Author response on manuscript cp-2018-17

This document contains a list of relevant changes in the manuscript, followed by point-to-point responses to review reports 1-3 and a markup file tracking changes in the manuscript. Note that our replies 1-3 refer to a general response that is not included in this document, instead we refer to our list of relevant changes. The general response can be found in the online discussion.

List of significant changes

In the revised manuscript, Martin Rypdal will be added as an author as he has contributed significantly to Sect. 2.2 on target data generation. Rypdal's contribution has grown since the original submission so that co-authorship is appropriate.

In addition to the changes listed below, references have been added in a number of places.

1. The new title of the manuscript is:

   How wrong is the BARCAST climate field rconstruction technique in reconstructing a climate with long-range memory?

2. The abstract has been rewritten.

3. Sect. 2.2 on target data generation has been rewritten to avoid confusion and misconceptions.

4. The title of Sect. 2.3 has been changed, and the definition of the periodogram has been moved to appendix A.

5. A paragraph from the discussion has become a separate subsection (4.4).

6. The CRPS skill metric has been elaborated in Sect. 4.4 and Appendix B.

7. We have discovered an error in the calculation of the average CRPS scores that influences our skill metric results. Sect 4.5 and 4.6 are rewritten, and Fig. 9 is removed.

   Major changes to the discussion involve:

8. References to which figures/tables we refer are inserted as the discussion proceeds.

9. Text editing to avoid confusion in several paragraphs, ref comments from reviewer 1 and 3 in particular.

10. More precise conclusions regarding the skill metric results using the CRPS and the subcomponents $\overline{\text{Reli}}$ and $\overline{\text{CRPS}}_{\text{pot}}$.

11. Mentioning other tree-ring processing techniques in Sect. 5.1, including the age-band decomposition and the signal free processing method.

12. Relating our results to to the topic of proxy system modelling, ref. Dee et al. (2017); Dolman and Laepple (2018) in Sect. 5.1.

13. Discuss prospective recommendations to improvements of BARCAST in Sect. 5.2

14. Final conclusions follow the suggestion from reviewer 1 and summarize how the present study sets a useful precedent for utilizing a carefully-designed, experimental structure to isolate specific reconstruction technique properties.

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.

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.

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.

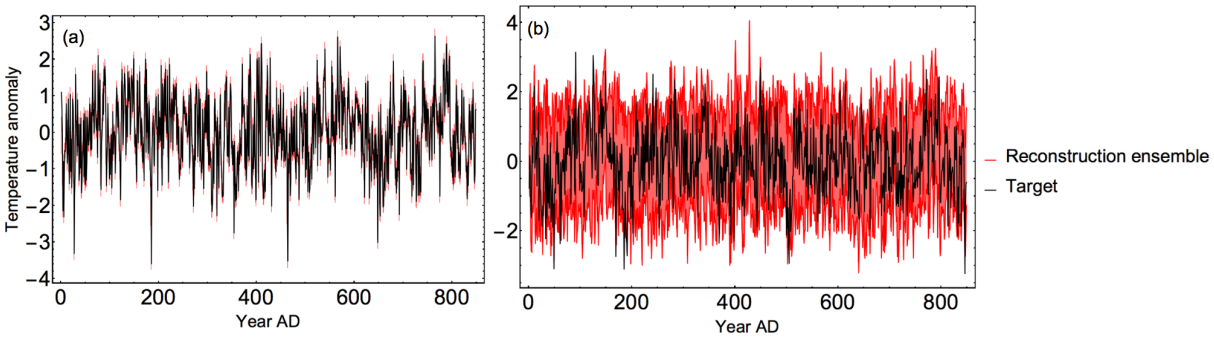


Figure 1: Example of local target (black curve) and 95% confidence range of reconstruction ensemble (red). Parameter $\beta = 0.75$, SNR=∞. (a) is for a proxy location, while (b) is between proxy locations.

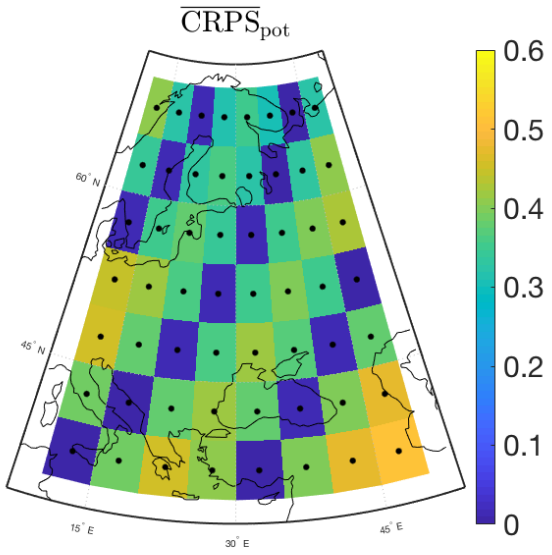


Figure 2: $\overline{\text{CRPS}}_{\text{pot}}$ for the parameters $\beta = 0.75$, SNR=∞.

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.

Response to reviewer 2

Thank you for providing general comments on the manuscript. The title and abstract of the revised manuscript will be changed, see the general response posted.
Reviewer: p.1 l.2 Citation needed for unsuitability of the CE and RE metrics.

Response: see our reply to reviewer one, who has similar concerns.

Reviewer: Figure 3 caption doesn't explain what each of the (a)-(b) panels are uniquely showing.

Response: The figure caption will be revised to make this clear.

Reviewer: I would recommend highlighting in both the abstract and the conclusions how the authors very nicely were able to test the issue of long-range memory in isolation by constructing the spatial fields statistically rather than through climate models. I think this is important to highlight because it's not usually (or ever yet?) done.

Response: Thank you for this feedback, we will bring the novelty of the data generation into focus as requested. The equations in Sect.2.2 will be moderately rewritten to avoid confusion and misconceptions. The methodology for data generation is unconventional but not unique, Werner and Tingley (2015) generate pseudo proxy data in a similar way, except the target data are formulated directly according to the BARCAST model equations.

References

J. P. Werner and M. P. Tingley. Technical Note: Probabilistically constraining proxy age-depth models within a Bayesian hierarchical reconstruction model. *Climate of the Past*, 11(3):533–545, 2015. doi: 10.5194/cp-11-533-2015.

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.

How wrong ~~are~~ is the BARCAST climate field reconstruction ~~techniques~~ technique in reconstructing a climate with long-range memory?

Tine Nilsen [1], Johannes P. Werner [2], Dmitry V. Divine [3, 1], and Martin Rypdal [1]

[1]Department of Mathematics and Statistics, UiT The Arctic University of Norway, Tromsø, Norway
[2]Bjerknes Centre for Climate Research and Department for Earth Science, University of Bergen, Bergen, Norway
[3]Norwegian Polar Institute, Tromsø, Norway

Correspondence to: Tine Nilsen (tine.nilsen@uit.no)

Abstract. The ~~Bayesian hierarchical model~~ skill of the state-of-the-art climate field reconstruction technique BARCAST ("Bayesian Algorithm for Reconstructing Climate Anomalies in Space and Time") ~~climate field reconstruction (CFR) technique, and idealized input data are used in the pseudoproxy experiments of this study. Ensembles of targets are generated from fields of~~ to reconstruct temperature with pronounced long-range ~~memory stochastic processes using a novel approach. The range of experiment setups include input data with different levels of persistence and levels of proxy noise, but without any form of external forcing. The input data are thereby a simplistic alternative to standard target data extracted from general circulation model (GCM) simulations. Ensemble-based temperature reconstructions are generated, representing the European landmass for a millennial time period. Hypothesis~~ memory (LRM) characteristics is tested. A novel technique for generating fields of target data has been developed and is used to provide ensembles of LRM stochastic processes with a prescribed spatial covariance structure. Based on different parameter setups, hypothesis testing in the spectral domain is ~~then~~ used to investigate if the field and spatial mean reconstructions are consistent with either the fractional Gaussian noise (fGn) null hypothesis used for generating the target data, or the autoregressive model of order one (~~AR(1))~~ AR1) null hypothesis which is the assumed ~~temperature model for this~~ temporal evolution model for the reconstruction technique. The study reveals that the resulting field and spatial mean reconstructions are consistent with the fGn hypothesis for ~~most of the parameter configurations~~some of the tested parameter configurations, while others are in better agreement with the AR(1) model. There are local differences in reconstruction skill and reconstructed scaling characteristics between individual grid cells, and ~~a generally better~~ the agreement with the fGn model is generally better for the spatial mean reconstruction than at individual locations. ~~The discrepancy from an fGn is most evident for the high-frequency part of the reconstructed signal, while the long-range memory is better preserved at frequencies corresponding to decadal time scales and longer. Selected experiment setups were found to give reconstructions consistent with the AR(1) model. Reconstruction skill is measured on an ensemble member basis using selected validation metrics. Despite the mismatch between the BARCAST temporal covariance model and the model of the target, the ensemble mean was in general found to be consistent with the target data, while the estimated confidence intervals are more affected by this discrepancy. Our results show~~ Our results demonstrate that the use of target data with a different spatiotemporal covariance

structure than the BARCAST model assumption can lead to a potentially biased CFR reconstruction and associated confidence intervals~~, because of the wrong model assumptions~~.

1 Introduction

Proxy-based climate reconstructions are major tools in understanding past and predicting future variability of the climate system~~over a range of timescales. Over the last few decades a considerable progress has been made, and a number of proxy/multiproxy reconstructions of different climate variables have been created.~~ . Target regions, spatial density and temporal coverage of the proxy network ~~varied~~ vary between the studies, with a general trend towards more comprehensive networks and sophisticated reconstruction techniques used. For example, ~~Jones et al. (1998); Moberg et al. (2005); Mann et al. (1998, 2008); PAGES 2k Consortium (~~ reconstructions of surface air temperatures (SAT) for different spatial and temporal domains. ~~The available reconstructions often~~ On the most detailed level, the available reconstructions tend to disagree on aspects such as specific timing, duration and amplitude of warm/cold periods, due to different methods, types and number of proxies, and regional delimitation used in the different studies, (Wang et al., 2015). There are also alternative viewpoints on a more fundamental basis considering how the level of high frequency versus low frequency variability is best represented, see e.g. Christiansen (2011); Tingley and Li (2012). ~~Discrepancies in the Fourier domain can occur among other things~~ In this context, differences between reconstructions can occur due to shortcomings of the reconstruction techniques, such as regression ~~dilution. This describes~~ causing variance losses back in time and bias of the target variable mean. These artifacts can appear as a consequence of noisy measurements used as predictors in regression techniques based on ordinary least squares (Christiansen, 2011; Wang et al., 2014), though Ordinary Least Squares still provides optimal parameter estimation when the predictor variable has error (Wonnacott and Wonnacott, 1979). The level of high/low frequency variability in reconstructions also depends on the type and quality of the proxy data used as input (Christiansen and Ljungqvist, 2017).

The concept of pseudoproxy experiments was introduced after millennium-long paleoclimate simulations from GCMs first became available, and has been developed and applied over the last ~~decade~~two decades, (Mann et al., 2005, 2007; Lee et al., 2008). Pseudoproxy experiments are used to test the skill of reconstruction methods and the sensitivity to the proxy network used, see Smerdon (2012) for a review. The idea behind idealized pseudoproxy experiments is to extract target data of an environmental variable of interest from long paleoclimate model simulations~~for an arbitrary reconstruction region~~. The target data is then sampled in a spatiotemporal pattern that simulates real proxy networks and instrumental data. The target data representing the proxy period is further perturbed with noise to simulate real proxy data in a systematic manner, while the ~~pseudo instrumental~~ pseudo-instrumental data are left unchanged or only weakly perturbed with noise of magnitude typical for the ~~real world~~ real-world instrumental data. The surrogate pseudoproxy and ~~pseudoinstrumental~~ pseudo-instrumental data are used as input to one or more reconstruction techniques, and the resulting reconstruction is then compared with the true target from the simulation. The reconstruction skill is quantified through statistical metrics, both for a calibration- and a much longer validation interval.

~~The available~~ Available pseudoproxy studies have to a ~~great~~ large extent used target data from the same GCM model simulations, subsets of the same spatially distributed proxy network and a temporally invariant pseudoproxy network (Smerdon, 2012). ~~The concept of extracting target data from simpler model simulations has not been widely explored.~~ In the present paper we extend the domain of pseudoproxy experiments to allow more flexible target data, with a range of explicitly controlled spatiotemporal characteristics. Instead of employing surrogate data from paleoclimate GCM simulations, ensembles of target fields are drawn from a field of stochastic processes with prescribed dependencies in space and time. In the framework of such an experiment design, the idealized temperature field can be thought of as an (unforced) control simulation of the Earth's surface temperature field with a simplified ~~spatial~~ spatiotemporal covariance structure. The primary goal of using these target fields is to test the ability of the reconstruction method to preserve the spatiotemporal covariance structure of the surrogates in the climate field reconstruction. ~~An example of such a study is Werner and Tingley (2015), where idealized target data were generated based on the BARCAST model equations introduced here~~ Earlier, Werner and Tingley (2015) generated stochastic target fields using the AR(1) model equations of BARCAST introduced in Sect. 2.1. We present for the first time a data generation technique for fields of long-range memory (LRM) target data.

~~In addition~~Additionally, we test the reconstruction skill on an ensemble member basis using standard metrics including the correlation coefficient and the root-mean-squared error (RMSE). ~~We also employ the~~ The continuous ranked probability score (CRPS) ~~, which is a suitable skill metric for ensemble-based reconstructions , in contrast to the often-used coefficient of efficiency (CE) and reduction of error (RE)~~is also employed, this is a skill metric composed of two subcomponents recently introduced for ensemble based reconstructions (Gneiting and Raftery, 2007).

Temporal dependence in a stochastic process over time t is described as persistence or memory~~, given that the process has a Gaussian probability distribution. A long-range memory (LRM)~~ . An LRM stochastic process exhibits an autocorrelation function (ACF) and a power spectral density (PSD) of a power-law form: $\mathrm{C}(t) \sim t^{\beta-1}$, and $\mathrm{S}(f) \sim f^{-\beta}$ respectively. The power-law behavior of the ACF and the PSD indicates the absence of a characteristic time scale in the time series; the record is *scale invariant* (or just *scaling*). The spectral exponent β determines the strength of the persistence. The special case $\beta = 0$ is the white noise process, which has a uniform PSD over the range of frequencies. For comparison, another model often used to describe the background variability of the Earth's SAT is the autoregressive process of order 1 (~~AR(1))~~ AR1) (Hasselmann, 1976). This process has a Lorentzian power spectrum (steep slope at high frequencies, constant at low frequencies) and thereby does not exhibit long-range correlations.

