Response to reviewer 1

Specific comments

Reviewer page 1, title: The title needs to be adjusted to reflect that this study is specific to BARCAST, and does not address CFRs generally.
Response: The title of the revised manuscript will be changed, see general response.

Reviewer page 1, line 14: Does "Selected" here mean that some of the experiments were found to give reconstructions consistent with the AR(1) model, but others did not? If that is the case it needs to be described as such.
Response: The abstract will be moderately rewritten, see general response.

Reviewer page 2, line 11: This is not necessarily the case, which is why the word "can" needs to be added in line 10.
Ordinary Least Squares still provides optimal parameter estimation when the predictor variable has error, but the predictor variable is not correlated with the predictand error. (Econometrics, Wonnocott and Wonnocott, 2nd ed., 1979, John Wiley and Sons)
In multiple regression, the combined effect across the several predictors can be both reduction or enhancement of the estimated coefficients. (Applied Regression Including Computing and Graphics, Cook and Weinberg, 1999, John Wiley and Sons)
Response: Thank you for this additional information, the sentences will be moderately rewritten.

Reviewer page 3, line 10: Add a brief parenthetical expression here to denote to the readers what a Lorentzian power spectrum is.
Response: A Lorentzian power spectrum has a steep slope at high frequencies, and is flat at low frequencies.

Reviewer page 3, line 17: Add a reference here for this statement.
Response: References to Fraedrich and Blender (2003); Fredriksen and Rypdal (2016) are inserted.

Reviewer page 3, line 17: Note here why Gaussianity is important in this context. It does not seem to follow from the long-range memory properties theme of the rest of the paragraph.
Response: The fractional Gaussian noise follows a Gaussian distribution by definition. In order to model the Earth's surface temperature using the fractional Gaussian noise stochastic process, the temperature data cannot deviate too strongly from a Gaussian distribution. This will be made more clear in the revision.
There exists another, more general class of stochastic processes exhibiting LRM but not following a Gaussian distribution, but it is outside the scope of our paper to consider such processes.

Reviewer page 3, line 25: Add a reference here for this statement.
Response: References to Franke et al. (2013); Fredriksen and Rypdal (2016); Nilsen et al. (2016) are inserted.

Reviewer page 5, line 29: A brief explanation of what conjugate means in this context should be added here.
Response: conditionally conjugate priors means that the prior and the posterior distribution has the same parametric form. Added in the revision.

Reviewer page 3, line 31: Give a reference here for MCMC, the Gibbs sampler, and the Metropolis step. Would Gelman et al., 2003, be appropriate?
Response: Yes, reference to Gelman et al. (2003) has been inserted.

Reviewer page 7, Eq. 5: The nature of how eq. 6 is derived from eq. 5 may not be obvious/clear to some readers, and perhaps can be explained in a very brief appendix, or possibly by some additional text here.
Response: Eq. 6 is not derived from Eq. 5 per se. The exponential kernel function $\exp^{-(1-\alpha)\mathbf{I}(t-s)}$ is replaced with the power-law kernel $(t-s)^{\beta/2-1}$. This is the necessary operation to obtain the desired LRM properties of the target data, the idea stems from Rypdal (2012); Rypdal and Rypdal (2014).

Reviewer page 7, line 7: This statement, that there is no contribution from T(0), seems at odds with what is said in the next sentence, where the solution for t > 0 depends not ONLY on the initial condition, but the entire time history of T(t).
[Emphasis of ONLY added.]
This apparent contradiction should be resolved, so the reader is not confused.
Response: Sect.2.2 will be modified for clarity. The nature of the long-memory model is that the temperature at time t depends on all past values $[-\infty, t]$. The original form of Eq. 6 is therefore:

$$\mathbf{T}(t) = \int_{-\infty}^{t} (t-s)^{\beta/2-1} \boldsymbol{\epsilon}_{\mathrm{s}} \mathrm{ds} \tag{1}$$

The contribution before time $t = 0$ is then neglected, this means $\boldsymbol{T}(0)$=0 and we get the equation in the same form as Eq. 6 in the manuscript. We apologize for the confusion. Neglecting the contribution prior to $t = 0$ is a choice that is justified in this case.

Reviewer page 7, below line 13: Explain why epsilon(sub s) goes away here.
Response: $\epsilon_{\mathbf{s}}$ is a function of s only and not of t. For the kernel, on the other hand, we have:

$$\lim_{s \to t} (\mathrm{t} - \mathrm{s})^{\beta/2-1} = \infty$$

Hence, this is the term that needs to be adjusted to avoid the singularity.

