Climate of the Past Discussions

Clim. Past Discuss., referee comment RC2
https://doi.org/10.5194/cp-2022-5-RC2, 2022

Comment on cp-2022-5

Anonymous Referee #2

Referee comment on "Evaluation of statistical climate reconstruction methods based on pseudoproxy experiments using linear and machine-learning methods" by Zeguo Zhang et al., Clim. Past Discuss., https://doi.org/10.5194/cp-2022-5-RC2, 2022

Summary: This manuscript applies a machine learning method known as Bidirectional Long-Short-Term Memory Neural Network (Bi-LSTM) to derive climate field reconstructions (CFRs) within the context of pseudoproxy experiments built on last millennium simulations generated with three different fully-coupled climate models. The results from Bi-LSTM are compared to results from two traditional CFR methods: principle component regression (PCR) and canonical correlation analysis (CCA). The authors find that all of the methods perform similarly, but skill metrics indicate that Bi-LSTM is not as skillful as the other two traditional methods. The one bright spot for Bi-LSTM is that is appears to capture cold extremes in NH and regional means better than the other two methods.

General Comments: The investigation of a non-linear machine learning method in the context of the CFR problem is interesting and important. It is useful to see how the method compares to more traditional methods and to see some well-established behaviors reproduced in the authors implementation of PCR and CCA, as well as in the Bi-LSTM method. Those behaviors include the loss of variance with the addition of realistic noise levels in the pseudoproxies and the tendency for skill to concentrate in well sampled areas of the pseudoproxy network. These findings are nevertheless not new and the presentation of the characteristics does not advance the field in a significant way. The application of Bi-LSTM is new, but the authors do not make a strong case for why this method should be applied. Of all the machine learning methods, why this one? Is Bi-LSTM particularly well suited for the CFR problem? Is the non-linear nature of the method or its incorporation of serial correlation important for the problem? Without strong arguments for why these characteristics are useful, the application of Bi-LSTM has the feeling of just being the method that the authors had sitting on the shelf. This should be remedied.

Another general concern is that the manuscript is largely just an application of the methods and a description of the results. There is little insight into *why* the methods might behave the way that they do. Does the Bi-LSTM perform similar or worse to the PCR and CCA methods simply because the problem isn't strongly non-linear? Does Bi-LSTM capture the cold extremes *because* it is non-linear? If that is the case, why does it better capture the cold extremes and not the warm extremes? These and other questions

are simply not taken up and the descriptive presentation of the results is not commensurate with the state of the science in terms of how the results are interpreted to understand how and why methods perform the way that they do. Consider some of the recent work in paleo data assimilation, in which the motivation is to incorporate new information in the form of climate model constraints, or in the Yun et al. (2021, https://doi.org/10.5194/cp-17-2583-2021) paper in which the authors seek to diagnose why various methods perform the way that they do. My argument is not to say that Bi-LSTM is not a worthwhile method, it is just to say that there is very little investigation in this paper that provides insight; it largely boils down to the application of a method and a descriptive report on how it performed. This is a rather modest accomplishment, especially if there is not a compelling reason why the method was applied in the first place.

In addition to the above general concerns, I provide comments on specific lines in the manuscript below.

Specific Comments:

Ln 7: There are many places in the manuscript that use "summer season temperature." Summer season is redundant and can be changed to simply summer temperature.

Ln 22: This is a bit of a strange list of references for this general statement. I suggest the authors just list the many review articles on CE climate over the last decade or more: Mann and Jones (2003); Jones et al. (2009); Frank et al. (2010); Christiansen and Ljungqvist (2012); Smerdon and Pollack (2016); Christiansen and Ljungqvist (2017)

Ln 24: Again, this is a strange list of references for the sentence it is supporting. More appropriate references are: Hegerl et al. (2006, 2007); Schurer et al. (2013, 2014); Anchukaitis et al. (2012, 2017); Tejedor et al. (2021 PNAS and PP)

Ln 26: "hinders to capture" is not grammatically correct.

Ln 28: "ice cores), etc." is incorrect structure.

Ln 41-40: "Many scientific studies that employ pseudo-proxies and real proxies have focused on global and hemisphere climate field or climate index reconstructions..." What else is there? This is basically everything unless the authors are thinking about recons of dynamic indices or want to point out that the majority have focused on specifically temperature recons (with the exceptions of the data assimilation methods that have tested multiple variable recons in pseudoproxy studies).

Ln 51: Data assimilation isn't mentioned at all in the Introduction, which overlooks a rapidly expanding area of CFR research and production in the field right now. It is relevant here inasmuch as the method does not assume temporally stable relationships between proxies and the targeted climate variables.