For the instrumental time period, studies have shown that detrended local and spatially averaged surface temperature data exhibit long-range memory properties on time scales from months up to decades, (Koscielny-Bunde et al., 1996; Rybski et al., 2006; Fredriksen and Rypdal, 2016). For proxy/multiproxy SAT reconstructions, studies indicate persistence up to a few centuries or millennia, ~~(Rybski et al., 2006; Lovejoy and Schertzer, 2012; Nilsen et al., 2016)~~(Rybski et al., 2006; Huybers and Curry, 2006; N
The exact strength of persistence varies between data sets and depends on the degree of spatial averaging, but in general $0 < \beta < 1.3$ is adequate. The value of $\beta > 1$ is usually associated with sea surface temperature, which features stronger persistence due to effects of oceanic heat capacity ~~. The deviation from Gaussianity of instrumental temperatures varies with latitude~~

~~(Franzke et al., 2012), and the nonlinearity in some types of proxy records also result in nongaussianity (Emile-Geay and Tingley, 2016).~~ (Fraedrich and Blender, 2003; Fredriksen and Rypdal, 2016).

Our basic assumption is that the background temporal evolution of Earth's surface air temperature can be modelled by the persistent Gaussian stochastic model known as the fractional Gaussian noise (fGn) (Beran et al., 2013)[Chapter 1 and 2], (Rypdal et al., 2013). This process is stationary, and the persistence is defined by the spectral exponent ~~$0<\beta<1.$~~ $0 < \beta < 1$. The synthetic target data are designed as ensembles of ~~LRM-processes~~ fGn-processes in time, with an exponentially decaying spatial covariance structure. In contrast to using target data from GCM simulations, this gives us the opportunity to vary the strength of persistence in the target data, retaining a simplistic and temporally persistent model for the signal covariance structure. The persistence is varied systematically to mimic the range observed in actual ~~reconstructions~~ observations over land, typically $0 < \beta < 1$ (Franke et al., 2013; Fredriksen and Rypdal, 2016; Nilsen et al., 2016). The pseudoproxy data quality is also varied by adding levels of white noise corresponding to signal-to-noise ratios by standard deviation (SNR)$= \infty, 3, 1, 0.3$. For comparison, the signal to noise ratio of observed proxy data is normally between 0.5-0.25 (Smerdon, 2012). However, in ~~?~~Werner et al. (2018), most tree-ring series were found to have SNR > 1.

The fGn model is appropriate for many observations of SAT data, but we acknowledge that there are also some deviations. The theoretical fGn follow a Gaussian distribution, but for instrumental SAT data the deviation from Gaussianity varies with latitude (Franzke et al., 2012). Some temperature-sensitive proxy types are also characterized by nonlinearities and non-gaussianity (Emile-Geay and Tingley, 2016).

Since the target data are represented as an ensemble of independent members generated from the same stochastic process, there is little value in estimating and analyzing ensemble means from the target and reconstructed time series themselves. Anomalies across the ensemble members will average out, and the ensemble mean will simply be a time series with non-representative variability across scales. Instead we will focus on averages in the spectral sense. The ~~means~~ median of the ensemble member-based metrics are used to quantify the reconstruction skill.

The reconstruction method to be tested is the "Bayesian Algorithm for Reconstructing Climate Anomalies in Space and Time" (BARCAST), based on a Bayesian Hierarchical Model (Tingley and Huybers, 2010a). This is a state-of-the-art paleoclimate reconstruction technique, described in further detail in Sect. 2.1. The motivation for using this particular reconstruction technique in the present ~~pseudoproxy~~ study is the contrasting background assumptions for the temporal covariance structure. BARCAST assumes that the temperature evolution follows an AR(1) process in time, while the target data are generated according to the fGn model. The consequences of using an incorrect null hypothesis for the temporal data structure are illustrated in Fig. 1. Here, the original time series in Fig. 1a follows an fGn structure. The corresponding ~~power spectrum~~ 95% confidence range of power spectra is plotted in blue in Fig. 1c. Using the incorrect null hypothesis that the data are generated from an AR(1) model, we estimate the AR(1) parameters from the time series in Fig. 1a using Maximum Likelihood estimation. A realization of an AR(1) process with these parameters is plotted in Fig. 1b, with the ~~power spectrum~~ 95% confidence range of power spectra shown in red in Fig. 1c. The characteristic timescale indicating the memory limit of the system is evident as a break in the red AR(1) spectrum. This is an artifact that does not stem from the original data, but simply occurs because an incorrect assumption was used for the temporal covariance structure.

A particular advantage of BARCAST as a probabilistic reconstruction technique lies in its capability to provide an objective error estimate as the result of generating a distribution of solutions for each set of initial conditions. The reconstruction skill of the method has been tested earlier and compared against a few other CFR techniques using pseudoproxy experiments. Tingley and Huybers (2010b) use instrumental temperature data for North America, and construct pseudoproxy data from some of the longest time series. BARCAST is then compared to the RegEM method used by Mann et al. (2008, 2009). The findings are that BARCAST is more skillful than RegEM if the assumptions for the method are not strongly violated. The uncertainty bands are also narrower. ~~The other~~ Another pseudoproxy study is described in Werner et al. (2013), where BARCAST is compared against ~~a CFR method based on~~ the canonical correlation analysis (CCA) CFR method. The pseudo proxies in that paper were constructed from a millennium-long forced run of the NCAR CCSM4 model. The results showed that BARCAST outperformed the CCA method over the entire reconstruction domain, being similar in areas with good data coverage. There is an additional pseudoproxy study by Gómez-Navarro et al. (2015), targeting precipitation which has a more complex spatial covariance structure than SAT anomalies. In that study, BARCAST was not found to outperform the other methods.

In the following, we describe the methodology of BARCAST and the target data generation in Sect. 2. The spectral estimator used for persistence analyses is also introduced here. Sect. 3 ~~comprises~~ is comprised of an overview of the experiment setup and explains the hypothesis testing procedure. Results are presented in Sect. 4 after performing hypothesis testing in the spectral domain of persistence properties in the local and spatial mean reconstructions. The skill metric results are also summarized. Finally, Sect. 5 ~~discuss~~ discusses the implications of our results and provides concluding remarks.

2 Data and methods

2.1 BARCAST methodology

BARCAST is a climate field reconstruction method, described in detail in Tingley and Huybers (2010a). It is based on a Bayesian hierarchical model with three levels. The true temperature field in BARCAST, $\mathbf{T}_t$ is modelled as a multivariate first-order autoregressive model (AR(1)) in time. Model equations are defined at the process level:

$$\mathbf{T}_t - \mu\mathbf{1} = \alpha(\mathbf{T}_{t-1} - \mu\mathbf{1}) + \boldsymbol{\epsilon}_t \tag{1}$$

Where the scalar parameter μ is the mean of the process, α is the AR(1) coefficient, and $\mathbf{1}$ is a vector of ones. The subscript t indexes time in years, and the innovations (increments) $\boldsymbol{\epsilon}_t$ are assumed to be IID normal draws $\boldsymbol{\epsilon}_t \sim N(0, \boldsymbol{\Sigma})$, where

$$\Sigma_{ij} = \sigma^2 \exp(-\phi|\mathbf{x}_i - \mathbf{x}_j|) \tag{2}$$

is the spatial covariance matrix depicting the covariance between locations $\mathbf{x}_i$ and $\mathbf{x}_j$.

The spatial e-folding distance is $1/\phi$ and is chosen to be $\sim$ 1000 km for the target data. This is a conservative estimate resulting in weak spatial correlations for the variability across a continental landmass. (North et al., 2011) estimate that the decorrelation length for a 1-year average of Siberian temperature station data is 3000 km. On the other hand, Tingley and

Huybers (2010a) estimate a decorrelation length of 1800 km for annually mean global land data. They further use annual mean instrumental and proxy data from the North American continent to reconstruct SAT back to 1850, and find a spatial correlation length scale of approximately 3300 km for this BARCAST reconstruction. Werner et al. (2013) use $1/\phi \sim$1000 km as the mean for the lognormal prior in the BARCAST pseudoproxy reconstruction for Europe, but the reconstruction has correlation lengths between 6000-7000 km. The reconstruction of ~~?~~ Werner et al. (2018) has spatial correlation length slightly longer than 1000 km.

On the data level, the observation equations for the instrumental and proxy data are:

$$\mathbf{W}_t = \begin{pmatrix} \mathbf{H}_{I,t} \\ \beta_1 \cdot \mathbf{H}_{P,t} \end{pmatrix} \mathbf{T}_t + \begin{pmatrix} \mathbf{e}_{I,t} \\ \mathbf{e}_{P,t} + \beta_0 \mathbf{1} \end{pmatrix} \tag{3}$$

Where $\mathbf{e}_{I,t}$ and $\mathbf{e}_{P,t}$ are multivariate normal draws $\sim N(0, \tau_I^2 \mathbf{I})$ and $\sim N(0, \tau_P^2 \mathbf{I})$. $\mathbf{H}_{I,t}$ and $\mathbf{H}_{P,t}$ are selection matrices of ones and zeros which at each year select the locations where there are instrumental/proxy data. β_0 and β_1 are parameters representing the ~~scaling factor and bias~~ bias and scaling factor of the proxy records relative to the temperatures. Note that these two parameters have no relation to the spectral parameter β. The BARCAST parameters are distinguished by their indices, the notation is kept as it is to comply with existing literature.

The remaining level is the prior. Weakly informative but proper prior distributions are specified for the scalar parameters and the temperature field for the first year in the analysis. The priors for all parameters except ϕ are conditionally conjugate, meaning the prior and the posterior distribution has the same parametric form. The Markov-Chain Monte Carlo (MCMC) algorithm known as the Gibbs sampler (with one Metropolis step) is used for the posterior simulation (Gelman et al., 2003). Table C1 sums up the prior distributions and the choice of hyperparameters for the scalar parameters in BARCAST. The CFR version applied here has been updated as described in Werner and Tingley (2015). The updated version allows inclusion of proxy records with age uncertainties. This property will not be used here directly, but it implies that proxies of different types may be included. Instead of estimating one single parameter value of τ_P^2, β_0 and β_1, the updated version estimates individual values of the parameters for each proxy record ~~(?)~~(Werner et al., 2018).

~~In the present study, the~~ The Metropolis-coupled MCMC algorithm is run for 5000 iterations, running three chains in parallel. Each chain is assumed equally representative for the temperature reconstruction if the parameters converge. There are a number of ways to investigate convergence, for instance one can study the variability in the plots of draws of the model parameters as a function of step number of the sampler, as in Werner et al. (2013). However, a more robust convergence measure can be achieved when generating more than one chain in parallel. By comparing the within-chain variance to the between-chain variance we get the convergence measure $\hat{R}$, (Gelman et al., 2003, Chapter 11). $\hat{R}$ close to one indicates convergence for the scalar parameters.

There are numerous reasons why the parameters may fail to converge, including inadequate choice of prior distribution and/or hyperparameters or using an insufficient number of iterations in the MCMC algorithm. It may also be problematic if the spatiotemporal covariance structure of the observations or surrogate data deviate strongly from the model assumption of BARCAST.

~~Since BARCAST is a probabilistic reconstruction technique it~~ BARCAST was used to generate an ensemble of reconstructions, in order to achieve a mean reconstruction as well as uncertainties. In our case, the draws for each temperature field and parameter are thinned so that only every 10 of the 5000 iterations are saved; this secures independence of the draws.

The output temperature field is reconstructed also in grid cells without observations, ~~which~~ this is a unique property compared to other well-known field reconstruction methods such as the regularized expectation maximum technique (RegEM) applied in Mann et al. (2009). Note that the assumptions for BARCAST should generally be different for land and oceanic regions, due to the differences in characteristic timescales and spatiotemporal processes. BARCAST is so far only configured to ~~deal with~~ handle continental land data, (Tingley and Huybers, 2010a).

2.2 Target data generation

While generating ensembles of synthetic LRM processes in time is straightforward using statistical software packages, it is more complicated to generate a field of persistent processes with prescribed spatial covariance. Below we describe a novel technique that fulfills this goal, which can be extended to include more complicated spatial covariance structures. Such a spatiotemporal field of stochastic processes ~~may potentially have many~~ has many potential applications, both theoretical and practical~~applications~~.

Generation of target data begins with reformulating ~~eq. (1)~~ Eq. 1 so that the temperature evolution is defined from a power-law function instead of an AR(1). The continuous-time version of Eq. ~~(1)~~ (1 (with $\mu = 0$) is the ~~ordinary~~ stochastic (ordinary) differential equation:

$$\frac{d\mathbf{T}}{dt} = -(1-\alpha)d\mathbf{I}\mathbf{T}_t = -\lambda\mathbf{T}dt + d\epsilon\mathbf{W}_t, \quad \text{Where } \mathbf{I} \text{ Is the identity matrix.} \tag{4}$$

~~with the solution:~~

$$\mathbf{T}(t) = \int_0^t \exp^{-(1-\alpha)\mathbf{I}(t-s)} \boldsymbol{\epsilon}_s \mathrm{ds}$$

~~The exponential kernel is then replaced~~ where $\mathbf{T}_t = (T_{t,1}, \ldots, T_{t,n})$, and $T_{t,i}$ is the temperature at time t and spatial position $\mathbf{x}_i$. The noise term $d\mathbf{W}_t = (dW_{t,1}, \ldots, dW_{t,n})$ is a vector of (dependent) white-noise measures. Spatial dependence is given

by Eq. 2 when $\boldsymbol{\epsilon}_t = \mathbf{W}_{t+\Delta t} - \mathbf{W}_t$ and $\Delta t = 1$ yr. If $\mathbf{I}$ denotes the identity matrix, the stationary solution of Eq. 4 is

$$\mathbf{T}_t = \int_{-\infty}^{t} \exp\Big(-\lambda \mathbf{I}(t-s)\Big) d\mathbf{W}_s, \tag{5}$$

which defines a set of dependent Ornstein-Uhlenbeck processes (the continuous-time versions of AR(1) processes with $\alpha = e^{-\lambda}$, $\alpha = 1-\lambda$). Eq. 4 assumes that the system is characterized by a single eigenvalue λ, and consequently that there is only one characteristic time scale $1/\lambda$. It is well-known that surface temperature exhibits variability on a range of characteristic time scales, and more realistic models can be obtained by generalizing the response kernel as a weighted sum of exponential functions (Fredriksen and Rypdal, 2017):

$$\mathbf{T}_t = \int_{-\infty}^{t} \left[\sum_k c_k \exp\Big(-\lambda_k \mathbf{I}(t-s)\Big)\right] d\mathbf{W}_s. \tag{6}$$

An emergent property of the climate system is that the temporal variability is approximately scale invariant (Rypdal and Rypdal, 2014, 2016 the multi-scale response kernel in Eq. 6 can be approximated by a power-law function to yield:

$$\mathbf{T}(t)_t = \int_{0\,-\infty}^{\;t} (t-s)^{\beta/2-1} d\epsilon \mathbf{W}_s \mathrm{d}s_s. \tag{7}$$

This expression describes the long-memory response to the noise forcing~~after time $t = 0$. Note that there is no contribution from the initial condition $\mathbf{T}(0)$. This is because $\mathbf{T}(t)$~~. We note that this should be considered as a formal expression since the stochastic integral is divergent due to the singularity at $t = s$. Also note that $\mathbf{T}_t$ in Eq. ~~(8)~~ 7 in contrast to Eq. ~~(5)~~ 5 is no longer a solution to an ordinary differential equation, but ~~rather~~ to a fractional differential equation~~, whose solution for $t > 0$ depends not only on the initial condition but the entire time history of $\mathbf{T}(t)$, $t \in (-\infty, 0)$. Eq. 8 effectively corresponds to~~. By neglecting the contribution from the noisy forcing prior to $t = 0$ ~~.~~

~~In discrete form , the convolution integral in Eq. 8 is approximated over an index s:~~ we obtain

$$\mathbf{T}_t = \int_0^t (t-s)^{\beta/2-1} d\mathbf{W}_s, \tag{8}$$

which in discrete form can be approximated by

$$\mathbf{T}_t = \sum_{s=0}^{t} (t-s+\tau_0)^{\beta/2-1} \epsilon_s \boldsymbol{\epsilon}_s. \tag{9}$$

~~Note that here the~~ The stabilizing term τ_0 is added to avoid the singularity at $s = t$. The optimal choice would be to choose τ_0 such that the term in the sum arising from $s = t$ represents the integral ~~in~~ over the interval $s \in (t-1, t)$, i.e.,

$$\tau_0 = \int_0^{\tau_0} \tau^{\beta/2-1} d\tau,$$

which has the solution $\tau_0 = \beta/2$.