Reviewer page 7, line 14: It can be unclear here to some readers why G is a function of tau, rather than t.
Please explain how this comes to be, and how tau here relates to the tau variance terms in the Normal distributions defined for the e values in Eq 3.
Response: A detail was missing in the text, namely that $\tau = t - s$. This τ is not related to the BARCAST parameters τ_I^2 and τ_P^2. In the revision we will use $t - s$ instead of τ. **G** must be a function of the time step $t - s$, with the unit step function

$$\Theta(t-s) = \begin{cases} 0, & t-s < 0 \\ 1, & t-s \geq 0 \end{cases}$$

Reviewer page 8, line 6: Explain this mathematical description more for the readers for whom this expression will not be meaningful as written.
Again, a brief appendix might be useful.
Response: The formulas for estimating the periodogram have been slightly rewritten and moved to an appendix. The alternative formulation is less compact and hopefully more readible:
The periodogram is defined here in terms of the discrete Fourier transform H_m as

$$S(f_m) = \left(\frac{2}{N}\right)|H_m|^2, \quad m = 1, 2, \ldots, N/2$$

For evenly sampled time series $x_1, x_2, ...x_N$. The sampling time is an arbitrary time unit, and the frequency is measured in cycles per time unit: $f_m = \frac{m}{N}$. $\Delta f = \frac{1}{N}$ is the frequency resolution and the smallest frequency which can be represented in the spectrum.

Reviewer page 9, line 14: Wouldn't it be more appropriate here if the confidence range for the theoretical spectrum is generated from the distribution of analogous MEAN spectra generated by the MC process?
Response: The approach we are using is meaningful and appropriate. The simulated fGn Monte Carlo ensemble has, on average, the desired statistical properties of theoretical fGn processes. The hypothesis testing involves finding out if the reconstruction ensemble on average has the same statistical properties. The spectral shape is the prominent feature we investigate. The spectral shape of the full fGn Monte Carlo ensemble is practically identical to that of the mean spectrum of the same ensemble. However, the width of the confidence range would be drastically narrowed if the latter type was used, since the mean fGn spectrum is less noisy than the spectrum of an individual ensemble member. Such a narrow confidence range is impractical for our experiment and most other real-world applications, where the data at hand may follow the general power-law

spectral shape, but are not expected to be strictly identical at every single point.

Reviewer page 11, line 26: Explain why it is improper in this context.
Response: The reduction of error (RE) and coefficient of efficiency (CE) are not suitable for ensemble-based reconstruction in general (Gneiting and Raftery, 2007). For probabilistic forecasts, scoring rules are used to measure the forecast accuracy, and preferably proper scoring rules. Proper scoring rules are built around the concept of reward systems, encouraging the forecaster to be honest, use the state of knowledge or personal beliefs. Proper in this sense means that the maximum reward (CRPS score=0) is given when the true probability distribution is reported. The RE and CE are improper scoring rules, meaning they measure the accuracy of a forecast, but the maximum score is not necessarily given if the true probability distribution is reported. For climate reconstructions, RE=1 and CE=1 implies a deterministic forecast, the maximum score is obtained when the mean (a point measure) within a probability distribution P is used instead of the predictive distribution P itself.

Reviewer page 11, line 30: It would be good to add distribution graphics of the skill metrics across their relevant ensembles.
Response: This is a good idea, we suggest to include a boxplot in every panel of Figs. 6-9, similar to Figs. 2-5 in Werner et al. (2013).

Reviewer page 12, Eq. 11: Explain what the E(sub F) notation means here.
This is done – to an extent – in context in the third sentence following (starting in line 4), but the nomenclature E(sub F) itself should be fully described.
It is not clear what the E operator does.
Response: $\mathbb{E}$ is the expectation value. $\mathbb{E}_F$ denotes the expectation value of the cumulative distribution function. Another form of this equation will be used in the revision.

Reviewer page 12, line 12: The explanations for average potential CRPS and Reliability need to be augmented with how these are identified in the terms of eq. 11 itself.
This is not clear, and since these are relatively new metrics vis-a-vis paleoclimate reconstruction, it will be highly useful for them to be more concretely explained in terms of the formal mathematical terms of eq. 11.
Response: Given the reviewer's interest in these concepts we have decided to rewrite section 4.4 in the revision and include additional explanatory text in the appendix so that it is clear why the RE and CE are improper, and elaborating on the CRPS. See suggested revision in the general reply.