Ln 53: Coats et al. (2013, doi:10.1002/grl.50938) and Yun et al. (2021) specifically take up the stationarity assumption.

Ln 71: Decadally filtered after the forced global warming signal has been removed (usually via detrending).

Ln 91: It is not clear what is being combined here. It was stated above that they use the PAGES2k network combined with a tree-ring network from St. George. Here they say they are combining mollusk shell records with PAGES2k and Luterbacher et al. The inclusion of the mollusk shell records is a bit random, as I am not sure they have been included in a large-scale CFR to this point. In a synthetic experiment like this, it seems a bit ad hoc to create a sampling based on a theoretical combination of proxies that, to my knowledge, has never been adopted before (I am not aware of a large-scale application of the St. George assessment, unless the authors are using that reference to refer to a large-scale sampling from the ITRDB). Just using the PAGE2k sampling seems sufficient and straightforward here.

Ln 95: The choice of climate model has been definitively shown to impact the pseudoproxy results: Smerdon et al. (2011, 2015) and Parsons et al. (2021, https://doi.org/10.1029/2020EA001467)

Ln 132: Which ensemble member? Also from not form.

Ln 134: The CCSM4 model is presented as if it is distinct from the CESM, when in fact they share a very close lineage. This should be mentioned and does not make what the authors have done to be three truly independent models because of the close lineage between CCSM4 and CESM.

Ln 150: The grid cell that contains the proxy location is probably more accurate.

Ln 156: It is useful to point out that white noise is not realistic and that there have been attempts to use other noise colors or noise simulated by proxy system models. Noise colors were investigated in the seminal von Storch et al. (2004) paper, Wang et al. (2014, https://doi.org/10.5194/cp-10-1-2014) investigated various noise structures, and Evans et al. (2014, https://doi.org/10.1002/2014GL062063) investigated pseudoproxy

experiments with noise from proxy system models.

Ln 165: I strongly disagree with what the authors have done to split up the calibration and validation intervals. They use a much longer training interval than would ever be possible in the real world and they calibrate outside of the 20th century when the strong trend therein may have important impacts on their methodological performance. Given the descriptive nature of this study, it is weakened even more if the conditions under which the methods are tested are far outside of what is possible with real data. I strongly encourage the authors to complete the study over a more realistic calibration interval length and in the 20th century. Absent these more realistic constraints, the skill measures the authors provide are probably inflated and impossible to interpret for more realistic frameworks.

Equation 1: The PCR and CCA formalism is inexplicably written in series form. Why not use the much more traditional formulation using matrix notation? The relationship between PCR and CCA is also more evident using matrix notation, in which PCR is simply a special form of CCA, i.e. it does not reduce the rank of the cross correlation matrix. This relationship should also be noted.

Ln 192: Residual term with what assumed properties?

Paragraph starting on Ln 279: This paragraph is full of undefined jargon that is not cited. It is meaningless for the uninitiated. Please correct.

Ln 299: This was first noted in Smerdon et al. (2010, 2011).

Ln 310: "reduction in skill" as opposed the vague use of degradation here?

Ln 320: In the spirit of my general comments, one curiosity is why CCA does not perform better than PCR with regard to the cc metric. CCA is designed to optimize the correlation, which is why it can sometimes yield larger variance losses. It is therefore curious why it doesn't universally beat out PCR in the cc metric.

Ln 327: The variance losses have a relatively straightforward interpretation for the traditional regression approaches. When analyzing the mean results, the variance losses reflect loss of signal (reflected in the mean) and increases in the variance associated with the error term. It would be useful to know if the machine learning method can be interpreted in the same way, or if there is an alternative way to think about variance losses for that method.

Ln 390: Why should complexity translate to improved skill? I am aware of no grand postulate that makes this case.

Ln 402: The relationship (or lack thereof) between the skill of the mean indices and spatial skill was first discussed in Smerdon et al. (2010, 2011) and further highlighted in Smerdon's 2012 pseudoproxy review.

Ln 418: This is vague. What about alternative methods might be useful in the context of the CFR problem? There are lots of methods out there. What direction can the authors provide, based on the work they have done, that might represent useful characteristics in other machine-learning methods to try in the context of this problem?

Ln 451: CCA is a classic linear-based CFR method. This structure is awkward.

Ln 460: "Reservoir Computing methods-Echo State Network" is screaming for a reference so that the rest of us can figure out what it is.

Figures 6 and 7: Much of the text in this figure would only be legible by Ant Man. I strongly suggest increasing the size of the legend, fonts, and axis labels.

Figures 8 and 9: I find these figures very hard to read. Why include the bar plots for the data bins? It would be much clearer to simply show the estimated PDFs, which characterize the behavior well enough.