~~Summation for time steps $s, t = 1, 2, ...N$ and $\tau = t - s$~~ Summations over time steps $s = 1, 2, ...N$ of (9) results in the matrix ~~G with terms:~~

$$\mathbf{G}(\tau) = (\tau + \beta/2)^{(\beta/2-1)}\Theta(\tau)$$

~~Where $\Theta(\tau)$ is the~~ product:

$$T_{t,i} = \sum_{s=1}^{N} G_{ts}\epsilon_{s,i},$$

where $\mathbf{G} = (G_{ts})$ is the $N \times N$ matrix

$$G_{t,s} = (t - s + \beta/2)^{(\beta/2-1)}\Theta(t - s), \tag{10}$$

and $\Theta(t)$ is the unit step function. ~~ϵ_t is kept identical as in eq. (1) and (2). The~~

If we for convenience omit the spatial index i from Eq. 10, the model for the target temperature field $\mathbf{T}$ at time t can be ~~calculated as:~~ written in the compressed form

$$\mathbf{T}_t = \mathbf{G}_t\boldsymbol{\epsilon}_t. \tag{11}$$

$$\mathbf{T}_t = \mathbf{G}_t\boldsymbol{\epsilon}_t$$

2.3 ~~Estimation of power-spectral density~~Scaling analysis in the spectral domain

The temporal dependencies in the reconstructions are investigated to obtain detailed information about how the reconstruction technique may alter the level of variability on different scales, and how sensitive it is to the proxy data quality. Persistence properties of target data, pseudoproxies and the ~~reconstruction~~ reconstructions are compared and analyzed in the spectral domain using the periodogram as the estimator. See appendix A for details on how the periodogram is estimated.

~~The periodogram is defined here in terms of the discrete Fourier transform H_m as $S(f_m) = (2/N)|H_m|^2, m = 1, 2, \ldots, N/2$. The sampling time is an arbitrary time unit, and the frequency is measured in cycles per time unit: $f_m = m/N$. $\Delta f = 1/N$ is the frequency resolution and the smallest frequency which can be represented in the spectrum.~~

Power spectra are visualized in log-log plots since the spectral exponent then can be estimated by a simple linear fit to the spectrum. The raw and log-binned periodograms are plotted, and β is estimated from the latter. Log-binning of the periodogram is used here for analytical purposes, since it is useful with a representation where all frequencies are weighted equally with respect to their contributions to the total variance.

It is also possible to use other estimators for scaling analysis, such as the detrended fluctuation analysis (DFA, Peng et al. (1994)), or wavelet variance analysis (Malamud and Turcotte, 1999). ~~Each estimation technique has benefits and deficiencies,~~

and one One can argue for the superiority of methods other than PSD or the use of a multi-method approach. However, we consider the spectral analysis to be adequate for our purpose and refer to Nilsen et al. (2016) for a discussion Rypdal et al. (2013); Nilsen et al. (discussions on selected estimators for scaling analysis.

3 Experiment setup

The experiment domain configuration is selected to resemble that of the continental landmass of Europe, with $N = 56$ grid cells of size 5°x5°5° x 5°. The reconstruction period is 1000 years,. reflecting the last millennium. The reconstruction region and period are inspired by the BARCAST reconstructions in Werner et al. (2013); Luterbacher et al. (2016) and approximate the density of instrumental and proxy data in reconstructions of the European climate of the last millennium. The temporal resolution for all types of data is annual. By construction the target fGn data are meant to be an analogue of the unforced SAT fieldand hence can be considered as representing GCM control simulations. We will study both the field and spatial mean reconstruction.

Pseudoinstrumental data cover the entire reconstruction region for the time period 850-1000 and are identical to the noise-free values of the true target variables. The spatial distribution of the pseudoproxy network is highly idealized as illustrated in Fig. 2, the data covers every fourth grid cell for the time period 1-1000. The pseudoproxies are constructed by perturbing the target data with white noise according to Eq. 3. The variance of the proxy observations is τ_P^2, and the SNR is calculated as:

$$\mathrm{SNR} = \frac{\beta_1^2 \mathrm{Var}(T_t)}{\tau_P^2} \tag{12}$$

Our set of experiments is summarized in Table 1 and comprises target data with three different strengths of persistence, $\beta = 0.55, 0.75, 0.95$ and pseudoproxies with four different signal to noise ratios by standard deviation (SNR): SNR=∞, 3, 1; SNR $= \infty, 3, 1$, and 0.3. In total, 20 realizations of target pseudoproxy and pseudoinstrumental data are generated for each combination of β and SNR and used as input to BARCAST. The reconstruction method is probabilistic and generates ensembles of reconstructions for each input data realization. In total, 30 000 ensemble members are constructed for every parameter setup.

3.1 Hypothesis testing

Hypothesis testing in the spectral domain is used to determine which pseudoproxy/reconstructed data sets can be classified as fGn with the prescribed scaling parameter, or as AR(1) with parameter α estimated from BARCAST. The power spectrum for each ensemble member of the local/spatial mean reconstructions is estimated, and the mean power spectrum is then used for further analyses. The first null hypothesis is that the data sets under study can be described using an fGn with the prescribed scaling parameter for the target data at all frequencies, $\beta_{\mathrm{target}} = 0.55, 0.75$ and 0.95 respectively. For testing we generate a Monte Carlo (MC) ensemble of fGn series with a value of the scaling parameter identical to the target data. The power spectrum of each ensemble member is estimated, and the confidence range for the theoretical spectrum is then calculated using the 2.5

and 97.5 quantiles of the log-binned periodograms of the ~~MC~~ Monte Carlo ensemble. The null hypothesis is rejected if the log-binned mean spectrum of the data is outside of the confidence range for the fGn model at any point.

The second null hypothesis tested is that the data can be described as an AR(1) process at all frequencies, with the parameter α estimated from BARCAST. Distributions for all scalar parameters including the AR(1) parameter α are provided through the reconstruction algorithm. The mean of this parameter was used to generate a Monte Carlo ensemble of AR(1) processes. The ~~MC~~ Monte Carlo ensemble and the confidence range is then based on log-binned periodograms for this theoretical AR(1) process.

Figure 3 presents an example of the hypothesis testing procedure. The fGn 95% confidence range is plotted as a shaded gray area in the log-log plot together with the mean raw and mean log-binned periodograms for the data to be tested. Blue curve and dots represent mean raw and log-binned PSD for pseudoproxy data, red curve and dots represent mean raw and log-binned PSD for reconstructed data. The gray, dotted line is the ensemble mean.

~~Note that the formulation of the~~ The two null hypotheses ~~gives~~ give no restriction about the normalization of the fGn and AR(1) data used to generate the ~~MC ensemble~~Monte Carlo ensembles. Particularly, they do not have to be standardized in the same manner as the pseudoproxy/reconstructed data. This makes the ~~experiment~~ experiments more flexible, as the ~~spectral~~ confidence range of the ~~MC~~ Monte Carlo ensemble can be shifted vertically to better accommodate the data under study. A standard normalization of data includes subtracting the mean and normalizing by the standard deviation. This was sufficient to support the null hypotheses in many of our experiments. A different normalization ~~could also be used, for instance if one considers only the high- or only the low frequency variability to be representative of the true variability in the time series. This is often the case with proxy-based reconstructions. In this case, it would be useful to calculate the variance in the high- or low frequency range by filtering the data, and then normalize the unfiltered data by this variance.~~ had to be used in other experiments.

4 Results

BARCAST successfully estimates posterior distributions for all reconstructed temperature fields and scalar parameters. Convergence is reached for the scalar parameters despite the inconsistency of the input data temporal covariance structure with the default assumption of BARCAST. Table C2 lists the true parameter values used for the target data generation, and ~~Tab.~~ Table C3 summarizes the mean of the posterior distributions estimated from BARCAST. Studying the parameter dependencies, it is clear that the posterior distributions of α and σ^2 depend on the prescribed β and to a lesser extent, SNR for the target data. ~~The mean values of the α distributions were used to generate Monte Carlo ensembles of AR(1) processes for hypothesis testing. For the parameters τ_P^2, β_0 and β_1, there are individual posterior distributions for each of the local proxy records.~~ Instead of listing the posterior distributions of τ_P^2 and β_1 we have estimated the local reconstructed SNR at each proxy location using Eq. 12.

Further results concern the spectral analyses and skill metrics. ~~For each ensemble member of the input dataset and temperature reconstruction, the PSD is estimated and the mean spectrum is used in further analyses.~~ All references to spectra in the fol-

lowing correspond to mean spectra. Analyses of the reconstruction ~~skills~~ skill presented below are performed on a grid point basis~~as well as for the~~ , and for the correlation and RMSE also for the spatial mean reconstruction. While the latter provides an aggregate summary of the method's ability to reproduce specified properties of the climate process on a global scale, the former evaluates ~~the BARCAST~~BARCAST's spatial performance.

4.1 Isolated effects of added proxy noise on scaling properties in the input data

The scaling properties of the input data are modified already when the target data are perturbed with white noise to generate pseudoproxies. The power spectra shown in blue in Fig. 3 are used to illustrate these effects for one arbitrary proxy location and ~~β~~$\beta = 0.75$. Figure 3a shows the spectrum for SNR$= \infty$, which is the unperturbed fGn signal corresponding to ideal proxies. Panels 3~~b, c and d~~ b-d show spectra for SNR=3, 1 and 0.3 respectively. The effect of added white noise in the spectral domain is manifested as flattening of the-high frequency part of the spectrum equal to $\beta = 0$, and a gradual transition to higher β for lower frequencies. The ~~pseudoproxies in panels 3b, c and d all deviate from the confidence range on the highest frequencies, while the log-binned spectrum in Fig. 3(d) is outside on lower frequencies as well. The hypothesis testing~~ results for $\beta_{\text{target}} = 0.55$ and 0.95 are ~~the same~~similar (figures not shown).

4.2 Memory properties in the field reconstruction

Hypothesis testing was performed in the spectral domain for the field reconstructions, with the two null hypotheses formulated as follows:

1: The reconstruction is consistent with the fGn structure in the target data for all frequencies.
2: The reconstruction is consistent with the AR(1) model used in BARCAST for all frequencies.

Table 2 summarizes the results for all experiment configurations at local grid cells, both directly at and between proxy ~~grid cells. Figures are provided only for one β, all SNR.~~ locations. Figure 3 shows the mean power spectra generated for ~~the experiment $\beta = 0.75$ at~~ one arbitrary proxy grid cell of the reconstruction in red. The fGn model is adequate for SNR=∞, 3 and 1, shown in panel 3a-c. For the lowest SNR presented in panel d, the reconstruction spectrum falls outside the confidence range of the theoretical spectrum for one single log-binned point. Not unexpectedly, the difference in shape of the PSD between the pseudoproxy and reconstructed spectra increases with decreasing SNR. The difference is largest for the noisiest proxies with SNR=0.3. This figure does not show the hypothesis testing for the reconstructed spectrum using the AR(1) null hypothesis. Results show that this null hypothesis is rejected for all cases except SNR=0.3.

The hypothesis testing results vary moderately between the individual grid cells. PSD analyses of the local reconstructions using the same β but in an arbitrary non-proxy location are displayed in Fig. 4. Here, the reconstructed mean spectrum is plotted in gray together with both the fGn 95% confidence range (blue) and the AR(1) confidence range (red). Hypothesis testing using null hypothesis 1 and 2 is performed systematically. Wherever the reconstructed log-binned spectrum is consistent with the fGn/AR(1) model, the edges of the associated confidence range are plotted with solid lines. We find that all reconstructed

log-binned spectra are consistent with the AR(1) model, ~~while only the cases~~ and for SNR=∞ and 3 ~~are~~ the reconstructions are also consistent with the fGn model.

4.3 Memory properties in the spatial mean reconstruction

The spatial mean reconstruction is calculated as the mean of the local reconstructions for all grid cells considered, weighted by the areas of the grid cells. The reconstruction region considered is $37.5° - 67.5°$N, $12.5° - 47.5°$E~~.~~, as shown in Fig. 2. Figure 5 shows the raw and log-binned periodogram of the spatial mean reconstruction for ~~$\beta_{target} = 0.75$~~ $\beta_{\text{target}} = 0.75$ in gray, together with the 95% confidence range of fGn generated with $\beta = 0.75$ (blue) and AR(1) confidence range (~~pink~~red). All hypothesis testing results for the spatial mean reconstruction are summarized in Table 3. Results show that the fGn null hypothesis is suitable for all values of β and SNR, while the AR(1) null hypothesis is also supported for the case $\beta = 0.55, 0.75$, SNR=0.3.

