1. General comments

A: We would like to thank Referee 1 for their detailed and constructive comments.

Regarding their general comments, we agree that it would be beneficial to deepen our exploration of the LSTM sensitivity to different architectures and tunable parameters, as is common for deep learning applications. In the revised version, we will evaluate the LSTM methodology using a range of architectures and explore more deeply the performance of LSTM (see also answer to Specific Comment 2). We will also put our results in context with paleo-data assimilation methods, and explore the reasons why the Bi-LSTM achieves similar reconstruction skill as PCR. We will also take care to add more climatological insight to our analysis.

In the following responses, we specify how we plan to address those points.

2. Specific comments

1. I'm still unclear as to whether this article is proposing Bi-LSTMs as a viable alternative to traditional statistical methods or whether its trying to show that they don't work well enough. If this article is intended to propose using Bi-LSTMs (or to refute their use) then this stance needs to be made more clear up front and in the results. Right now the article presents the results as a neutral comparison of methods", but this makes it difficult to reach a conclusion about the proposed method.

A: The revised version will more broadly explore the Bi-LSTM method, which will allow us to achieve more robust conclusions on the Bi-LSTM. The results achieved so far indicate that the method is not superior to the traditional linear methods, with the possible exception of the replication of cold extremes, but the planned deeper exploration with a range of architectures may modulate this conclusion.

2 Follow up to Q1-If the article is proposing (or refuting) Bi-LSTMs then it needs a more comprehensive evaluation of the Bi-LSTM model. The results in Appendix B are a good start, but it would have been nice to see how varying the depth of the network, using dropout, weight decay, or other regularization techniques, or varying the learning rates (or using a scheduler), number of epochs, and other aspects of the training procedure such as the loss function effect generalization. From this, a reader could draw broader conclusions about the effectiveness of Bi-LSTM models rather than the effectiveness of the single model presented.

A: We will explore a range of architectures, different network depths, introducing dropout layers, using different learning rates, and employing different loss functions to provide a more comprehensive evaluation of the Bi-LSTM performance and effectiveness. Those results will be rather detailed and - depending on the outcome - we will need to take care not to overload the manuscript with figures and tables of sensitivity studies. We therefore plan to include most of these results in an appendix, and retain in the main manuscript only those results/configurations that help support the main conclusions on the Bi-LSTM method.

3 The statement "The reconstruction of mean temperature series could provide a general assessment of the skill to reconstruct extreme temperature phases" needs either a citation or experimental results in Section 3. I think extremal behavior could be quite different than behavior near the mean? It would be interesting to see if the Bi-LSTM can model quantiles of the distribution better than PCR or CCA.

A: This sentence was meant to justify the consideration of the Bi-LSTM method in the first place. This method, in contrast to the usual set-up with the traditional PCR and CCA methods, naturally incorporates the serial correlations of the inputs. The working hypothesis is that this could improve the reconstruction of extremes, or at least provide a different reconstruction skill than for the mean values. Our objective was

to test this potential difference. We will expand in the revised version the justification of the use of the Bi-LSTM method and the analysis of the skill of all three methods for the specific replication of extremes. The evaluation of the replication of extremes will also need to address not only the amplitude of spatial averages or of the indices, but also the spatial patterns that occur during those extremes. For instance, it will be interesting to show whether deficiencies in the replication of the amplitude of extreme indices or spatial averages is brought about by a general overestimation of the extreme amplitudes or by an incorrect replication of the sign of spatially resolved anomalies, which would cause a subdued extreme after spatial averaging.

4 What is the rationale for training on 850-1425 and then testing on the later period of 1426-2000 (where did 2000-2005 go?), rather than the reverse as in Steiger et al. (2014)? I think that in a paleoclimate experiment we would be more interested in the performance of our method in the relative past, rather than the relative future. Would the performance of different methods change with the temporal order of training and testing?

A: We agree and will replace the calibration period with the 20 century, and use the rest of the past last millennium time period as validation (also in response to a comment by Reviewer 2). This will provide an even more realistic test of the methods and also address the next comment.

5 In practice, the Bi-LSTM would need to be trained on real proxies and real observations, which would limit the training period to 1850 onwards. Will this be enough observations, over a long enough time horizon, to train an LSTM model? Comparisons between the various methods under this limited data setting would be helpful. Also, how will you account for the significant covariate shift between the post-industrial land pre-industrial periods?

A: This is a relevant question, which partially explains our unusual choice of the calibration period. We agree with the reviewer that our choice of calibration period cannot be implemented in a real application. We wanted, however, to ameliorate the problem of the estimation of the covariances when using a small (and anthropogenically contaminated) sample size. This would have allowed us to better identify the methodological differences per se. In view of the opinion of the reviewers, we will now replace our choice of the calibration period and use the 20th century, as indicated in the response to the previous comment. We will do PPEs limiting, for example, the training period to 1900 onwards to check its performance. Degradation of performance may well be expected when a limited dataset is used to train a neural network model (Najafabadi et al., 2015). Also splitting data inappropriately might cause unexpected effects on the general performance of one neural network model. In typical neural network tests in the ML context, one splits the data at random for most tasks but available data in the paleoclimate context is of course not random. As the reviewer rightly states, the data might contain trends or shifts in covariance in time – these could result from changes in the way the data was gathered or from varying choices over what information to collect (Riley, 2019) as well as from changes in climate conditions. This is a problem that is inherent to most data-driven climate reconstruction methods, which assumes that the covariance patterns learned from the training data (usually in the 20th century) are stationary enough to also represent the covariance during the reconstruction period (as we have mentioned also in ll.196-197 and ll.220-221 for the PCA and CCA). One objective of the pseudo-proxy experiments is precisely to test the consequences of this assumption.

