Review of

‘A systematic comparison of bias correction methods for palaeoclimatic simulations’

by R. Beyer, M. Krapp and A. Manica

Recommendation: reject, invite resubmission of a revised version

This manuscript presents a comparison of three bias correction (BC) methods applied to GCM simulations for the present, Mid-Holocene, and Last Glacial Maximum. The method applied are a delta change (DC) method, a generalized linear model (GLM), and quantile mapping (QM).

Exploring BC methods in the context of palaeoclimate simulations is a useful contribution to the research area, and I thus in principle support publication of studies related to this topic. However the use of BC in this manuscript is mechanistic, uncritical and superficial, and the overall approach, methods and specific setups are poorly explained, There is also a potential implementation error for QM. The lack of clarity on what has actually been done is so large that I cannot assess whether the results are in principle suitable for publication. I thus think that a revised manuscript should be sent for a full review again.

Specific comments

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.

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.

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.

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).

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.

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.

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.

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.

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

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?

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

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 cross-validation, 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 proxy-based reconstructions. Please add statements and if suitable figures on this.

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.

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.

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 BC-corrected 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.

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.