4.4 Effects from BARCAST on the reconstructed signal variance

The power spectra can also be used to gain information about the fraction of variance lost/gained in the reconstruction compared with the target. This fraction is in some sense the bias of the variance, and was found by integrating the spectra of the input and output data over frequency. The spatial mean target/reconstructions were used, and the mean log-binned spectra. The total power in the spatial mean reconstruction and the target were estimated, and the ratio of the two provides the under/overestimation of the variance: $\text{R}_{\text{Var}} = \frac{\text{Var}(T_{t\,(\text{rec})})}{\text{Var}(T_{t\,(\text{target})})}$. A ratio less than unity implies that the reconstructed variance is underestimated compared with the target. Our analyses for the total variance reveal that the ratio varies between 0.83-1.05 for the different experiments and typically decreases for increasing noise levels. How much the ratio decreases with SNR depends on β, with higher ratios for higher β values. For example, R~ 1 for all β, SNR$= \infty$ and progressively decreases to R=0.83, 0.89 and 0.94 for SNR=0.3, $\beta = 0.55, 0.75, 0.95$ respectively. In other terms, there are larger variance losses in the reconstruction for smaller values of β than for higher β. We also divided the spectra into three different frequency ranges as shown in Fig. 6 to test if the fraction of variance lost/gained is frequency-dependent. The sections separate low frequencies corresponding approximately to centennial timescales, mid frequencies corresponding to timescales between decades and centuries, and high frequencies corresponding to timescales shorter than decadal. The results show no systematic differences between the frequency ranges associated with the parameter configuration.

4.5 Assessment of reconstruction skill

It is common practice in paleoclimatology to evaluate reconstruction skill using metrics such as the Pearson's correlation coefficient r, the root-mean squared error (RMSE), ~~and~~ the coefficient of efficiency (CE) ~~(Smerdon et al., 2011; Wang et al., 2014). However,~~ and reduction of error (RE) (Smerdon et al., 2011; Wang et al., 2014). The two former skill metrics will be used in this study, but the CE and RE metrics are not proper scoring rules and are therefore unsuitable for ensemble-based reconstructions in general (Gneiting and Raftery, 2007; Tipton et al., 2015), see also Appendix B. Instead, the ~~CE metric is improper for~~

reconstructions based on the Bayesian framework (?), and will not be used. Insteadwe will use the continuous ranked probability score (CRPS) , (Gneiting and Raftery, 2007), which was earlier used in (Werner and Tingley, 2015; ?). Skill values are estimated on an ensemble member basis, but results given below are mean values for the entire ensemble.

The skill of the reconstruction method is measured using the RMSE, the Pearson's correlation coefficient (r) and the CRPS. will be used (Gneiting and Raftery, 2007). The metric was used specifically for probabilistic climate reconstructions in Tipton et al. (2015); Werner and Tingley (2015); Werner et al. (2018). Since the CRPSis , its average and subcomponents are less well-known than the two former methodsskill metrics, we define it briefly these terms in Appendix B and refer to Gneiting and Raftery (2007) Gneiting and Raftery (2007); Hersbach (2000) for further details.

$$\mathrm{CRPS} = \frac{1}{2}\mathbb{E}_F|X - X'| - \mathbb{E}_F|X - x|$$

Where F refer to the cumulative distribution function. X and X' are ensemble members of the reconstruction, and x is the target variable. The first term represents the mismatch between the cumulative distribution functions of all pairs of ensemble member reconstructions, while the second term measures the mismatch between all ensemble members and the target cumulative distribution function. The CRPS score is given for each individual time step. The estimates are given in the same unit as the variable under study, here surface temperature. In the following we are handling the two subcomponents of the CRPS: the temporally averaged score metric, called the average potential CRPS, (CRPS results below are shown for the $\overline{\mathrm{CRPS}}$ represented via the sum of the subcomponents $\overline{\mathrm{CRPS}}_{\mathrm{pot}}$) and the Reliability, representing the validity of the uncertainty bands (Hersbach, 2000). The $\overline{\mathrm{CRPS}}_{\mathrm{pot}}$ metric is akin to the Mean Absolute Error of a deterministic forecast. The Reliability metric tests whether for all cases in which a certain probability p was forecast, on average, the event occurred with that fraction p. Or, using other words, it is tested whether the ensemble is capable of generating cumulative distributions that have, on average, the desired statistical property. Note that a perfectly reliable system has Reliability=0. and $\overline{\mathrm{Reli}}$ score.

4.5.1 Skill measure results

The figures 7, 8, 9 and ?? Figures 7-9 display the spatial distribution of the ensemble mean skill metrics for the experiment β=0.75 and all noise levels. All figures show a spatial pattern of dependence on the proxy availability, with the best skill attained at proxy sitesexcept in Fig. ??. . This is the most important result for all the spatially distributed skill metrics. For the BARCAST CFR method, the signal at locations distant from proxy information by design cannot be skillfully reconstructed, as the amount of shared information on the target climate field between the two locations decreases exponentially with distance (Werner et al., 2013). See Sect.5 for further details on the relation between skill and co-localization of proxy data.

Figure 7 shows the local correlation coefficient r between the target and the localized reconstruction for the verification period 1-18492-1849. The correlation is highest for the ideal-proxy experiment in Fig. 7a, and gradually decreases at all locations as the noise level rises in panels b-d. Fig. 8 shows the local RMSE. Note that Fig. 8-9 use the same color bar as in Fig. 7, but best skill is achieved where the RMSE/$\overline{\mathrm{CRPS}}_{\mathrm{pot}}$ is low. Fig. 9 shows the distribution of $\overline{\mathrm{CRPS}}_{\mathrm{pot}}$. The contribution from $\overline{\mathrm{Reli}}$ is generally $< 1 * 10^{-2}$, indicating excellent correspondence between the predicted and the reconstructed confidence intervals. The $\overline{\mathrm{CRPS}}_{\mathrm{pot}}$ therefore dominates the average CRPS metric. The minimum estimate for the $\overline{\mathrm{CRPS}}_{\mathrm{pot}}$ at proxy

locations in Fig. 9a is ~~0.15, which indicates~~ $2 * 10^{-2}$, indicating a low error between the temporally averaged reconstruction and the target. For the remaining locations in Fig. 9a-d, the estimates are between ~~0.61-0.67~~0.24-0.55 given in the same unit as the target variable. The temperature unit has not been given for our ~~pseudoproxy~~ reconstructions, but for real-world reconstructions the unit will typically be degrees Celsius (°C) or Kelvin (K). ~~The Reliability shown in Fig. ?? is generally low if the proxy locations in ??a are neglected. Except for these grid cells, the local Reliability score ranges between $1 * 10^{-3}$ to 0.32, with lowest (best skill) estimate for the lowest SNR (strongest noise). The improved Reliability for higher noise scenarios is apparently due to a better consistency between the BARCAST model assumption and the LRM signal which is deteriorated with a high additive noise level. The maximum Reliability score at the proxy locations in ??a is 0.93, which indicates poor reconstruction skill when the validity bands are considered. For these locations, the contrasting skill scores obtained for the $\overline{\mathrm{CRPS}}_{\mathrm{pot}}$ and the Reliability indicate that the reconstructions are on average in good agreement with the target, but the confidence range is not.~~

Table 4 summarizes the ~~mean~~ median local skill for all experiments and skill metrics. BARCAST is in general able to reconstruct major features of the target field. A general conclusion that can be drawn is that the skill metrics vary with SNR, but are less sensitive to the value of β. For the highest noise-level SNR=0.3, the values obtained for r and the RMSE are in line with those listed in Table 1 of Werner et al. (2013).

Table 5 sums up the ensemble ~~mean~~ median skill values of r and RMSE for the spatial mean reconstructions. The skill ~~values are considerably higher~~ is considerably better than for the local field reconstructions. ~~The CRPS scores have not been evaluated for the spatial mean reconstruction.~~

5 Discussion

In this study we have tested the capability of BARCAST to preserve temporal ~~long-range memory~~ LRM properties of reconstructed data. Pseudoproxy and ~~pseudoinstrumental~~ pseudo-instrumental data were generated with a prescribed spatial covariance structure and LRM temporal persistence using a new method. The data were then used as input to the BARCAST reconstruction algorithm, which by construction ~~uses~~ use an AR(1) model for temporal dependencies in the input/output data. The spatiotemporal availability of observational data was kept the same for all experiments in order to isolate the effect of the added noise level and the strength of persistence in the target data. The mean spectra of the reconstructions ~~are~~ were tested against the null hypotheses that the reconstructed data can be represented as LRM processes using the parameters specified for the target data, or as AR(1) processes using the parameter estimated from BARCAST.

We found that despite the default assumptions in BARCAST, not all local and spatial mean reconstructions were consistent with the AR(1) model. ~~Typically, the~~ Figures 3-5 and Tables 2-3 summarize the hypothesis testing results; the local reconstructions at grid cells between proxy locations ~~are shown to follow~~ follow to large extent the AR(1) model, while the local reconstructions directly at proxy locations are more similar to the original fGn data. However, the ~~parameter setup~~ simulated proxy quality is crucial for the spectral shape of ~~these~~ the local reconstructions, with higher noise levels indicating better agreement with the AR(1) model than ~~the~~ fGn.

~~Moreover, the hypothesis testing results show that all~~ All spatial mean reconstructions are consistent with the fGn null hypothesis according to Table 3. For the two cases $\beta = 0.55, 0.75$, SNR=0.3, the spatial mean reconstructions are also consistent with the AR(1) null hypothesis. This is clear from the spatial mean reconstruction spectra (gray curves in Fig. 5) and from comparing the hypothesis testing results in Table 2 and 3. The improvement in scaling behavior with spatial averaging is expected, as the small-scale variability denoted by ϵ_t in Eq. 11 is averaged out. Eliminating local disturbances naturally results in a more coherent signal. However, the spatial mean of the target data set does not have a significantly higher β than local target values. This is due to the relatively short spatial correlation length chosen: $1/\phi = 1000$ km. In observed temperature data, spatial averaging tends to increase the scaling parameter β (Fredriksen and Rypdal, 2016).

~~The fact that the reconstructed LRM properties are better preserved directly at proxy locations than between proxies is an expected result. By construction, BARCAST estimates the posterior distributions of τ_P^2 and β_1, which are related to the signal to noise ratio through Eq. 12. Tab. C3 shows that the estimated reconstructed SNR is close to the true SNR. The reconstruction is designed to follow the model equations 1 and 2 of the true temperature field. The temperatures between proxies therefore have to be generated through stochastic infilling using all the estimated parameters, while directly at proxy sites the only necessary operation is to remove the assumed proxy noise.~~

~~Due to the interdependence of the BARCAST parameters, the increase in the estimated AR(1) parameter α and β_0 is accompanied by a decrease of σ^2 for noisy input data. These erroneous estimates influence the resulting reconstruction.~~

The power spectra in Fig. 3~~, 4 and 5~~ -5 b-d show that the temporal covariance structure of the reconstructions is altered compared with the target data for all experiments where noisy input data were used. Furthermore, the spectra of the pseudoproxies ~~and the reconstructions~~ in 3b-d all deviate from the target in the high frequency range, but for ~~different reasons~~a different reason. The pseudoproxy data deviate from the target due to the white proxy noise component, while the reconstruction deviate because BARCAST quantifies the proxy noise from an AR(1) assumption. ~~This has important implications for how paleoclimate reconstructions should be interpreted.~~ Real-world proxy data are generally noisy, and the noise level is normally at the high end of the range studied here. We demonstrate that the variability-level of the reconstructions does not exclusively reflect the characteristics of the target data, but is also influenced by the fitting of noisy data to a model that is not necessarily correct. ~~Other reconstruction~~ At present, there exists no reconstruction technique assuming explicitly that the climate variable follows an LRM process.

In addition to BARCAST, other reconstruction techniques that may experience similar deficiencies ~~is~~ for LRM target data are the regularized expectation-maximization algorithm (RegEM), (Schneider, 2001; Mann et al., 2007), and all related models (CCA, PCA~~), that assumes~~ , GraphEM). These models assume observations at subsequent years are independent (Tingley and Huybers, 2010b). 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 methodology on precipitation data in Wahl et al. (2017).

~~Our results further suggest that the spatial mean reconstructions are to a small extent more consistent with the LRM null hypothesis than local values. This is clear from the spatial mean reconstruction spectra (gray curves in Fig. 5) and from~~

~~comparing the hypothesis testing results in Table 2 and 3. The improvement in scaling behavior is expected, as~~ When an incorrect statistical model is used to reconstruct a climate signal, the ~~small-scale variability denoted by ϵ_t in Eq. 11 is averaged out. Eliminating local disturbances naturally results in a more coherent signal. However, the spatial mean of the target data set does not have a significant higher β than local target values. This is due to the relatively short spatial correlation length chosen: $1/\phi = 1000$ km. In observed temperature data, spatial averaging tends to increase the scaling parameter β (Fredriksen and Rypdal, 2016)~~temporal correlation structure is likely to be deteriorated in the process. For the range of different reconstructions available, such effects may contribute to discussions on a number of questions under study, including the possible existence of different scaling regimes in paleoclimate, see Huybers and Curry (2006); Lovejoy and Schertzer (2012); Rypdal and R

~~If the~~ The criteria for the hypothesis testing used in this study are strict, and may be modified if reasonable arguments are provided. For example, if the first null hypothesis used here was modified so that only the low-frequency components of the spectra were required to fall within the confidence ranges, more of the reconstructions would be consistent with the fGn model. ~~After all, we have the information that the high-frequency component is affected by noise, and we know the color and level of this noise.~~ However, from studying the spectra in Fig. 3~~, 4 and~~ -5, it is generally unclear where one should set a threshold, since the spectra show a gradual change with a lack of any abrupt breaks. Considering real-world proxy records, the noise color and level is generally unknown or not quantified. We know there are certain sources of noise ~~that are not related to climate~~ influencing different frequency ranges ~~. However,~~ that are unrelated to climate, but it is difficult to decide when the noise becomes negligible compared with the effects of climate driven processes. The decision to use all frequencies for the hypothesis testing in this idealized study is therefore a conservative and objective choice.

