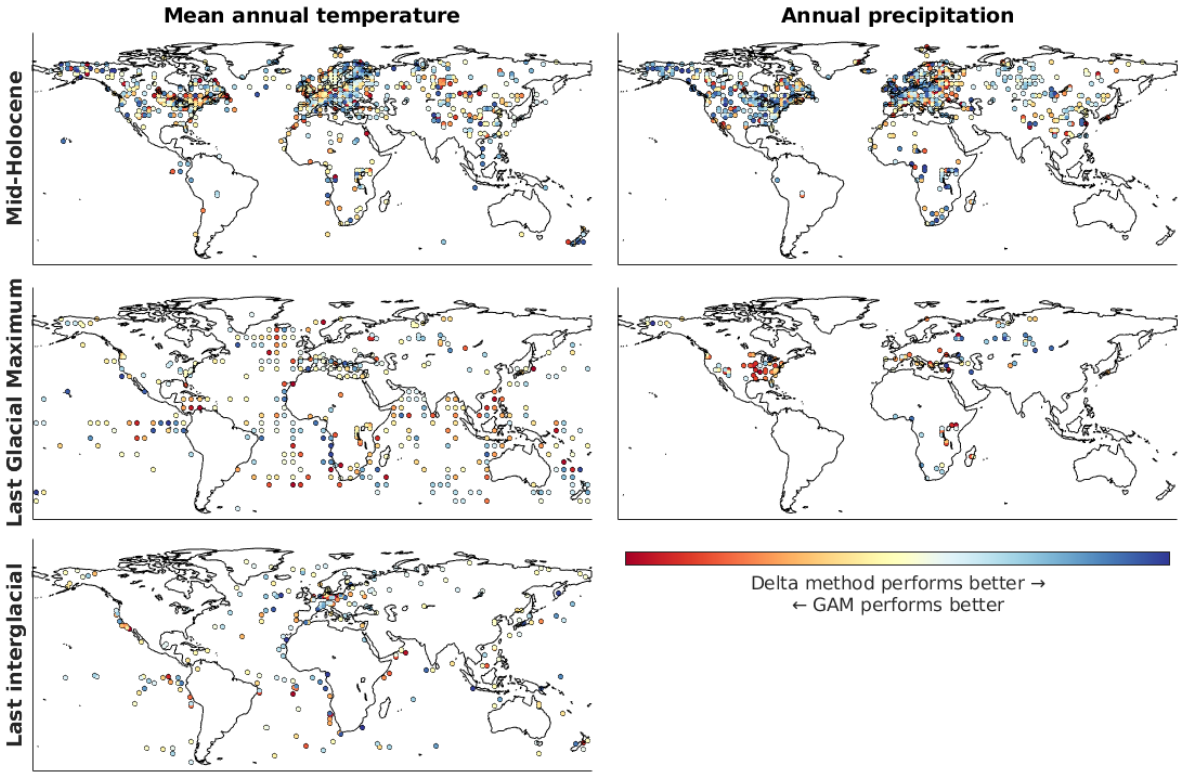


Figure 4. Relative performances of the ~~delta method~~ Delta Method and ~~a~~ the GAM approach in terms of debiasing simulated mean annual temperature (left column) and annual precipitation (right column). The colour spectrum represents the interval [0,1], and marker colours are calculated as the ratio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

method is only appropriate 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~~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 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. **??** 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~~ 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 (**??**).

5 Conclusions

Our comparison of global debiased ~~paleosimulation~~ palaeosimulation data and empirical reconstructions suggests that, overall, the ~~delta method~~ Delta Method provides slightly better performance at debiasing compared to GAMs ~~, whilst quantile mapping is a poor choice for removing bias~~and Quantile Mapping – though not in all cases to at a statistically significanty extent. Given our results, we suggest that the ~~delta method~~ Delta Method is good starting point for bias removal of simulated Late Quaternary climate data ~~, bearing in mind that its effectiveness varies regionally.~~ 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 (**?**).