Reviewer page 12, line 29: Help explain for the reader what this combination of good agreement with the target, but not with the confidence confidence range, means.
Response: We have discovered an error in the calculation of the CRPS, see our general reply. The above statement is no longer correct and will be removed. The recalculated Reliability is consistently low and very close to zero, the average CRPS is therefore to a large extent dominated by the $\overline{\mathrm{CRPS}}_{\mathrm{pot}}$. Sect. 4.4.1 will be rewritten, and Figure 9

removed from the revision.
For visualization of the average CRPS reconstruction skill, see Fig. 1 and 2 below. Figure 1 shows two examples of local target fGn times series with $\beta = 0.75$, SNR $= \infty$ (black) and the 95% confidence range of the reconstruction ensemble in red, based on 1500 ensemble members. The plot (a) is for a proxy location, while (b) is for a location between proxies. For (a) the confidence range is extremely narrow, as the reconstructions are nearly identical to the target. For (b), the ensemble has a larger spread, a few target values fall outside of the confidence range and are therefore considered outliers.

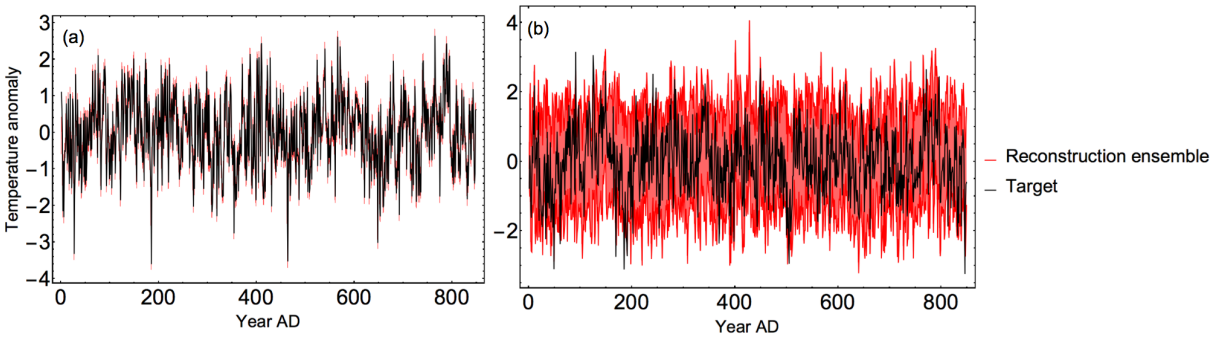

Figure 1: Example of local target (black curve) and 95% confidence range of reconstruction ensemble (red). Parameter $\beta = 0.75$, SNR=∞. (a) is for a proxy location, while (b) is between proxy locations.

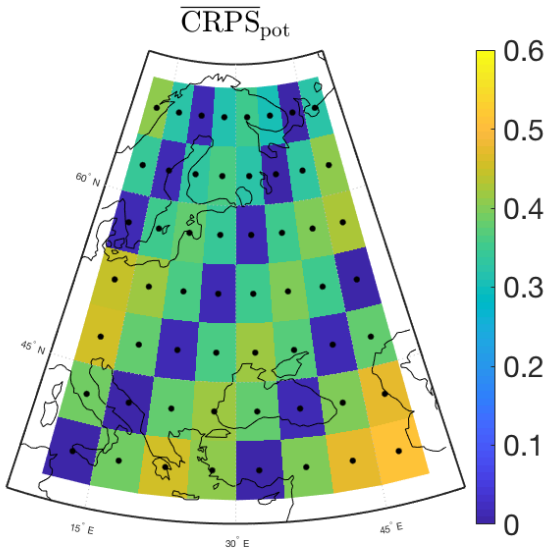

Figure 2: $\overline{\text{CRPS}}_{\text{pot}}$ for the parameters $\beta = 0.75$, SNR=∞.

Figure 2 shows the $\overline{\text{CRPS}}_{\text{pot}}$ for the same parameter setup and all locations. This skill score reflects the average accuracy of the reconstructed forecast, best at proxy locations and decreasing away from proxy sites. The low Reliability is not visualized, but in general this indicates that the predicted confidence ranges are in line with the actual reconstructed ensemble. This means that the target value is generally located in the centre of the confidence range, in contrast to high Reliability occurring if the target to large extent is biased towards the upper/lower end of the confidence range. The combination of a low $\overline{\text{CRPS}}_{\text{pot}}$ and high Reliability is still a possible outcome for real reconstructions, as observed for a few locations of Fig. A1 of Werner et al. (2018).