The skill metrics used to validate the reconstruction skill are the ~~RMSE~~Pearson's correlation coefficient r, the RMSE, and $\overline{\text{CRPS}}$, ~~r and CRPS,~~ the latter divided into the $\overline{\text{CRPS}}_{\text{pot}}$ and the ~~Reliability. We stress that even though the estimates of RMSE, correlation and $\overline{\text{CRPS}}_{\text{pot}}$ indicate skillful mean reconstructions, this does not necessarily imply a reliable reconstruction in terms of correct confidence intervals. The Reliability reflects this uncertainty, which is an important measure for probabilistic reconstruction techniques.~~

~~The power spectra can also be used to gain information about the fraction of variance lost~~$\overline{\text{Reli}}$. Skill metric results are illustrated in Fig. 7-9 and summarized in Table 4-5. The reconstruction skill is sensitive to the proxy quality, and highest at sites with co-localized proxy information. This is an expected result, due to the BARCAST model formulation and our choice of a relatively short decorrelation length $1/\phi \sim 1000$ km. Contrasting results of high skill away from proxy sites and poor skill close to proxy sites have been documented in Wahl et al. (2017), although care must be taken for the comparison as that paper used a different reconstruction methodology (truncated EOF-principal component spatial regression) and the target variable was precipitation/~~gained in the reconstruction compared with the target. This fraction is in some sense the bias of the variance, and was found by integrating the spectra of the input and output data over frequency. The spatial mean target/reconstructions were used, and the mean log-binned spectra. The total power in the spatial mean reconstruction and the target were estimated, and the ratio of the two provides the under/overestimation of the variance: $\text{R}_{\text{Var}} = \frac{\text{Var}(T_{t\,(\text{rec})})}{\text{Var}(T_{t\,(\text{target})})}$. A ratio less than unity implies that the reconstructed variance is underestimated compared with the target. Our analyses for the total variance reveal that the ratio varies between 0.83-1.05 for the different experiments and typically decreases for increasing noise levels. How much~~

the ratio decreases with SNR depends on β, with higher ratios for higher β values. For example, R$\sim$ 1 for all β, SNR$=\infty$ and progressively decreases to R=0.83, 0.89 and 0.94 for SNR=0.3, $\beta = 0.55, 0.75, 0.95$ respectively. In other terms, there are larger variance losses in the reconstruction for smaller values of β than for higher β. We also divided the spectra into three different frequency ranges as shown in Fig. 6 to test if the fraction of variance lost/gained is frequency-dependent. The sections separate low frequencies corresponding approximately to centennial timescales, mid frequencies corresponding to timescales between decades and centuries, and high frequencies corresponding to timescales shorter than decadal. The results show no systematic differences between the frequency ranges associated with the parameter configuration. hydroclimate instead of SAT. Without performing dedicated pseudoproxy experiments it is difficult to resolve the main cause of these contrasting results for spatial skill.

Previously, the scaling properties of millennium-long paleoclimate reconstructions have been studied in e.g. Lovejoy and Schertzer (2012 These papers present different viewpoints on scaling models used to represent Earth's surface temperature variability on a range of timescales. Lovejoy and Schertzer (2012) suggests that climate variability on timescales from months to centuries can be denoted "Macroweather" and described using a scaling parameter $\beta \sim 0.2$, while variability on centennial timescales and longer is "Climate" with $\beta \sim 1.4$. This concept involving a separation of scaling regimes around centennial timescales was challenged by Nilsen et al. (2016). It was demonstrated that a spread in scaling parameters follows naturally from analyzing a range of proxy-based reconstructions for the Holocene covering different spatial regions and dynamical regimes. The occurrence of a second scaling regime was exclusively observed when analyzing timeseries including the last glacial period, which are nonstationary and involves nonlinearities that are not present for the Holocene climate. In the present paper it has been shown that both proxy noise and the BARCAST reconstruction technique contribute to alteration of the memory properties of the reconstructed data, introducing artifacts that may be interpreted as scale-breaks. None of these effects are intrinsic to the target data signal, but are introduced through non-climatic effects. Observing only the reconstruction may not give the complete answer on the temporal structure of the true temperature signal.

5.1 Implications for real proxy data

The spectral shape of the input pseudo proxy pseudoproxy data plotted in blue in Fig. 3 are similar to spectra of observed proxy data as observed in e.g. some types of tree-ring records, (Franke et al., 2013; Zhang et al., 2015; **?**). (Franke et al., 2013; Zhang et al., 2015; In particular, Franke et al. (2013); Zhang et al. (2015) found that the scaling parameters β were higher for tree-ring based reconstructions than for the corresponding instrumental data for the same region. **?** Werner et al. (2018) present a new spatial SAT reconstruction for the Arctic, using the BARCAST methodology. The reconstruction is based on annually layered records and layer counted archives with age uncertainties. The persistence properties of the input proxy records and the reconstructed temperatures were investigated using the same spectral techniques as here. analyses shown in Fig. A4 in **?** presents a map over the Arctic and an overview of the spatial distribution and type of proxy record. It also indicates if the proxy record is consistent with an AR(1) process null hypothesis or an fGn based on hypothesis testing. The analyses demonstrated demonstrate that several of the tree-ring records could not be categorized as neither AR(1) or scaling processes, but featured spectra similar to the pseudoproxy spectrum in our Fig. 3c-d. The characteristic flat spectrum at high frequencies, and the increased power on

~~bidecadal frequencies and lower can give the impression that the low-frequency power is inflated. However, from the presented experiments we know it is rather the high frequency power that is affected by the added white noise.~~ We hypothesize that the possible mechanism(s) altering the variability can be due to effects of the tree-ring processing techniques, specifically the methods applied to eliminate the biological tree aging effect on the growth of the trees ~~(Briffa et al., 1992)~~(Cook et al., 1995). The actual ~~tree ring~~ tree-ring width is a superposition of the age-dependent curve, which is individual for a tree, and a signal that can often be associated with climatic effects on the tree growth process. To correct for the biological age-effect, the raw tree-ring growth values are often transformed into proxy indices using the Regional Curve Standardization technique (RCS, ~~Briffa et al. (1992)). This technique attempts~~ Briffa et al. (1992); Cook et al. (1995)), Age Band Decomposition (ABD, Briffa et al. (2001)), the signal-free processing (Melvin and Briffa, 2008) or other techniques. These techniques attempt to eliminate biological age effects on tree-growth while preserving low frequency variability. As an example, consider the RCS processing of tree-ring width as a function of age (Helama et al., 2017). For a number of individual tree-ring records, each record is aligned according to their biological years. The mean of all the series is then modelled as a negative exponential function (the RCS curve). To construct the RCS chronology, the raw, individual tree-ring width curves are divided by the mean RCS curve for the full region. The RCS chronology is then the average of the index individual records. It is likely that the shape of a particular tree-ring width spectrum reflects the uncertainty in the ~~RCS~~ standardization curve, which is expected to be largest at the timescales corresponding to the initial stage of a tree growth, where the slope of the growth curve is generally steeper (i.e. of the order of a few decades). In particular, there may be slightly different climate processes affecting the growth of different trees, causing localized nonlinearities that limit the representativeness of the derived ~~exponential RCS~~chronology. We therefore suggest that the observed excess of LRM properties in some of the tree ring-based proxy records could be an artifact of the fitting procedure.

Our study further suggests that for a proxy network of high quality and density, exhibiting LRM properties, the BARCAST methodology is without modification capable of constructing skillful reconstructions with LRM preserved across the region. This is because the data information overwhelms the vague priors. The availability of well-documented proxy records therefore helps the analyst select an appropriate reconstruction method based on the input data. For quantification and assessment of real-world proxy quality, forward proxy modelling is a powerful tool that models proxy growth/deposition instead of the target variable evolution, also taking known proxy uncertainties and biases into consideration. See for example Dee et al. (2017) for a comprehensive study on terrestrial proxy system modeling, and Dolman and Laepple (2018) on forward modelling of sediment-based proxies.

5.2 Concluding remarks

~~A natural continuation of the pseudoproxy study presented here would be to~~ Several extensions to the presented work appears relevant for future studies, including (a) implementing external forcing and responses to these forcings in the target data to make the numerical experiments more realistic, (b) generate target data using a more complex model ~~. The~~ than described in Sect.2.2, and (c) reformulate BARCAST model Eq. 1-2. to account for LRM properties in the target data. In addition, there is a

possibility of repeating the experiments from this study using a different reconstruction technique, and experiments with more complicated spatiotemporal design of the multiproxy network can also be considered (Smerdon, 2012; Wang et al., 2014).

The alternatives (a) and (b) can be implemented together. Relevant advancements for target data generation can be obtained using the class of stochastic-diffusive models, such as the models described in North et al. (2011); Rypdal et al. (2015)~~make interesting candidates because of the~~ . The alternative method for generating spatial covariance ~~. The reconstruction~~ stands in contrast to what is done in the present study. The data generation technique used in this paper and also in Werner and Tingley (2015) generate a signal without spatial dynamics, where the spatial covariance is defined through the noise term. On the other hand, the stochastic-diffusive models generate the spatial covariance through the diffusion, without spatial structure in the noise term. The latter model type may be considered more physically correct and intuitive than the simplistic model used here. North et al. (2011) use an exponential model for the temporal covariance structure, while Rypdal et al. (2015) use an LRM model. ~~However, the intention of the present study was to conduct experimentswhere the target data follows all the model assumptions of BARCAST, except for the temporal correlation structure. Since this small modification had a pronounced effect on the reconstructions it is likely that using a different model would have even larger influence. Using the stochastic diffusive models in either North et al. (2011) or Rypdal et al. (2015) it would also be possible to implement external forcing and responses to these forcings in the target data to make the numerical experiments more realistic.~~

~~Another extension proposed for BARCAST already in Tingley and Huybers (2010a) was to generalize~~ For the BARCAST CFR methodology, reformulation of model Eq. 1-2 would drastically improve the performance in our experiments. However, at present 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~~using the Matérn covariance function. This would make the assumptions of BARCAST more realistic with respect to teleconnections~~ , accounting for teleconnections introduce similar computational challenges. The more general Matérn covariance family form (Tingley and Huybers, 2010a already been implemented for BARCAST, but was not used in this study. Another problem is the potential temporal instability of teleconnections; it is possible that major climate modes might have changed their configuration through time. ~~The change has been implemented but slows the algorithm down substantially in its present form.~~

~~Smerdon (2012) further proposed a number of improvements to be implemented in future pseudoproxy experiments, including accounting for temporal nonuniform availability of proxy data. Most studies overestimate the proxy sampling in the earlier part of the reconstruction period when using temporally invariant pseudoproxy networks. BARCAST is fully capable of taking such networks as input. Wang et al. (2014) later performed pseudo proxy experiments using four different CFR techniques, and tested two types of proxy networks: one that is fixed through the entire reconstruction period and one where the number of proxy records is reduced back in time, (staircase network). The effect of temporal heterogeneities is unexpectedly that the reconstruction skill does not decrease strictly following the proxy availability. Strong forcing events has a larger impact on the skill according to that study~~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 pseudoproxy study presented here ~~is based on simplistic target data generated using a novel technique~~sets a powerful example for how to construct and utilize an experimental structure to isolate specific properties of paleoclimate reconstruction techniques. The generation of the input data requires far less computation power and time than for GCM paleoclimatic simulations, but also results in less realistic target temperature fields. ~~However, we~~ We demonstrate that there are many areas of use for these types of data, including statistical modelling and hypothesis testing. ~~In particular, the pseudoproxy experiment presented here may be replicated using different index or field reconstruction techniques.~~

Appendix A: ~~**Information on true parameters, prior and posterior distributions of BARCAST parameters**~~

Appendix A: **Estimation of the periodogram**

The periodogram is defined here in terms of the discrete Fourier transform H_m as (Malamud and Turcotte, 1999):

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

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.

Appendix B: Continuous ranked probability score (CRPS) for a reconstruction ensemble

For probabilistic forecasts, scoring rules are used to measure the forecast accuracy, and proper scoring rules secure that the maximum reward is given when the true probability distribution is reported. In contrast, the reduction of error (RE) and coefficient of efficiency (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 the probability distribution P is used instead of the predictive distribution P itself.

The concept behind the CRPS is to provide a metric of the distance between the predicted (forecasted) and occurred (observed) cumulative distribution functions of the variable of interest. The lowest possible value for the metric corresponding to a perfect forecast is therefore CRPS=0. Following Sect. 4.b of Hersbach (2000), (elaborated for clarity) the definition of the CRPS and its subcomponents can be defined as follows:

$$\mathrm{CRPS}(P, x_{\mathrm{target}}) = \int_{-\infty}^{\infty} [P(x) - \Theta(x - x_{\mathrm{target}})]^2 dx \tag{B1}$$

Where x is the variable of interest, x_{target} denotes target (validation) data, Θ is the unit step function and $P(x)$ is the cumulative distribution function of the forecast ensemble with a probability density function (PDF) of $\rho(y)$:

$$P(x) = \int_{-\infty}^{x} \rho(y) dy \tag{B2}$$

In the case of a reconstruction ensemble at each spatial location j and time step t (omitted for convenience), the CPRS can be evaluated as:

$$\mathrm{CPRS} = \sum_{i=0}^{N} \int_{x_i}^{x_{i+1}} \left[\frac{i}{N} - \Theta(x - x_{\mathrm{target}}) \right]^2 dx \tag{B3}$$

Where $x_i < x < x_{i+1}$ refer to members of the locally ordered reconstruction ensemble of length N. For this study, x corresponds to the ensemble of local reconstructed values of T.

For the average over K time- and/or grid points, the average CRPS ($\overline{\mathrm{CPRS}}$) is defined as a weighted sum with equal weights, yielding:

$$\overline{\mathrm{CPRS}} = \sum_{i=0}^{N} \overline{g}_i \left[(1 - \overline{o}_i)(\frac{i}{N})^2 + \overline{o}_i (1 - \frac{i}{N})^2 \right] \tag{B4}$$

Here, $\overline{g}_i$ and $\overline{o}_i$ define two quantities characterizing the reconstruction ensemble and its link with the verifying target data. The quantity $\overline{g}_i = \overline{x_{i+1} - x_i}$, $i \in (0, N)$ represents the average over K distances between the neighboring members i and $i+1$ of the locally ordered reconstruction ensemble, and essentially quantifies the ensemble spread. The quantity $\overline{o}_i$ in turn is related with the average over K the frequency of the verifying target analysis x_{target} to be below $\frac{1}{2}(x_i + x_{i+1})$, and should ideally match the forecasted probability of i/N.

