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 suitable 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 in terms of its 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 these 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 complete, unprocessed 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 now also motivate our evaluation approach in greater detail by means of the following new paragraph

> 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. […] 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).

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 a newly added figure.

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 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 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 (Fig. 2). 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 the other methods for all variables at present day.

We have also highlighted these results in Fig. 2, and have added caveats in the Discussion stating that the slightly better overall performance of the Delta Method in several cases 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 increased the resolution and size of the maps. We have increased the font sizes in all figures.

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

We have specified 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’ 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.