Reviewer page 12, line 33: The most general conclusion is that BARCAST performs much better, except re: reliability, where there are co-localized proxy data. This needs to be clearly mentioned here.
This stands in contrast to some other CFR results, where good skill is obtained in regions not co-localized with the predictor data.
E.g., cf. reconstruction of precipitation in California (Wahl et al., 2017, J Clim, DOI: 10.1175/JCLI-D-16-0423.1.)
Response: We thoroughly inspected Wahl et al. (2017) which the reviewer is referring to and two more publications (together with supplementary materials) that could be relevant in the context of this comment, namely Wahl and Smerdon (2012) and Diaz and Wahl (2015). Unfortunately, we found no indication that the issue raised by the reviewer has been discussed directly in any of the aforementioned studies. One can assume the reviewer refers to Fig S2 and S4 from Wahl et al. (2017), and the corresponding tree ring proxy network for the region depicted in Fig S2 in Wahl and Smerdon (2012) From these figures, we may infer a somewhat lower, yet mostly non-negative, reconstruction skill for the area centered at ca. 36N 117.5W, featuring a relatively high proxy data density. At the same time, higher skill is obtained in regions without proxy data coverage. However, there is no discussion on the source of these spatial discrepancies (at least we could not find one) other than that in the spatial mean sense, the method demonstrated good performance, which indeed is the case. We take the liberty to speculate that the discrepancies in the reconstruction spatial skill can be caused by a number of factors other than the apparent effect of the choice made on the climate field reconstruction technique. First, target data processing may cause spatial variability in the reconstruction skill. The choice of regridding/interpolation/extrapolation method will have an effect on the target data variance across timescales, especially in the data-poor regions. The second point is that the area with lower RE/CE is associated with proxies from mountainous regions, where regridding might have an effect on the target data sets since climate divides are present.
Should the reviewer find the question critical to be elaborated further, we kindly ask to clarify where this problem is discussed. However, in our opinion this would be fairly difficult to resolve without dedicated experiments with truncated EOF-principal components spatial regression (TEOF-PCSR) on the equivalent target/pseudoproxy datasets.

Reviewer page 13, line 4: Indicate which figures and/or tables provide the information

being described as the Discussion section proceeds.
Response: This will be done in the revision.

Reviewer page 13, line 24: This statement needs to be further illuminated for the reader. Aren't the same equations used for the reconstruction at all sites?
Response: These sentences will be rewritten, sorry for the confusion. Yes, the same equations are used for all grid cells, what was meant is that there is already information from the data at proxy sites, overwhelming the priors.

Reviewer page 13, line 26: Explain this statement further, in terms of how the resulting reconstruction is influenced.
Response: The three levels of BARCAST connects the model equations with the proxy and instrumental data, and with the priors for all parameters and the temperature field. The parameters are therefore interdependent, so adjustment of one parameter must be compensated by the other parameters. The effect of interdependence between the AR(1) coefficient and σ^2 is demonstrated in Fig. 9 in Tingley and Huybers (2010), where the posterior draws of ϕ and σ^2 are negatively correlated. Tingley and Huybers (2010) claim that for their version of barcast the pdfs of the final draws are not ill-defined, they rather converge to some specific values with relatively narrow pdfs.

For the reconstructions, increasing the AR(1) parameter α means changing the temporal correlation structure of the reconstruction at all frequencies, where the high-frequency component is forced to have stronger temporal correlations and the low-frequency component is forced to have weak or no temporal correlations. Increasing β_0 means the proxy bias is increased, so the relationship between the proxy and the true temperature is influenced. At last, decreasing σ^2 means that the white-noise innovations are given less weight, hence the coherent signal is less influenced by local, stochastic perturbation.

Reviewer page 13, line 28: This result is not very apparent in Fig. 5 a-c.
Response: The sentence mentions specifically "noisy input data", so b-d in all figures. But the effect is indeed minimal in Fig. 5b-c, this will be changed in the revision.