It can be demonstrated that the spatially and/or temporally averaged CRPS can further be broken into two parts: the average reliability score metric ($\overline{\mathrm{Reli}}$) and the average potential CRPS ($\overline{\mathrm{CRPS}}_{\mathrm{pot}}$):

$$\overline{\mathrm{CPRS}} = \overline{\mathrm{Reli}} + \overline{\mathrm{CRPS}}_{\mathrm{pot}}$$

where

$$\overline{\mathrm{Reli}} = \sum_{i=0}^{N} \overline{g}_i \left(\overline{o}_i - \frac{i}{N} \right)^2 \tag{B5}$$

$$\overline{\mathrm{CRPS}}_{\mathrm{pot}} = \sum_{i=0}^{N} \overline{g}_i \overline{o}_i (1 - \overline{o}_i) \tag{B6}$$

Eq. B5 suggests that $\overline{\mathrm{Reli}}$ summarizes second order statistics on the consistency between the average frequency of $\overline{o}_i$ of the verifying analysis to be found below the middle of interval number i and i/N, estimating thereby how well the nominal coverage rates of the ensemble reconstructions correspond to the empirical (target-based) ones. Hence $\overline{\mathrm{Reli}}$ represents the metric for assessing the validity of the uncertainty bands. $\overline{\mathrm{Reli}}$ can also be interpreted as the MSE of the confidence intervals, which in a perfectly reliable system has $\overline{\mathrm{Reli}}$=0.

$\overline{\mathrm{CRPS}}_{\mathrm{pot}}$ in turn measures the accuracy of the reconstruction itself, quantifying the spread of the ensemble and the mismatch between the best estimate and the target variable. Eq. B6 demonstrates that the smaller $\overline{g}_i$; indicative of a more narrow reconstruction ensemble, the lower the resulting $\overline{\mathrm{CRPS}}_{\mathrm{pot}}$ is. At the same time $\overline{\mathrm{CRPS}}_{\mathrm{pot}}$ takes into account the effect of outliers, i.e. the cases with $x_{\mathrm{target}} \notin [x_1, x_N]$. Although the reconstruction ensemble can be compact around its local mean, too frequent outliers will have a clear negative impact on the resulting $\overline{\mathrm{CRPS}}_{\mathrm{pot}}$. Note that this metric is akin to the Mean Absolute Error of a deterministic forecast which achieves its minimal value of zero only in the case of a perfect forecast.

Both scores are given in the same unit as the variable under study, here surface temperature.

Appendix C: Information on true parameters, prior and posterior distributions of BARCAST parameters

The forms of the prior PDF's for the scalar parameters in BARCAST are identical to those used in (Werner et al., 2013). The values of the hyperparameters were chosen after analyzing the target data. The forms of the priors and the values of the hyperparameters are listed in ~~Tab.~~Table C1.

The parameter values prescribed for the target data are listed in ~~Tab.~~Table C2. The instrumental observations are identical to the true target values, and the instrumental error variance τ_I^2 is therefore zero. The proxy noise variance τ_P^2 is varied systematically for the different SNR through the relation in Eq. 12

The mean of the posterior distributions of the BARCAST parameters $\alpha, \mu, \sigma^2, 1/\phi$, τ_I^2 and β_0 are listed in ~~Tab.~~Table C3, together with the reconstructed SNR.

Competing interests. The authors declare that they have no conflict of interest.

Acknowledgements. T.N. was supported by the Norwegian Research Council (KLIMAFORSK program) under grant no. 229754, and partly by Tromsø Research Foundation via the UiT project A31054.

D. V. D was partly supported by Tromsø Research Foundation via the UiT project A33020.

J.P.W. gratefully acknowledges support from the Centre for Climate Dynamics (SKD) at the Bjerknes Centre.

D.D.V., T.N. and J.P.W. also acknowledge the IS-DAAD project 255778 HOLCLIM for providing travel support. The authors would like to thank K. Rypdal for helpful discussions and comments.

References

Beran, J., Feng, Y., Ghosh, S., and Kulik, R.: Long-Memory Processes, Springer, New York, 884 pp., 2013.

Briffa, K. R., Jones, P. D., Bartholin, T. S., Eckstein, D., Schweingruber, F. H., Karlén, W., Zetterberg, P., and Eronen, M.: Fennoscandian summers from ad 500: temperature changes on short and long timescales, Climate Dynamics, 7, 111–119, doi:10.1007/BF00211153, 1992.

Briffa, K. R., Osborn, T., Schweingruber, F. H., Harris, I. C., Jones, P. D., Shiyatov, S. G., and Vaganov, E.: Low frequency temperature variations from a northern tree ring density network, Journal of Geophysical Research: Atmospheres, 106, 2929–2941, doi:10.1029/2000JD900617, 2001.

Christiansen, B.: Reconstructing the NH Mean Temperature: Can Underestimation of Trends and Variability Be Avoided?, Journal of Climate, 24, 674–692, doi:10.1175/2010JCLI3646.1, 2011.

Christiansen, B. and Ljungqvist, F. C.: Challenges and perspectives for large-scale temperature reconstructions of the past two millennia, Reviews of Geophysics, 55, 40–96, doi:10.1002/2016RG000521, 2017.

Cook, E. R., Briffa, K. R., Meko, D. M., Graybill, D. A., and Funkhouser, G.: The 'segment length curse' in long tree-ring chronology development for palaeoclimatic studies, The Holocene, 5, 229–237, doi:10.1177/095968369500500211, 1995.

Dee, S., Parsons, L., Loope, G., Overpeck, J., Ault, T., and Emile-Geay, J.: 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, doi:https://doi.org/10.1016/j.epsl.2017.07.036, http://www.sciencedirect.com/science/article/pii/S0012821X1730420X, 2017.

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

Emile-Geay, J. and Tingley, M.: Inferring climate variability from nonlinear proxies: application to palaeo-ENSO studies, Climate of the Past, 12, 31–50, doi:10.5194/cp-12-31-2016, 2016.

Fraedrich, K. and Blender, R.: Scaling of Atmosphere and Ocean Temperature Correlations in Observations and Climate Models, Phys. Rev. Lett., 90, 108 501, doi:10.1103/PhysRevLett.90.108501, 2003.

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

Franzke, C. L. E., Graves, T., Watkins, N. W., Gramacy, R. B., and Hughes, C.: Robustness of estimators of long-range dependence and self-similarity under non-Gaussianity, Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences, 370, 1250–1267, doi:10.1098/rsta.2011.0349, 2012.

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

Fredriksen, H.-B. and Rypdal, M.: Long-Range Persistence in Global Surface Temperatures Explained by Linear Multibox Energy Balance Models, Journal of Climate, 30, 7157–7168, doi:10.1175/JCLI-D-16-0877.1, 2017.

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

Gneiting, T. and Raftery, A. E.: Strictly Proper Scoring Rules, Prediction, and Estimation, Journal of the American Statistical Association, 102, 359–378, doi:10.1198/016214506000001437, 2007.

Gómez-Navarro, J. J., Werner, J., Wagner, S., Luterbacher, J., and Zorita, E.: Establishing the skill of climate field reconstruction techniques for precipitation with pseudoproxy experiments, Climate Dynamics, 45, 1395–1413, 2015.

Hasselmann, K.: Stochastic climate models Part I. Theory, Tellus, 28, 473–485, doi:10.1111/j.2153-3490.1976.tb00696.x, 1976.

Helama, S., Melvin, T. M., and Briffa, K. R.: Regional curve standardization: State of the art, The Holocene, 27, 172–177, doi:10.1177/0959683616652709, 2017.

Hersbach, H.: Decomposition of the Continuous Ranked Probability Score for Ensemble Prediction Systems, Weather and Forecasting, 15, 559–570, doi:10.1175/1520-0434(2000)015<0559:DOTCRP>2.0.CO;2, 2000.

Huybers, P. and Curry, W.: Links between annual, Milankovitch and continuum temperature variability, Nature, 441, doi:10.1038/nature04745, 2006.

Jones, P. D., Briffa, K. R., Barnett, T. P., and Tett, S. F. B.: High-resolution palaeoclimatic records for the last millennium: interpretation, integration and comparison with General Circulation Model control-run temperatures, The Holocene, 8, 455–471, doi:10.1191/095968398667194956, 1998.

Koscielny-Bunde, A. B., Havlin, S., and Goldreich, Y.: Analysis of daily temperature fluctuations, Physica A, 231, 393–396, doi:10.1016/0378-4371(96)00187-2, 1996.

Lee, T. C. K., Zwiers, F. W., and Tsao, M.: Evaluation of proxy-based millennial reconstruction methods, Climate Dynamics, 31, 263–281, doi:10.1007/s00382-007-0351-9, 2008.

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

Luterbacher, J., Werner, J. P., Smerdon, J. E., Fernández-Donado, L., González-Rouco, F. J., Barriopedro, D., Ljungqvist, F. C., Büntgen, U., Zorita, E., Wagner, S., Esper, J., McCarroll, D., Toreti, A., Frank, D., Jungclaus, J. H., Barriendos, M., Bertolin, C., Bothe, O., Brázdil, R., Camuffo, D., Dobrovolný, P., Gagen, M., García-Bustamante, E., Ge, Q., Gómez-Navarro, J. J., Guiot, J., Hao, Z., Hegerl, G. C., Holmgren, K., Klimenko, V. V., Martín-Chivelet, J., Pfister, C., Roberts, N., Schindler, A., Schurer, A., Solomina, O., von Gunten, L., Wahl, E., Wanner, H., Wetter, O., Xoplaki, E., Yuan, N., Zanchettin, D., Zhang, H., and Zerefos, C.: European summer temperatures since Roman times, Environmental Research Letters, 11, 024 001, doi:10.1088/1748-9326/11/2/024001, 2016.

Malamud, B. D. and Turcotte, D. L.: Self-affine time series: measures of weak and strong persistence, Journal of Statistical Planning and Inference, 80, 173–196, doi:https://doi.org/10.1016/S0378-3758(98)00249-3, 1999.

Mann, M. E., Bradley, R. S., and Hughes, M. K.: Global-scale temperature patterns and climate forcing over the past six centuries, Nature, 392, 779–787, 1998.

Mann, M. E., Rutherford, S., Wahl, E., and Ammann, C.: Testing the Fidelity of Methods Used in Proxy-Based Reconstructions of Past Climate, Journal of Climate, 18, 4097–4107, doi:10.1175/JCLI3564.1, 2005.

Mann, M. E., Rutherford, S., Wahl, E., and Ammann, C.: Robustness of proxy-based climate field reconstruction methods, J. Geophys. Res., 112, n/a–n/a, 2007.

Mann, M. E., Zhang, Z., Hughes, M. K., Bradley, R. S., Miller, S. K., Rutherford, S., and Ni, F.: Proxy-based reconstructions of hemispheric and global surface temperature variations over the past two millennia, Proceedings of the National Academy of Sciences, 105, 13 252–13 257, doi:10.1073/pnas.0805721105, 2008.

Mann, M. E., Zhang, Z., Rutherford, S., Bradley, R. S., Hughes, M. K., Shindell, D., Ammann, C., Faluvegi, G., and Ni, F.: Global Signatures and Dynamical Origins of the Little Ice Age and Medieval Climate Anomaly, Science, 326, 1256–1260, 2009.

Melvin, T. M. and Briffa, K. R.: A signal-free approach to dendroclimatic standardisation, Dendrochronologia, 26, 71 – 86, doi:https://doi.org/10.1016/j.dendro.2007.12.001, 2008.

Moberg, A., Sonechkin, D. M., Holmgren, K., Datsenko, N. M., and Karlén, W.: Highly variable Northern Hemisphere temperatures reconstructed from low-and high-resolution proxy data, Nature, 433, 613–617, doi:10.1038/nature03265, 2005.

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

North, G. R., Wang, J., and Genton, M. G.: Correlation models for temperature fields, J. Climate, 24, 5850–5862, doi:10.1175/2011JCLI4199.1., 2011.

PAGES 2k Consortium: Continental-scale temperature variability during the past two millennia, Nature Geosci, 6, 339–346, doi:10.1038/ngeo1797, 2013.

Peng, C.-K., Buldyrev, S. V., Havlin, S., Simons, M., Stanley, H. E., and Goldberger, A. L.: Mosaic organization of DNA, Physical Review E, 49, 1685–1689, doi:http://dx.doi.org/10.1103/PhysRevE.49.1685, 1994.

Rybski, D., Bunde, A., Havlin, S., and von Storch, H.: Long-term persistence in climate and the detection problem, Geophys. Res. Lett., 33, n/a–n/a, doi:10.1029/2005GL025591, 2006.

Rypdal, K., Østvand, L., and Rypdal, M.: Long-range memory in Earth's surface temperature on time scales from months to centuries, J. Geophys. Res., 118, 7046–7062, doi:10.1002/jgrd.50399, 2013.

Rypdal, K., Rypdal, M., and Fredriksen., H. B.: Spatiotemporal Long-Range Persistence in Earth's Temperature Field: Analysis of Stochastic-Diffusive Energy Balance Models, J. Climate, 28, 8379–8395, doi:10.1175/JCLI-D-15-0183.1, 2015.

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

Rypdal, M. and Rypdal, K.: Late Quaternary temperature variability described as abrupt transitions on a $1/f$ noise background, Earth System Dynamics, 7, 281–293, doi:10.5194/esd-7-281-2016, https://www.earth-syst-dynam.net/7/281/2016/, 2016.

Schneider, T.: Analysis of Incomplete Climate Data: Estimation of Mean Values and Covariance Matrices and Imputation of Missing Values, Journal of Climate, 14, 853–871, doi:10.1175/1520-0442(2001)014<0853:AOICDE>2.0.CO;2, 2001.

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

Smerdon, J. E., Kaplan, A., Zorita, E., González-Rouco, J. F., and Evans, M. N.: Spatial performance of four climate field reconstruction methods targeting the Common Era, Geophysical Research Letters, 38, n/a–n/a, doi:10.1029/2011GL047372, l11705, 2011.

Tingley, M. P. and Huybers, P.: A Bayesian Algorithm for Reconstructing Climate Anomalies in Space and Time. Part I: Development and Applications to Paleoclimate Reconstruction Problems, Journal of Climate, 23, 2759–2781, doi:10.1175/2009JCLI3015.1, 2010a.

Tingley, M. P. and Huybers, P.: A Bayesian Algorithm for Reconstructing Climate Anomalies in Space and Time. Part II: Comparison with the Regularized Expectation-Maximization Algorithm, Journal of Climate, 23, 2782–2800, doi:10.1175/2009JCLI3016.1, 2010b.

Tingley, M. P. and Li, B.: Comments on "Reconstructing the NH Mean Temperature: Can Underestimation of Trends and Variability Be Avoided?, Journal of Climate, 25, 3441–3446, doi:10.1175/JCLI-D-11-00005.1, 2012.

Tipton, J., Hooten, M., Pederson, N., Tingley, M., and Bishop, D.: Reconstruction of late Holocene climate based on tree growth and mechanistic hierarchical models, Environmetrics, 27, 42–54, doi:10.1002/env.2368, 2015.

Wahl, E. R., Diaz, H. F., Vose, R. S., and Gross, W. S.: Multicentury Evaluation of Recovery from Strong Precipitation Deficits in California, Journal of Climate, 30, 6053–6063, doi:10.1175/JCLI-D-16-0423.1, 2017.