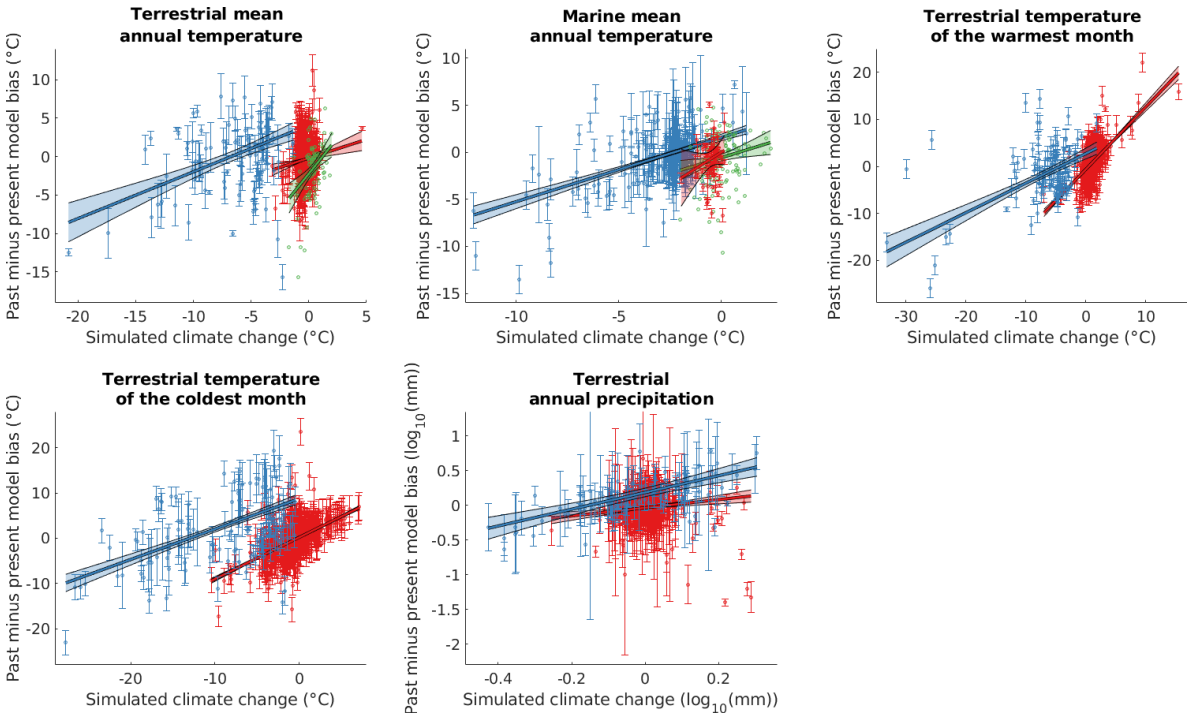


Figure 5. Differences between local past and present model bias (at locations for which reconstructions are available) against the local simulated climate change signal (i.e. the difference between past and present simulated value) of the variable of interest. Red, blue and green markers represent data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

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. ~~We believe that such~~ 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.

~~Empirical reconstructions of continental mean annual temperature against raw and debiased simulation data. Lines show 1:1 relationships.~~

~~Empirical reconstructions of marine mean annual temperature against raw and debiased simulation data. Lines show 1:1 relationships.~~

~~Empirical reconstructions of continental temperature of the warmest month against raw and debiased simulation data. Lines show 1:1 relationships.~~

~~Empirical reconstructions of continental temperature of the coldest month against raw and debiased simulation data. Lines show 1:1 relationships.~~

~~Empirical reconstructions of continental annual precipitation against raw and debiased simulation data. Lines show 1:1 relationships.~~

~~Comparison of present-day and past model biases in locations where reconstructions are available. Lines show 1:1 relationships.~~

~~Median errors of the simulated and bias-corrected climate data. Error bars represent 25% and 75% quantiles.~~

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

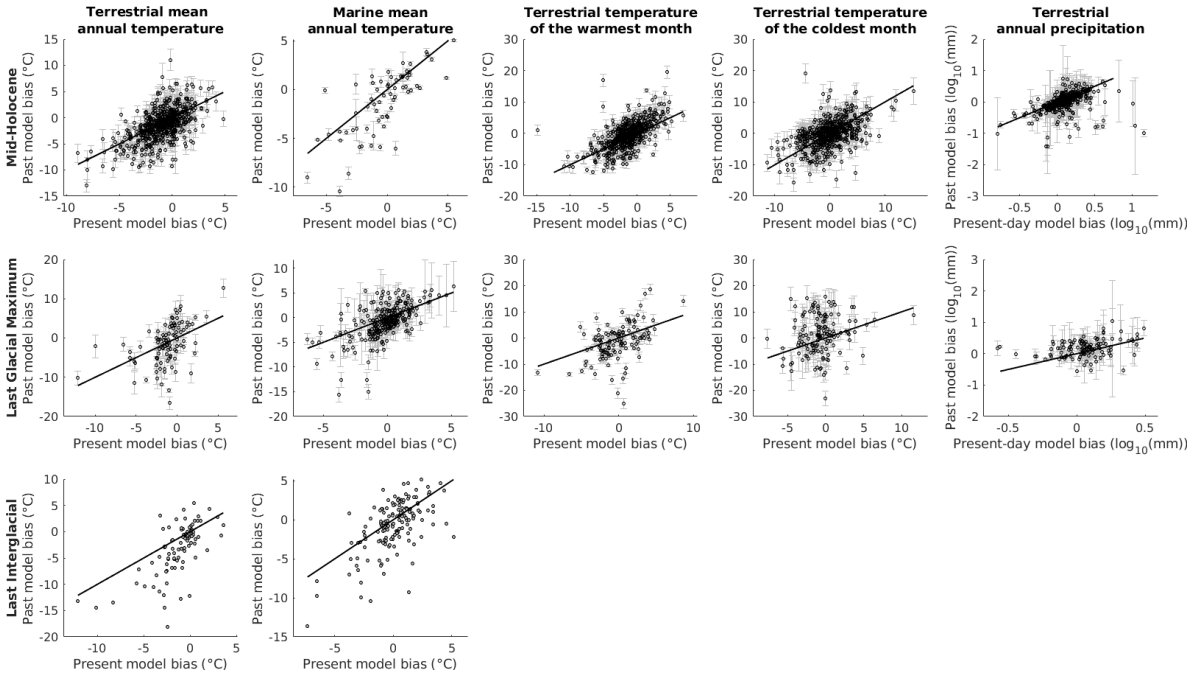


Figure A1. Comparison of present-day and past model biases (which the Delta Method assumes to be similar) from locations where reconstructions are available. Lines represent 1:1 relationships.

