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 oftentimes not present in the paleoclimatological context. This is because palaeoclimate data is most often provided at a medium-scale spatial resolution, and represents 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:

> [H]ere, 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 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 substantial variability of 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 of a specific type, i.e. it does not allow inference of the goodness of the climate model, or of the performance of each method in improving the representation of 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 that these 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.

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 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. We only show the statistical summary of these plots, in terms of the MAB, in Fig. 2.

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.

As suggested by the Reviewer, we have 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). A newly added figure shows that the differences between the methods in terms of bias-correcting the climate change signal are extremely small.

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” (e.g. like 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. We have corrected this 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”, $V_{emp}(x,t)$ - $V_{emp}(x,0)$.