Wang, J., Emile-Geay, J., Guillot, D., Smerdon, J. E., and Rajaratnam, B.: Evaluating climate field reconstruction techniques using improved emulations of real-world conditions, Climate of the Past, 10, 1–19, doi:10.5194/cp-10-1-2014, 2014.

Wang, J., Emile-Geay, J., Guillot, D., McKay, N. P., and Rajaratnam, B.: Fragility of reconstructed temperature patterns over the Common Era: Implications for model evaluation, Geophysical Research Letters, 42, 7162–7170, doi:10.1002/2015GL065265, 2015.

Werner, J. P. and Tingley, M. P.: Technical Note: Probabilistically constraining proxy age-depth models within a Bayesian hierarchical reconstruction model, Climate of the Past, 11, 533–545, doi:10.5194/cp-11-533-2015, 2015.

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

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

Wonnacott, R. J. and Wonnacott, T. H.: Econometrics, Probability and Mathematical Statistics, John Wiley and Sons, New York, 604, 1979.

Zhang, H., Yuan, N., Esper, J., Werner, J. P., Xoplaki, E., Büntgen, U., Treydte, K., and Luterbacher, J.: Modified climate with long term memory in tree ring proxies, Environmental Research Letters, 10, 084 020, doi:10.1088/1748-9326/10/8/084020, 2015.

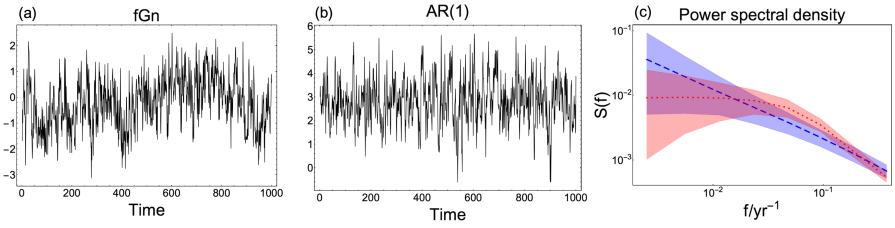


Figure 1. (a) Arbitrary fGn time series with $\beta = 0.75$. (b) Arbitrary time series of an AR(1) process with parameters estimated from the time series in (a) using Maximum Likelihood. (c) Log-log ~~spectrum~~ spectra showing 95% confidence ranges based on Monte Carlo ensembles of fGn with $\beta = 0.75$ (blue shaded area), and AR(1) processes with parameters estimated from the time series in (a) (red, shaded area). Dashed (dotted) lines mark the ensemble means of the fGn (~~AR(1~~AR1) ~~process)~~ spectra respectively.

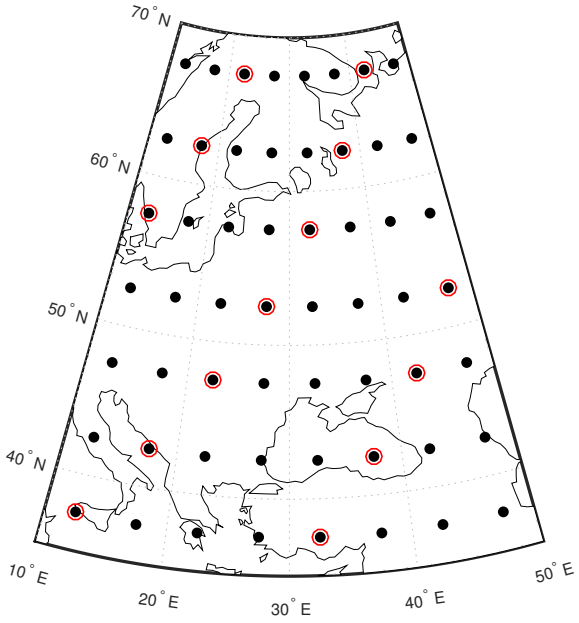


Figure 2. The spatial domain of the reconstruction experiments. Dots mark locations of instrumental sites, proxy sites are highlighted by red circles. The superimposed map of Europe provides a spatial scale.

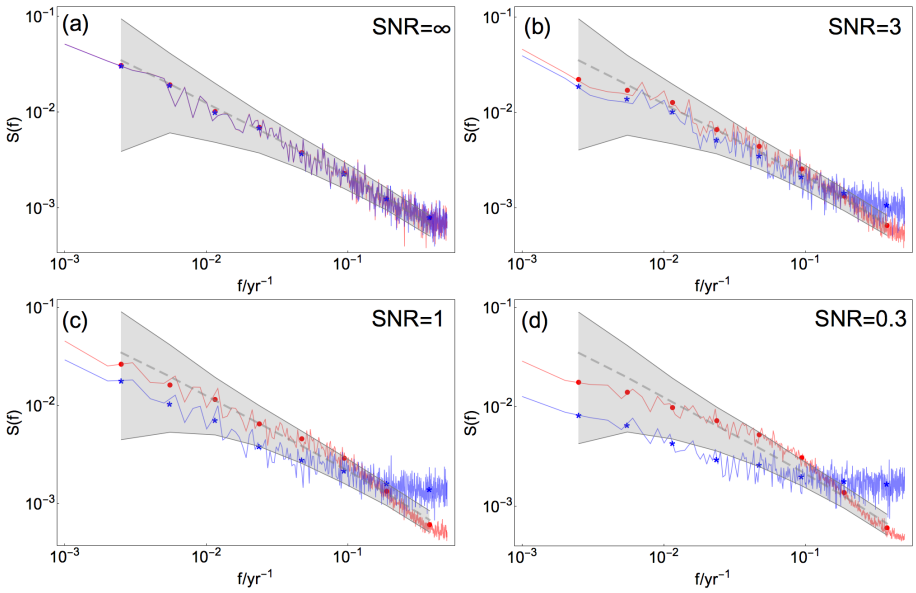


Figure 3. Mean raw and log-binned PSD for pseudoproxy data (blue curve and asterisks, respectively) and reconstruction at the same site (red curve and dots, respectively) generated from β_{target} =0.75 and different SNR indicated in the panels. Colored gray shadings and dashed, gray lines indicate 95% confidence range and the ensemble mean, respectively, for a Monte Carlo ensemble of fGn with β =0.75.

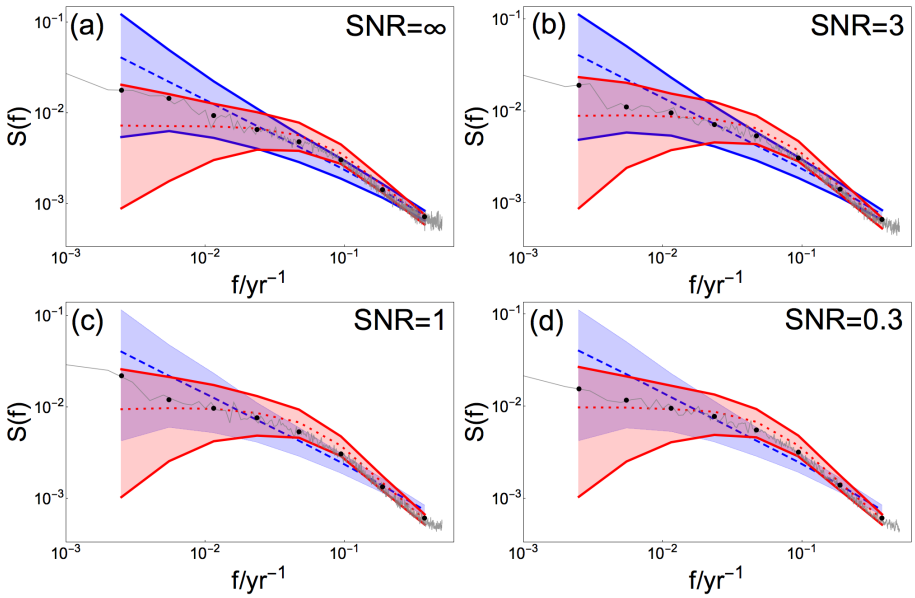


Figure 4. Mean raw and log-binned PSD for local reconstructed data at a site between proxies (gray curve and dots, respectively) generated from β_{target} =0.75 and different SNR. Colored shadings and dashed/dotted lines indicate 95% confidence range and the ensemble mean, respectively, for a Monte Carlo ensemble of fGn with β =0.75 (blue) and of AR(1) processes with α estimated from BARCAST (red). The confidence ranges found consistent with the data are drawn with solid lines.

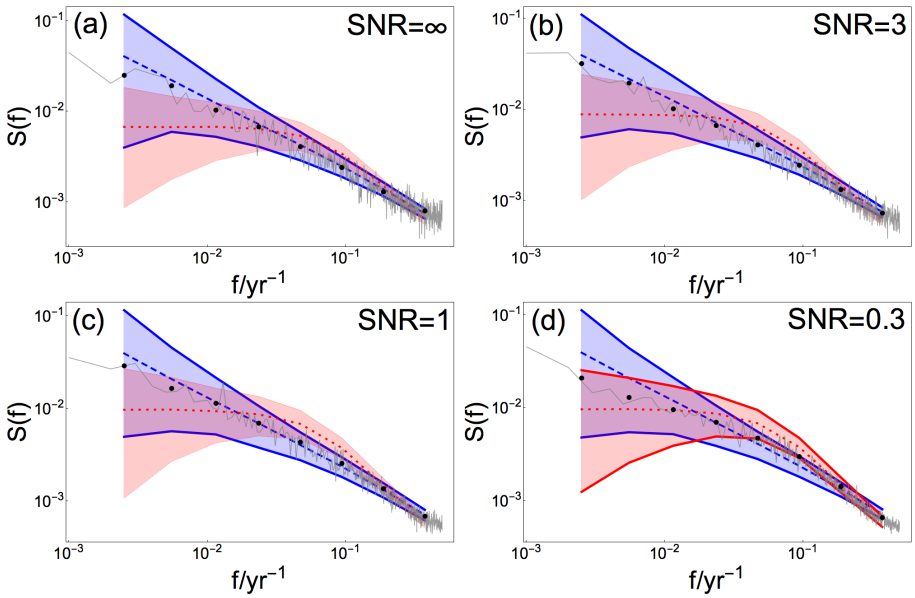


Figure 5. Mean raw and log-binned PSD for the spatial mean reconstruction (gray curve and dots, respectively), generated from β_{target} =0.75 and different SNR. Colored shadings and dashed/dotted lines indicate 95% confidence range and the ensemble mean, respectively, for a Monte Carlo ensemble of fGn with β =0.75 (blue) and of AR(1) processes with α estimated from BARCAST (red). The confidence ranges found consistent with the data are drawn with solid lines.

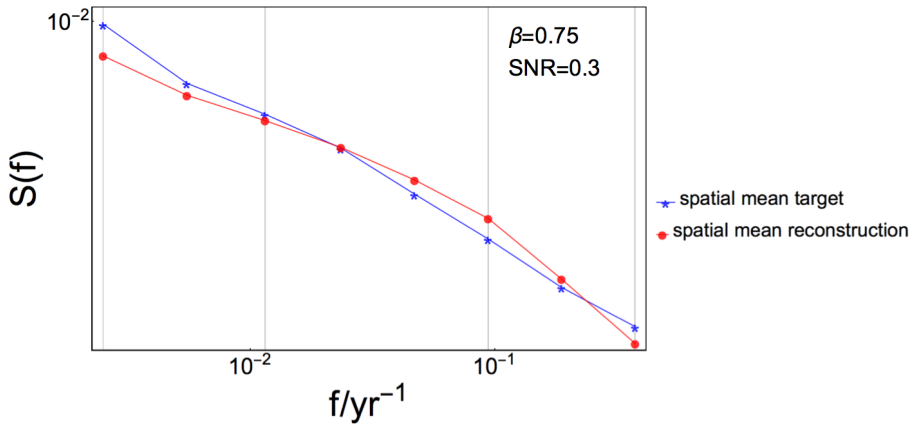


Figure 6. Log-log plot showing log-binned power spectra of spatial mean target (blue) and reconstruction (red) for one experiment. Vertical, gray lines mark the frequency ranges used to estimate bias of variance as referred to in Sect. 4.4.

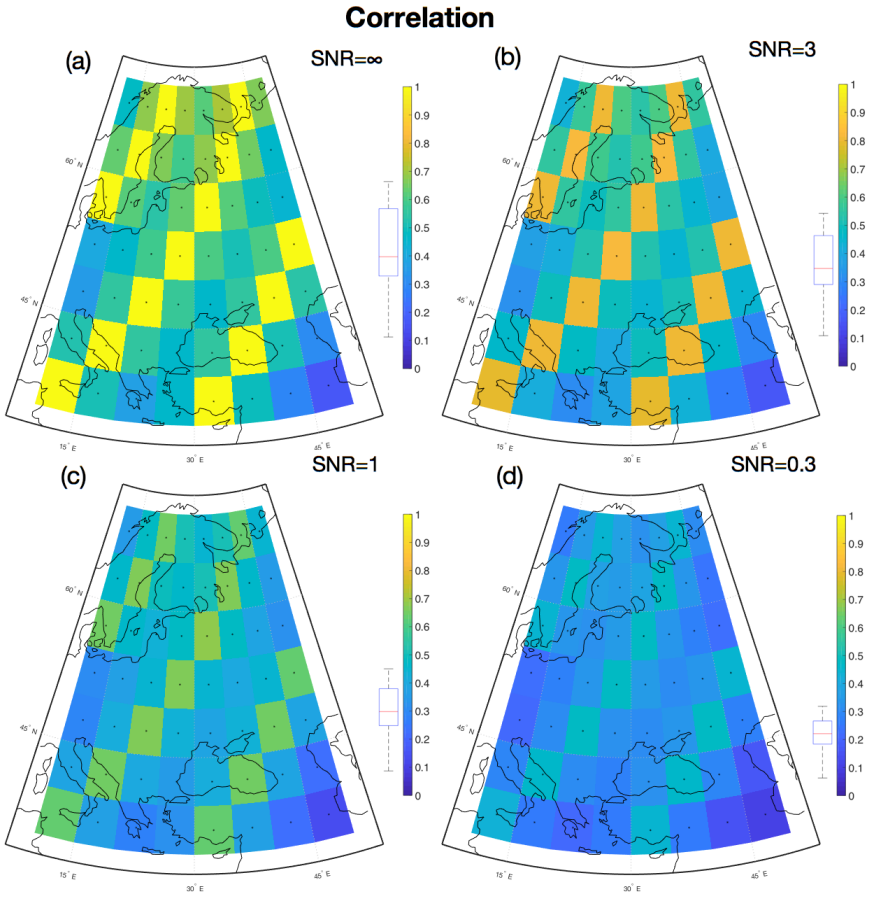


Figure 7. ~~Mean local~~ Local correlation coefficient between reconstructed temperature field and target field for the ~~reconstruction~~ verification period (ensemble mean). The boxplots left of the color bars indicate the distribution of grid point correlation coefficients. β =0.75

