We would like to thank the Editor for his detailed comments. In the following we explain how we have modified the manuscript to address his suggestions. All changes are highlighted in **yellow** front in the revised manuscript.

The original suggestions are written in bold font and our responses with normal font.

Referee #1 / General comments #2, #3 and #4 and Referee #2 / line 409: You have responded to these comments, but you do not appear to have made any relevant changes to the manuscript. If you have revised the manuscript, please highlight the changes that you have made. If you have not revised the manuscript, then please consider adding at least some of the information that you provide in your responses.

Referee 1 General comments:
I want to thank the authors for addressing many of my concerns and I commend them for more thoroughly investigating the LSTMs architecture and hyperparameters to try and get the LSTMs to work. I think the new discussion at the end of the paper is helpful for understanding why the LSTM's are fundamentally failing to improve over linear methods, i.e. the implied link is linear. I think that finding, and others from section 4.1, are important enough to warrant inclusion in the abstract since they're broader claims about non-linear methods for CFR that this paper found evidence for. Could the bi-LSTM then be seen as more of a tool for investigating the (non)amenability of CFR to non-linear methods in small data regimes, rather than just an "off the shelf" application that didn't pan out? If that understanding is correct, then I would consider emphasizing that a little more up front
A: In response to this general comment, we elaborated a little more than in our original version that an important objective of our study testing and verifying our working hypothesis that a more complex machine-learning method LSTM would provide better temperature reconstructions with the right amplitude of variations. We noticed that in this context, we were bound to use a limited amount of data to train the machine-learning method compared to other 'big data' applications.
In the revised manuscript, we have included in the abstract a summary of the new discussion contents in section 4.1. We have also added a sentence into the abstract that the Bi-LSTM could be a tool for exploring the amenability of CFRs especially in small data regimes.

2. ***Table 1 — Why do PCR and CCA improve in the noisy scenarios but LSTM deteriorates****?*
A: For the nonlinear machine learning methods, it is very sensitive to external noise. Kalapanidas et al., 2003 and Atla A, et al., 2011, demonstrated that linear regression can perform better results than nonlinear methods considering noise sensitivity studies. And some studies indicated that external interference or noise could damage the ability of neural networks (Heaven 2019). This is explained in lines 536.
We have also added sentence to elaborate more on the performance explanation in the Discussion section of the revised manuscript.

3. ***Line 527: "PCR and CCA exhibit overestimated reconstructions within noisy PPEs, the Bi-LSTM presents relatively robust reconstructions" — What are overestimated reconstructions? Is this indicating some advantage of LSTM****?*
A: The overestimated reconstruction here refers to the overestimation in the amplitude of variability. The overestimation is represented by the ratio of standard deviations, as a metric for assessing the reconstruction variance. A SD ratio close to 1 indicates we achieve perfect reconstructions with the same amplitude as the target. In these noisy PPEs, the linear regression method we employed, especially the PCR method, exhibits obvious overestimations (the value of SD ratio is bigger than 1 over some spatial regions as shown

in Fig 6-7). This comment is related to the previous comment, and we believe it is also addressed in lines 536.
We have added sentence to indicate that the overestimated reconstructions refer to the reconstructions in the amplitude of climatic variability. And added sentence to indicate the possible advantage of LSTM in CFRs.

4. ***Line 555: "Both ESN and LSTM belong to the family of RNN, yet ESN is much simpler than LSTM (Lukosevicius and Jaeger 2009), and has outperformed the RNN methods in other applications (Chattopadhyay et al., 2019; Nadiga, B. 2020)." — If ESN's are simpler and more promising, then why does this paper stick to the LSTM model, which clearly did not improve over simpler methods, and not just propose and evaluate ESN's, even if alongside the LSTM****?***
A: The results regarding the LSTM have been achieved along this study, and this experience lead us to think that the ESM could be more promising. This assumption is based on a few preliminary results, but not on a thorough testing. However, we cannot be sure at this stage that this will turn out to be correct. Our plan is to test the ESM in a follow-up publication.
The ESN method we mentioned in this manuscript is because we have already implemented further CFR experiments by employing this ESN and also compared it with the LSTM method. This is explained in lines 565.
We have added a sentence to emphasize that we will implement further steps to test ESN method in CFRs.

Refereces:
Kalapanidas E, Avouris N, Craciun M, Neagu D. (2003). Machine learning algorithms: a study on noise sensitivity. In Proc. 1st Balcan Conference in Informatics. pp. 356–365.
Atla A, et al. Sensitivity of different machine learning algorithms to noise. J Comput Sci Coll. 2011;26(5):96–103.
Heaven, D. Why deep-learning AIs are so easy to fool. Nature 574, 163–166 (2019).

Referee #2 / Lines 378-379: You have not responded to this comment. Please do so.
Ln 378-9: This does indeed support the stationarity of the teleconnection patterns, but also says something about the physical nature of the patterns, i.e. they are to some degree localized and do not share significant amounts of covariance outside of the regions where they are sampled*.*
A: We agree with this comment, and have included this suggestion into the revised manuscript. The changes are highlighted in yellow font in lines 380.

Ln 409: This is only true if the EOFs in the training interval are stationary, i.e. well represent the EOF patterns in the reconstruction interval as well*.*
A: In our manuscript, we have indicated that the EOF patterns derived from training interval are assumed to remain constant in time. This is also an assumption that made in real paleo reconstructions, otherwise the PCR method would not be valid. This stated in lines 410.
We have added sentence to state that the dominant EOF patterns are assumed to be stationary/constant with time in our PPEs.

Technical corrections:
Line 24-25: Please replace "simple" with "samples" and "dataset" with "datasets". I'm also not sure what you mean by "positive tests". Perhaps you could replace "positive tests on" with "that skill can be achieved even when" or similar.
A: We have corrected it in lines 25.