6 Lines 225-235 seem to motivate including temporal correlation in a model more generally, rather than LSTMs specially. Since methods like Data Assimilation already model time varying processes, what is the potential benefit of the LSTM? This section should contain a more clear and comprehensive justification of the LSTM to motivate it over existing time series techniques.

A: We will add a more clear justification for employing the Bi-LSTM in our manuscript. The main motivation of the usage of a purely data-driven method, in contrast to a data-assimilation methodology, is its simplicity in practice, as no climate model simulation is needed.

However, we do not totally agree with the reviewer that data-assimilation methods automatically take into account the serial correlation of the system. They do in the field of numerical weather prediction, for example, but off-line data assimilation in paleoclimate do not general consider serial correlations. A an usual set-up of the Kalman filter in paleoclimate data assimilation, the prior state PDF is sampled from a climate simulation and the posterior is updated using information from the proxy data. No information about the serial correlation flows into this updating step.

7 Comparing Table 1 and 2, why is SD ratio replaced with RMSE, particularly since RMSE was not mentioned as a comparison metric in the beginning of Section 3. Also, RMSE needs to be defined or at spelled out once.

A: We employ RMSE metric for quantifying the bias error between targets and reconstructions. We will define RMSE. The rationale for not using the RMSE metric for the spatially resolved reconstructions is that this metric is also determined by the variances of the target and the reconstructions, and this variance is spatially not uniform. This makes the interpretation of a spatial field of RMSE more difficult. As one of our main objectives is the correct replication of the amplitude of variations, we think that this information is better conveyed by separately showing the correlation and the ratio of variances. For singled valued spatial averages or indices, this problem is not present and we will then provide all three measures, correlation, ratio of variances and RMSE.

8 Line 156 states that "We then perturb the ideal pseudo-proxies with Gaussian white noise ... with signal-to-noise ratio (SNR) values of 0.25, 0.5 and 1", but then later on line 306 it states "More realistic pseudo-proxies are those containing 80% Gaussian white noise contamination.", and it would seem that the 80% contamination is used in all of the experiments. How is this 80% number connected the previously stated SNR values?

A: The noise level can be expressed using various definitions including SNR, variance of pure white noise (NVAR), and percentage of noise noise in the total variance variance (PNV) (Smerdon, 2012). Each of them can be readily translated: for example, 80% PNV corresponds to a SNR of 0.5. We will define PNV and use it uniformly through the manuscript to avoid confusion.

9 On line 395 – "The Bi-LSTM is able to capture periods of extreme cooling better than the other two methods but strongly underestimates the recent warming trend." Is it possible the LSTM is just biased towards colder temperatures?

A: We will test whether this property generally holds for different LSTM architectures or whether the LSTM is more sensitive than the other methods to the choice of the training period, since we used in the first version a generally colder training period. Another possibility is that the replication of extremes by this method is indeed not symmetrical. When analyzing the behavior of the methods and reconstructing extremes, we will also pay attention to any asymmetry in the reconstructed distribution functions. We will also investigate in a more detailed way whether the reconstruction skill is different for the multidecadal means or for extreme annual temperatures.

10 The figures need to be referenced more heavily in the text. Statements such as In addition, cc maps show higher values over regions where more pseudoproxies are located." and The Bi-LSTM and PCR methods

exhibit relatively consistent patterns with similar SD ratios" seem to refer to the content of a plot. Without an explicit reference though its hard to follow.

A: We will take care to reference the figures more frequently throughout the text.

11 I think section 2.2.1. Construction of pseudo-proxies should be grouped in with the Data section 2.1, rather than the Methods section 2.2.

A: This point can be considered a matter of taste. Actually, the data we used stem from climate simulations. The construction of pseudo-proxies is more a methodological issue, as pseudo-proxies can be constructed differently from the same underlying data. Our suggestions is to now include a Data and Methods section, separated in subsections.

3. Technical corrections

We will proofread the whole manuscript and correct grammatical and informal mistakes. We will make all changes considering the stated technical corrections.

Reference:

Najafabadi, M.M., Villanustre, F., Khoshgoftaar, T.M. et al. Deep learning applications and challenges in big data analytics. Journal of Big Data 2, 1 (2015). https://doi.org/10.1186/s40537-014-0007-7.

Riley, P. Three pitfalls to avoid in machine learning. Nature 572, 27–29 (2019).

Smerdon, J. E.: Climate models as a test bed for climate reconstruction methods: pseudoproxy experiments, Wiley Interdisciplinary Reviews, Clim. Change., 3, 63–77, https://doi.org/10.1002/wcc.149, 2012.