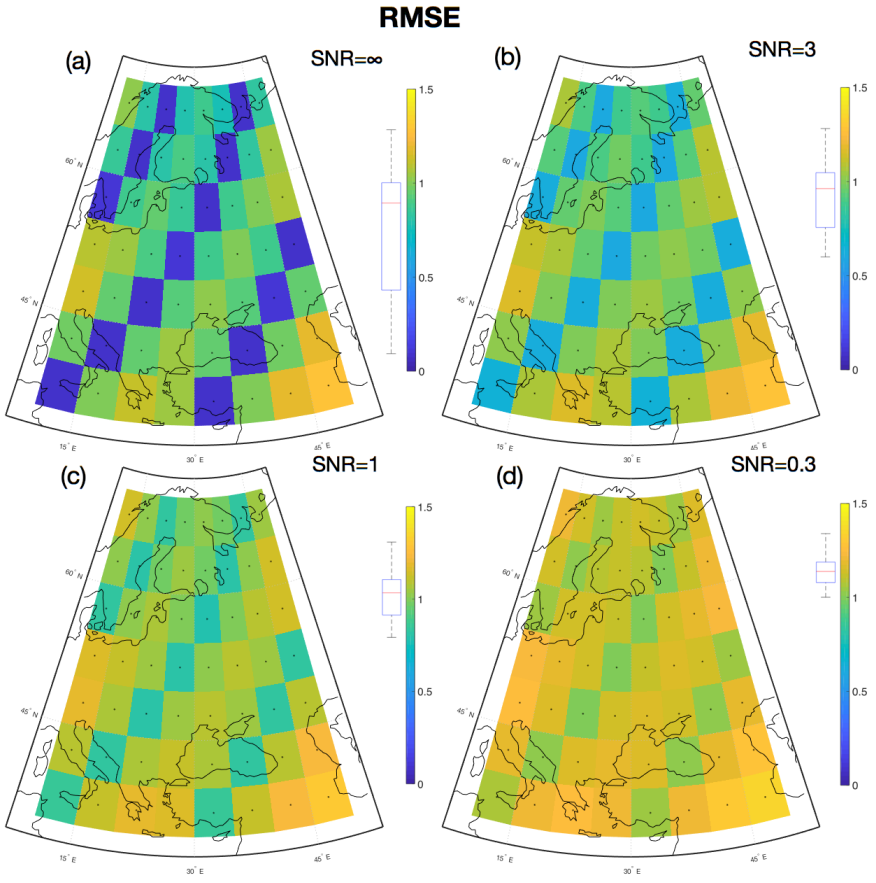
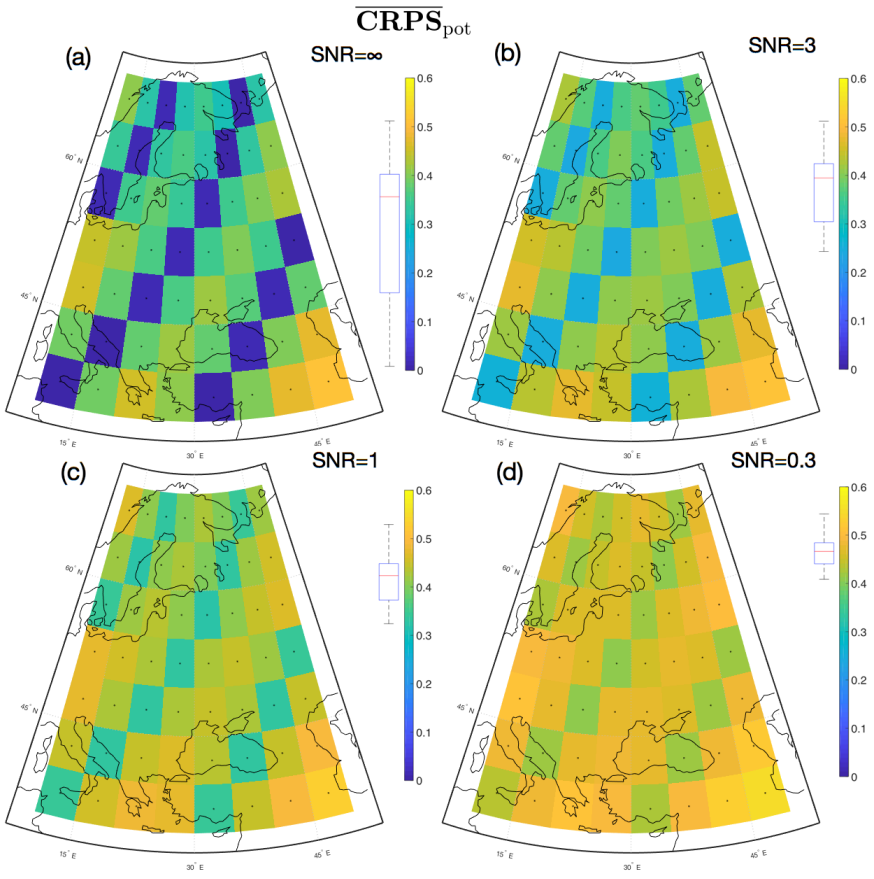


Figure 8. ~~Mean local~~ Local root-mean square error (RMSE) between reconstructed temperature field and target field for the ~~reconstruction~~ verification period. The boxplots left of the color bars indicate the distribution of grid point correlation coefficients. β =0.75.



~~Mean local $\overline{\mathrm{CRPS}}_{\mathrm{pot}}$ between reconstructed temperature field and target field. β =0.75.~~

Figure 9. ~~Mean local Reliability~~ Local $\overline{\mathrm{CRPS}}_{\mathrm{pot}}$ between reconstructed temperature field and target field ~~. β =0.75. Log-log plot showing log-binned power spectra of spatial mean target (blue) and reconstruction (red)~~ for ~~one experiment~~the verification period. ~~Vertical, gray lines mark~~ The boxplots left of the ~~frequency ranges used to estimate bias~~ color bars indicate the distribution of ~~variance as referred to in Se~~grid point correlation coefficients. ~~5~~β =0.75.

Table 1. Summary of the experiment setup.

Spatiotemporal resolution:	5x5 degrees /annual	
Strength of persistence:	β=0.55, 0.75, 0.95	
Noise level:	SNR=∞, 3, 1, 0.3	
Iterations before/after thinning:	5000/500	
	Input data	Reconstruction
Ensemble members per experiment	20	30 000

Table 2. Hypothesis testing results for local reconstructed data compared to Monte Carlo ensembles of fGn and AR(1) processes. The mark "x" in the table indicates that the null hypothesis cannot be rejected. The null hypotheses 1 and 2 are:

1: The reconstruction is consistent with the fGn structure in the target data for all frequencies.

2: The reconstruction is consistent with the AR(1) assumption from BARCAST for all frequencies.

Local field values				
SNR	∞	3	1	0.3
$\beta = 0.55$ Proxy site				
1	x	x	x	x
2:		x	x	x
$\beta = 0.55$ Between proxy sites				
1:	x	x	x	x
2:	x	x	x	x
$\beta = 0.75$ Proxy site				
1:	x	x	x	
2:				x
$\beta = 0.75$ Between proxy sites				
1:	x	x		
2:	x	x	x	x
$\beta = 0.95$ Proxy site				
1:	x	x	x	
2:				x
$\beta = 0.95$ Between proxy sites				
1:	x			
2:	x	x	x	x

Table 3. Hypothesis testing results for spatial mean reconstructed data compared to Monte Carlo ensembles of fGn and AR(1) processes. The null hypotheses 1 and 2 are the same as in Table 2, and the "x" has the same meaning.

Spatial mean values				
SNR	∞	3	1	0.3
$\beta = 0.55$				
1:	x	x	x	x
2:				x
$\beta = 0.75$				
1:	x	x	x	x
2:				x
$\beta = 0.95$				
1:	x	x	x	x
2:				

Table 4. ~~Mean~~ Median local skill measures

SNR	r	RMSE	$\overline{\mathrm{CRPS}}_{\mathrm{pot}}$ ~~Reliability~~
	$\beta = 0.55$		
∞	~~0.65~~0.60	~~0.74~~0.89	~~0.53~~0.30~~0.30~~0.36
3	~~0.54~~0.52	~~0.93~~0.97	~~0.63~~0.12~~0.12~~0.39
1	~~0.43~~0.42	~~1.04~~1.05	~~0.64~~6.6 $*$ ~~10~~$^{-2}$~~0.43~~
0.3	0.29	1.16	~~0.62~~3.1 $*$ ~~10~~$^{-2}$~~0.48~~
	$\beta = 0.75$		
∞	~~0.65~~0.60	~~0.74~~0.89	~~0.53~~0.3~~0.36~~
3	~~0.55~~0.52	~~0.92~~0.96	~~0.63~~0.13~~0.39~~
1	~~0.46~~0.45	~~1.02~~1.03	~~0.64~~7.6 $*$ ~~10~~$^{-2}$~~0.42~~
0.3	0.33	~~1.13~~1.14	~~0.63~~3.8 $*$ ~~10~~$^{-2}$~~0.47~~
	$\beta = 0.95$		
∞	~~0.66~~0.61	~~0.74~~0.88	~~0.53~~0.30~~0.35~~
3	~~0.56~~0.53	~~0.91~~0.95	~~0.63~~0.14~~0.39~~
1	~~0.49~~0.48	~~0.99~~1.01	~~0.65~~9.2 $*$ ~~10~~$^{-2}$~~0.41~~
0.3	~~0.39~~0.38	~~1.09~~1.10	~~0.64~~5.1 $*$ ~~10~~$^{-2}$~~0.45~~

Table 5. ~~Mean~~ Median skill measures for spatial mean

SNR	r	RMSE
	$\beta = 0.55$	
∞	~~0.97~~0.95	~~0.17~~0.18
3	~~0.86~~0.87	~~0.29~~0.28
1	~~0.75~~0.76	0.38
0.3	~~0.554~~0.564	~~0.52~~0.51
	$\beta = 0.75$	
∞	0.95	~~0.17~~0.18
3	~~0.85~~0.87	~~0.29~~0.28
1	~~0.8~~0.78	0.37
0.3	~~0.64~~0.60	~~0.51~~0.50
	$\beta = 0.95$	
∞	~~0.96~~0.94	~~0.17~~0.18
3	~~0.87~~0.88	0.28
1	~~0.82~~0.79	0.36
0.3	~~0.72~~0.66	~~0.45~~0.46

Table C1. List of parameters defined in BARCAST, form of prior and hyperparameters

Parameter	Form	Hyperparameters
α	Truncated normal	$\mathrm{N}_{[0,1]}(\alpha_\mu, \alpha_\sigma)$, $\alpha_\mu = 0.5$, $\alpha_\sigma = 0.1$
μ	Normal	N(μ_μ, μ_σ), $\mu_\mu = -0.4$, $\mu_\sigma = 0.1^2$
σ^2	Inv-gamma	shape=0.5, scale=0.5
ϕ	Lognormal	$\log\phi \sim N(\phi_\mu, \phi_\sigma)$, $\phi_\mu = -7$, $\phi_\sigma = 0.2$
τ_{I}^2	Inv-gamma	shape=0.5, scale=0.5
τ_{P}^2	Inv-gamma	shape=0.5, scale=0.5
β_0	Normal	N($\beta_{0,\mu}$, $\beta_{0,\sigma}$), $\beta_{0,\mu} = 0$, $\beta_{0,\sigma} = 0.04$
β_1	Normal	N($\beta_{1,\mu}$, $\beta_{1,\sigma}$), $\beta_{1,\mu} = 1.14$, $\beta_{1,\sigma} = 0.04$

Table C2. List of parameter values defined for the target data set. The four values of τ_P^2 listed are related to the four different signal-to-noise ratios: $\mathrm{SNR} = \frac{1}{\tau_P^2}$. ϵ_{mach} is machine epsilon, the smallest number represented by the computer which is greater than zero.

Parameter	Target value
μ	0
ϕ	1/1000
τ_I^2	0
τ_P^2	ϵ_{mach}, 0.333, 1, 3.33
β_0	0
β_1	1
β	0.5, 0.75, 0.95

Table C3. Mean of posterior distribution for each parameters

Persistence	$\mathrm{SNR}_{\mathrm{target}}$	$\mathrm{SNR}_{\mathrm{rec}}$	α	μ	σ^2	$1/\phi$	τ_I^2	β_0
$\beta = 0.55$	∞	117.6	0.40	$-2.2 * 10^{-2}$	0.83	1020	$2.3 * 10^{-3}$	$6.7 * 10^{-4}$
	3	3.05	0.43	$-2.6 * 10^{-2}$	0.78	1053	$2.4 * 10^{-2}$	$-2.7 * 10^{-3}$
	1	1.01	0.44	$-3.3 * 10^{-2}$	0.75	1064	$3.0 * 10^{-2}$	$3.8 * 10^{-3}$
	0.3	0.38	0.44	$-2.7 * 10^{-2}$	0.75	1053	$3.3 * 10^{-2}$	$3.2 * 10^{-3}$
$\beta = 0.75$	∞	115.8	0.57	$-3.5 * 10^{-2}$	0.68	1020	$2.3 * 10^{-3}$	$-8.2 * 10^{-4}$
	3	2.81	0.62	$-4.5 * 10^{-2}$	0.61	1111	$2.8 * 10^{-2}$	$5.1 * 10^{-3}$
	1	0.99	0.64	$-5.6 * 10^{-2}$	0.59	1136	$3.3 * 10^{-}2$	$8.3 * 10^{-3}$
	0.3	0.36	0.64	$-5.1 * 10^{-2}$	0.59	1136	$3.4 * 10^{-2}$	$3.8 * 10^{-3}$
$\beta = 0.95$	∞	112.9	0.71	$-4.8 * 10^{-2}$	0.5	1020	$2.4 * 10^{-3}$	$-5.3 * 10^{-4}$
	3	2.69	0.77	$-8.5 * 10^{-2}$	0.44	1205	$2.8 * 10^{-2}$	$3.0 * 10^{-3}$
	1	0.97	0.79	$-9.7 * 10^{-2}$	0.41	1235	$3.1 * 10^{-2}$	$1.1 * 10^{-2}$
	0.3	0.36	0.77	$-1.0 * 10^{-1}$	0.42	1190	$2.9 * 10^{-2}$	$1.6 * 10^{-2}$