Response to reviewer 3

Reviewer: Namely, I tend to agree with one of the other reviewers that there is not enough direct comparison between the strength of the BARCAST methodology and the other CFR techniques so as to further test the fallout of the assumptions regarding AR1 vs. fGn data. While the authors do review the other techniques in detail which is informative, a direct comparison on the data they've generated in this manuscript using the other available tools would make this much stronger.

Response: We agree that this type of comparison would be interesting, but we think that the results of our study are important enough to stand alone. The paper brings in a number of relevant aspects that justifies the anticipated length/extent. These will be elaborated and brought better into focus in the revision, and include the novel method of generating target data and the discussion of proper scoring rules/elaboration on the CRPS.

Though we realize undoubtedly that evaluation of skill for other CFR techniques with a similar set of experiments would be highly relevant, this is unfortunately not an option at this point. The title of the manuscript will be revised to reflect that only BARCAST is considered, see also our answers and consideration in the general reply letter.

Reviewer:
The authors also have not included any discussion of the GraphEM CFR technique despite it's inclusion in Wang et al., 2015 (which they cite) ? should this CFR method not also be discussed in terms of relative performance? There are also a number of other citations that I believe should be added to the Discussion which I have listed below.

Response: the GraphEM method will be mentioned explicitly in the revision together with the other EM methods as follows:

Other reconstruction techniques that may experience similar deficiencies is the regularized expectation-maximization algorithm (RegEM), (Schneider, 2001; Mann et al., 2007), and all related models (CCA, PCA, GraphEM);

The discussion will be rewritten, and the references mentioned (Laepple and Huybers, 2014b,a) are familiar to us. We will consider citing these papers if appropriate. Furthermore, Ault et al. (2013, 2014) consider precipitation/hydroclimate which do not necessarily exhibit similar persistence properties as surface temperature. As an example, from reviewer 1 we were informed about Wahl et al. (2017), where it is demonstrated that reconstructed precipitation and instrumental data follow a white-noise process in time. For these types of data the multivariate regression-based reconstruction methods may be considered appropriate.

Reviewer: Finally, for the final paragraph of the Concluding Remarks, I was really left hoping for a more forward - looking statement about the future of this field and what your work contributes towards a broader knowledge of our estimates of past climate variability from proxies using these techniques. I think an effort could be made to solidify

your findings and put them in a broader scope at the very end of the paper and put your work in context. How does your work enhance our ability as a field to interpret CFRs of past climate variability? What are the truly broad impacts or your work?

Response: The discussion and conclusions will be revised, these points will be addressed as they are similar to some comments of reviewer 1.

Reviewer: Overall, a very nice manuscript. As a general comment, as a person who is not a true expert in CFR, it would be nice if the authors could make an attempt to help readers who may not be as familiar with these topics with real-world examples or layman?s terms where appropriate.

Response: Thank you! Reviewer 1 had many specific comments on subjects that will be described in more detail in the revision. In particular we have tried to extend the text in a number of places to make it more easy to comprehend for potential readers not directly involved in specific studies on CFR methods. If there are other particular segments that are unclear we kindly ask the reviewer to point these out.

Specific comments

The grammar has been formatted to comply with the detailed comments, below we respond to the questions the reviewer asks.

Reviewer: 12.16 Figure 9 doesn't get much description or introduction but you partition it away for some reason. Can you please give more information about why you say, for example, except Fig. 9? what is different about Figure 9 exactly? This comes up again on line 12.23.

Response: We have discovered an error in the calculation of the CRPS, see the general reply for details. Figure 9 will therefore be removed in the revision.

Reviewer: Page 15 Lines 1-15 I think you need to beef up your discussion here about millennium- long paleoclimate reconstructions, because there are quite a few more citations whose work should be added to the discussion here, especially those which include model- data comparisons, including: (refs)

Response: The focus of this paragraph is very limited, we do not discuss millennium-long paleoclimate reconstructions or model-data comparison in general. It deals with a very specific topic of scale breaks in the power spectrum of Earth's surface temperature over a range of time scales (Lovejoy and Schertzer, 2012; Nilsen et al., 2016). Our results in the present paper demonstrate that such spectral scale-breaks may occur due to proxy noise and/or incorrect model selection, hence the breaks are unrelated to the true climate variability. This means the correlation structure of paleoclimate reconstructions may differ from that of instrumental observations of climate, and it is why the concept of universal scaling laws across a wide range of time scales is perhaps less meaningful.

Regarding the references you mention, some of these are familiar to us and others concern precipitation/hydroclimate which is not our climate variable in focus. In the revised discussion we will try to have better balance and express the above point in a more clear way.

Reviewer: In Section 5.1, you go into the issues with TRW and your estimates of the PSD. The issues with TRW and de-trending methods, especially how this alters the power spectrum, is discussed in Section 3.2.5. of Dee et al., 2017 as listed above and you should compare your analysis to that paper as needed.
Response: Thank you for making us aware of the Dee et al. (2017) paper which is highly relevant, we will make sure to cite it in the revision. See also the reply to reviewer 1.

References

T. R. Ault, J. E. Cole, J-T. Overpeck, G. T. Pederson, S. St. George, B-Otto-Bliesner, C-A. Woodhouse, and C. Deser. The continuum of hydroclimate variability in western north america during the last millennium. *Journal of Climate*, 26(16):5863–5878, 2013. doi: 10.1175/JCLI-D-11-00732.1.

T. R. Ault, J. E. Cole, J. T. Overpeck, G. T. Pederson, and D. M. Meko. Assessing the risk of persistent drought using climate model simulations and paleoclimate data. *Journal of Climate*, 27(20):7529–7549, 2014. doi: 10.1175/JCLI-D-12-00282.1.

T. Laepple and P. Huybers. Ocean surface temperature variability: Large model-data differences at decadal and longer periods. *P. Natl. A. Sci.*, 111(47):16682–16687, 2014a. doi: 10.1073/pnas.1412077111.

T. Laepple and P. Huybers. Global and regional variability in marine surface temperatures. *Geophysical Research Letters*, 41(7):2528–2534, 2014b. doi: 10.1002/2014GL059345.

S. Lovejoy and D. Schertzer. *Low Frequency Weather and the Emergence of the Climate*, pages 231–254. 196. American Geophysical Union, 2012. doi: 10.1029/2011GM001087.

T. Nilsen, K. Rypdal, and H.-B. Fredriksen. Are there multiple scaling regimes in holocene temperature records? *Earth Sys. Dynam*, 7(2):419–439, 2016. doi: 10.5194/esd-7-419-2016.

E.. R. Wahl, H. F. Diaz, R. S. Vose, and W. S. Gross. Multicentury evaluation of recovery from strong precipitation deficits in california. *Journal of Climate*, 30(15):6053–6063, 2017. doi: 10.1175/JCLI-D-16-0423.1.