Reviewer page 14, line 4: Explain why the characteristic of subsequent years being independent is involved in the discussion here.
That is, relate this characteristic to an incorrect statistical model.
Response: The end of this paragraph will be rewritten in the revision:
The assumption of temporal independence corresponds to yet another incorrect statistical model for our target data; a white noise process in time. Note that for target variables/data sets consistent with a white noise process, these types of reconstruction methods are appropriate, as demonstrated using the truncated EOF-principal components spatial regression (TEOF-PCSR) methodology on precipitation data in Wahl et al. (2017).

Reviewer page 14, line 23: Explain here the nature of the incorrectness in the confidence intervals.
Response: see reply above to comment on page 12, line 29.

Reviewer page 15, line 15: Are there thoughts that come from these experiments for how BARCAST itself might be improved?
Are there more realistic characterizations of the fundamental mathematical temporal and spatial specifications that could be recommended?
And if so, what would the numerical estimation ramifications also be?
These kinds of discussion components would be good to add, especially since this study is by construction an evaluation of BARCAST's performance.
Response: There are two aspects that will be addressed and elaborated in the revised discussion regarding this comment. It is not only BARCAST that could be improved, the availability of well-documented high-quality proxy records also helps the analyst select an appropriate reconstruction method based on the input data. Investigating the temporal correlation structure is an exercise that can be done at different stages of data manipulation, and we stress that the spatiotemporal covariance structure of reconstructions depends on model/method selection. Hopefully this article can draw the attention of scientists working with proxy data sampling and processing, reconstruction methodology as well as those using such reconstructions for statistical modeling. Our study is meant to improve understanding in these communities, create awareness and enhance communication between them. Modellers need to know as much as possible about what artifacts their data are subject to and which reconstructed time scales are considered most reliable. On the other hand, the producers of proxy data need to know which information is needed in order to provide this information if it exists.

Point 1 - Improvements regarding proxies
if the proxy network is of high quality and density, and exhibit LRM properties, the BARCAST methodology without modification should be capable of constructing skillful reconstructions with LRM preserved across the region. This is because the data information overwhelms the vague priors. For assessment of real-world proxy quality it is useful to quantify/model the uncertainties and noise affecting the different proxy types and/or specific reconstructions. Forward modelling of proxies are important tools for this task that we endorse, see for example Dee et al. (2017) for a comprehensive study on terrestrial proxy system modeling and the recent paper by Dolman and Laepple (2018) on forward modelling of sediment-based proxies.

Point 2 - Improvements of BARCAST
For the BARCAST CFR methodology, what would drastically improve the performance in our experiments would be reformulation of model Eq. 1-2. However, we cannot guarantee that modifications favoring LRM are practically feasible in the context of a Bayesian hierarchical model, due to higher computational demands. Changing the AR(1) model assumption to instead account for LRM would in the best scenario slow the algorithm down substantially, and in the worst scenario it would not converge at all.

Some cut-off time scale would have to be chosen to ensure convergence. Regarding the spatial covariance structure, accounting for teleconnections introduce similar computational challenges. The more general Matérn covariance family form has already been implemented for BARCAST, but was not used in this study. Another problem is the potential temporal instability of teleconnections. The fact that major climate modes might have changed their configuration through time is now considered realistic. Therefore, setting additional a priori constraints on the model may not be considered justified. The use of exponential covariance structure appears to be a conservative choice in such a situation.
The discussion will be rewritten in the revision to address requests also from reviewers 2 and 3.

Reviewer page 15, line 28: In Fig. 3c-d the increased power of the simulated proxy data is at sub-decadal frequencies, not at higher-to-bidecadal frequencies.
If the characteristics mentioned are for real proxy data, that needs to be clarified.
Response: The characteristics referred to are the spectra in fig 3c-d, not the real proxy data. To be more precise and general, we will rewrite this sentence:
The characteristic flat spectrum at high frequencies, and the increased power on (sub)-decadal frequencies and lower for (Fig 3c) and Fig. 3d respectively can give the impression that the low-frequency power is inflated.

Reviewer page 16, line 1: RCS is only one among a set of tree-ring processing techniques, and this should be made clearer here.
Response: You are absolutely right. The focus on the RCS is due to the fact that it is a commonly used method which may cause artifacts as those we discuss. For better balance we will also mention other methods in the revision, such as the age band decomposition (ABD) and signal-free processing mentioned in your next comment.

