Comment on CP-2022-5

1 General comments

This paper compares the reconstructive skill of three climate field reconstruction (CFR) methods in an array of pseudoproxy experiments. The authors compare the traditional approaches of principal component regression (PCR) and canonical correlation analysis (CCA) against a proposed Bi-directional Long Short Term Memory (Bi-LSTM) neural network method. Results show that PCR tends to out perform CCA and the Bi-LSTM in reconstruction NAE and NH summer temperature field. However, the Bi-LSTM shows some promise due to its improved ability to estimate the lower tails of the distribution.

I think that the application of LSTMs for CFR an interesting idea, but I do not think this paper goes far enough in its analyses to draw strong conclusions about its effectiveness. Only one LSTM architecture (with varying widths) is presented, while it is customary in deep learning research to evaluate a range of architectures and report the results for a collection of models. Additionally, by only comparing against PCR and CCA, the paper omits some potentially interesting comparisons against more sophisticated techniques like Data Assimilation (DA), which can improve over PCR in pseudoproxy experiments (Steiger et al., 2014). Finally, analysis of the results is largely descriptive of the plots, but does not delve deeply into possible explanations. For instance, its not clear why might the proposed Bi-LSTM model behave so similarly to PCR or why the Bi-LSTM captures the low end of the distribution better than PCR.

Overall, I think a deeper exploration of Bi-LSTM architecture settings, comparisons against DA methods, and more climatological insight into the results are necessary to make the results compelling and generalizable.

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.

2. Followup 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.

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.

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?

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 and pre-industrial periods?

6. Lines 225-235 seem to motivate including temporal correlation in a model more generally, rather than LSTMs specifically. 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.

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.
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?
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?
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.
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.

3 Technical corrections

There are many grammatical mistakes and informal statements in this paper. It would be good if the authors could thoroughly proofread the paper once more and correct them. I list a few examples here:

1. Line 26 – “which hinders to capture” should be “which fails to capture”
2. Line 27 – “in earlier centuries” is too informal and needs to be specified. Do you mean prior to 1850?
3. Line 27 – “(such as tree rings, ice cores), etc.” should be “(such as tree rings, ice cores, etc.)” and should have a citation.

4. Line 75 – “over NH and NAE” these acronyms need to be defined explicitly.

5. Line 98 – remove “ Besides,”

6. Line 136 – “labelled past1000 and historical and labelled r1i1p1” makes it sounds arbitrary. I think you mean you “combined past1000 and historical simulations from ensemble member r1i1p1”

7. Line 266 – “Object function” should be objective function. Similarly on line 268 – “Object loss” should be “The loss” or just “We minimize the loss with gradient descent”

8. Line 295 – “all the spatial cc patterns exhibit similarities” this is too informal.

9. Line 324 – “These regions barely contain any pseudoproxies,” also too informal.

References

Steiger, N. J., Hakim, G. J., Steig, E. J., Battisti, D. S., and Roe, G. H. (2014). Assimilation of time-averaged pseudoproxies for climate reconstruction. *Journal of Climate*, 27(1):426–441.