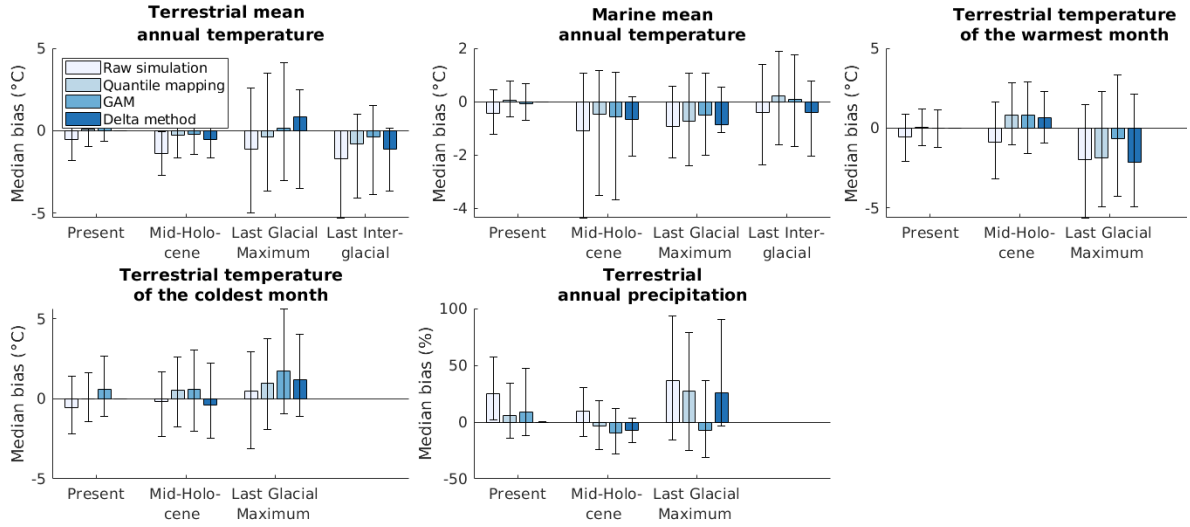


Figure A2. Median biases of the raw and bias-corrected climate simulation data. Error bars represent 25% and 75% weighted quantiles of the local biases available for the given climatic variable and point in time.

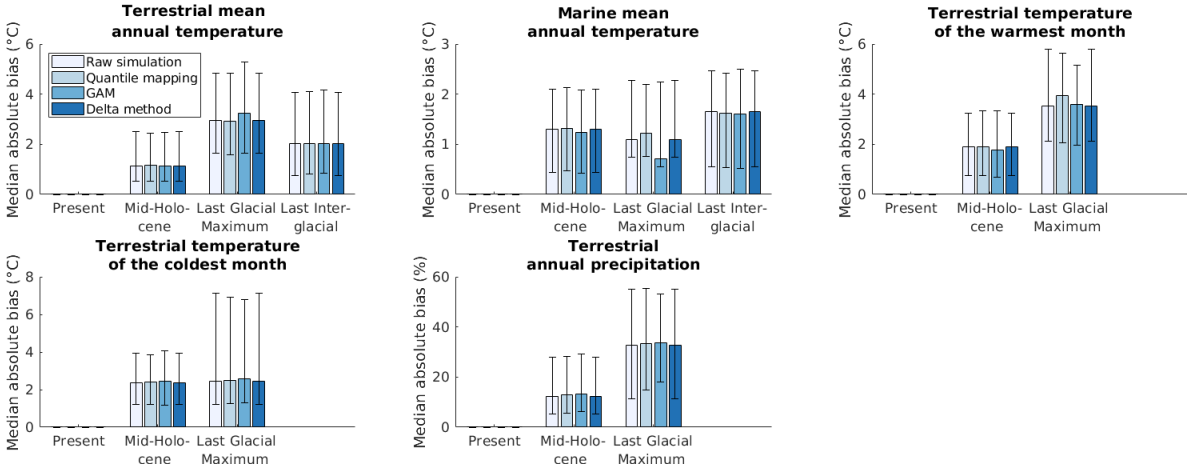


Figure A3. Median absolute biases of the climate change signal (*CCMAB*, Eq. (**??**)). Error bars represent 25% and 75% weighted quantiles of the local absolute climate change biases available for the given climatic variable and point in time.