Reviewer page 16, line 11: The "signal free processing" method of tree ring standardization should also be discussed here.
To the extent that the tree ring records involved are also sensitive to moisture, then persistence from one year to the next that is related to biological growth responses to soil moisture persistence can also affect these records. This is different from the standardization issue, pre se.
Cf. Bunde et al., 2013, Nature Clim. Change, doi:10.1038/nclimate1830.
Response: This is a good point, which we can state this in the revision for clarity. We are aware of the paper by Bunde et al. 2013, and we will consider citation.

Reviewer page 17, line 7: It would be good at this closure point to note how the present study sets a useful precedent for utilizing a carefully-designed, experimental structure to isolate specific reconstruction technique properties – akin to controlled laboratory experiments.
This is noted briefly in line 20 of page 16, but it would be good to highlight this point more broadly, since it is the real power and value-added of this paper.

Response: The discussion will be rewritten, and this particular comment will be addressed.

Reviewer page 22, Figure 1: Make the lines for the ensemble means significantly wider, so they can be seen better by the reader.
Response: Thank you for noting this, it will be fixed for the revision.

References

S.G. Dee, L.A. Parsons, G.R. Loope, J.T. Overpeck, T.R. Ault, and J. Emile-Geay. Improved spectral comparisons of paleoclimate models and observations via proxy system modeling: Implications for multi-decadal variability. *Earth and Planetary Science Letters*, 476:34 – 46, 2017. doi: https://doi.org/10.1016/j.epsl.2017.07.036.

H. F. Diaz and E. R. Wahl. Recent california water year precipitation deficits: A 440-year perspective. *Journal of Climate*, 28(12):4637–4652, 2015. doi: 10.1175/JCLI-D-14-00774.1.

A. M. Dolman and T. Laepple. Sedproxy: a forward model for sediment archived climate proxies. *Climate of the Past Discussions*, 2018:1–31, 2018. doi: 10.5194/cp-2018-13.

K. Fraedrich and R. Blender. Scaling of Atmosphere and Ocean Temperature Correlations in Observations and Climate Models. *Phys. Rev. Lett.*, 90:108501, 2003. doi: 10.1103/PhysRevLett.90.108501.

J. Franke, D. Frank, C. C. Raible, J. Esper, and S. Bronnimann. Spectral biases in tree-ring climate proxies. *Nature Clim. Change*, 3(4):360–364, 2013. doi: 10.1038/NCLIMATE1816.

H.-B. Fredriksen and K. Rypdal. Spectral Characteristics of Instrumental and Climate Model Surface Temperatures. *Journal of Climate*, 29(4):1253–1268, 2016. doi: 10.1175/JCLI-D-15-0457.1.

A. Gelman, J. Carlin, H. Stern, and D. Rubin. *Bayesian Data Analysis. 2nd ed.* Chapman & Hall, New York, 2003. 668 pp.

T. Gneiting and A. E Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477):359–378, 2007. doi: 10.1198/016214506000001437.

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.

K. Rypdal. Global temperature response to radiative forcing: Solar cycle versus volcanic eruptions. *Journal of Geophysical Research: Atmospheres*, 117(D6), 2012. doi: 10.1029/2011JD017283.

M. Rypdal and K. Rypdal. Long-memory effects in linear-response models of Earth's temperature and implications for future global warming. *J. Climate*, 27(14):5240–5258, 2014. doi: 10.1175/JCLI-D-13-00296.1.

M. P. Tingley and P. Huybers. A bayesian algorithm for reconstructing climate anomalies in space and time. part —: Development and Applications to Paleoclimate Reconstruction problems. *Journal of Climate*, 23(10):2759–2781, 2010. doi: 10.1175/2009JCLI3015.1.

E. R. Wahl and J. E. Smerdon. Comparative performance of paleoclimate field and index reconstructions derived from climate proxies and noiseonly predictors. *Geophysical Research Letters*, 39(6), 2012. doi: 10.1029/2012GL051086.

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.

J. P. Werner, J. Luterbacher., and J. E. Smerdon. A Pseudoproxy Evaluation of Bayesian Hierarchical Modeling and Canonical Correlation Analysis for Climate Field Reconstructions over Europe*. *J. Climate*, 26(3):851–867, 2013. doi: 10.1175/JCLI-D-12-00016.1.

J. P. Werner, D. V. Divine, F. Charpentier Ljungqvist, T. Nilsen, and P. Francus. Spatio-temporal variability of Arctic summer temperatures over the past 2 millennia. *Climate of the Past*, 14:527–557, 2018. doi: 10.5194/cp-14-527-2018.