1 General comments to all reviewers and the editor

We are very grateful for the input from the three anonymous reviewers and hope that all major points have been answered satisfactory. Please see the individual responses to reviewers 1, 2 and 3 on matters that are not discussed in this general reply.

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.

The most important changes requested by all three reviewers involve (1) changing the title, (2) rewriting the abstract and (3) rewrite discussion/conclusions to be more precise and informative. Other relevant modifications we find necessary is to (4) rewrite/clarify Sect. 2.2 on target data generation to avoid confusion and misconceptions, (5) elaborate on the CRPS skill metric in Sect. 4.4, and inserting selected formulas into the appendix. At last, (6) we have discovered an error in the calculation of the average CRPS scores that influences our skill metric results.

(1) The new suggested 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 will be rewritten and shortened in the revision, as requested. Here is a preliminary suggestion:

> The skill of the state-of-the-art climate field reconstruction technique BARCAST ("Bayesian Algorithm for Reconstructing Climate Anomalies in Space and Time") to reconstruct climate with pronounced LRM characteristics is tested. A novel technique for generating the target data has been developed and is used to provide ensembles of long-range memory stochastic processes with a prescribed spatial covariance structure. Based on different parameter setups, hypothesis testing in the spectral domain is 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 (AR1) null hypothesis which is the assumed temperature model for this reconstruction technique. The study reveals that the resulting field and spatial mean reconstructions are consistent with the fGn hypothesis for most of the tested parameter configurations. There are local differences in reconstructed scaling characteristics between individual grid cells, and the agreement with the fGn model is generally better for the spatial mean reconstruction than at individual locations. Some parameter setups were found to give reconstructions consistent with the AR(1) model, while others were not. Our results show 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.

(3) Given the opportunity to submit a revised manuscript, the discussion and conclusions will be rewritten to address the comments from all three reviewers. Reformulating this section will take more time, but in general the major changes will be:

- References to which figures/tables we refer to will be made clear as the discussion proceeds.
- Text editing to avoid confusion in several paragraphs, ref comments from reviewer 1 and 3 in particular.
- More precise conclusions regarding the skill metric results using the CRPS and the subcomponents $\overline{\text{Reli}}$ and $\overline{\text{CRPS}}_{\text{pot}}$ (see also 6).
- Consideration of other tree-ring processing techniques, including the age-band decomposition and the signal free processing method.
- Relating our results to to the topic of proxy system modelling, ref. Dee et al. (2017); Dolman and Laepple (2018). Also discuss prospective recommendations to improvements of BARCAST.
- Final conclusions will 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.

(4) Rewriting Sect.2.2, see suggestion at page 5-6 of this document.

(5) See suggestion at page 7-9 of this document for elaboration on the CRPS, its subcomponents and explaining why the RE/CE metrics are improper for probabilistic reconstruction methods.

(6) When calculating the skill score metrics, the first year should not be included, since it is used to initialize the reconstruction algorithm and it is not identical to the target value. By accident this year was included when calculating the average CRPS in the original manuscript, and being an outlier of a large magnitude relative to the mean, it had a profound negative effect on the skill metrics we used. After recalculations we find some changes to our results. The $\overline{\text{CRPS}}_{\text{pot}}$ is moderately reduced, see Fig. 1 replacing Fig. 8 in the manuscript.

On the other hand, the Reliability (using overline to indicate average from now on): $\overline{\text{Reli}}$ is drastically reduced after removing the initial year, implying a good consistency

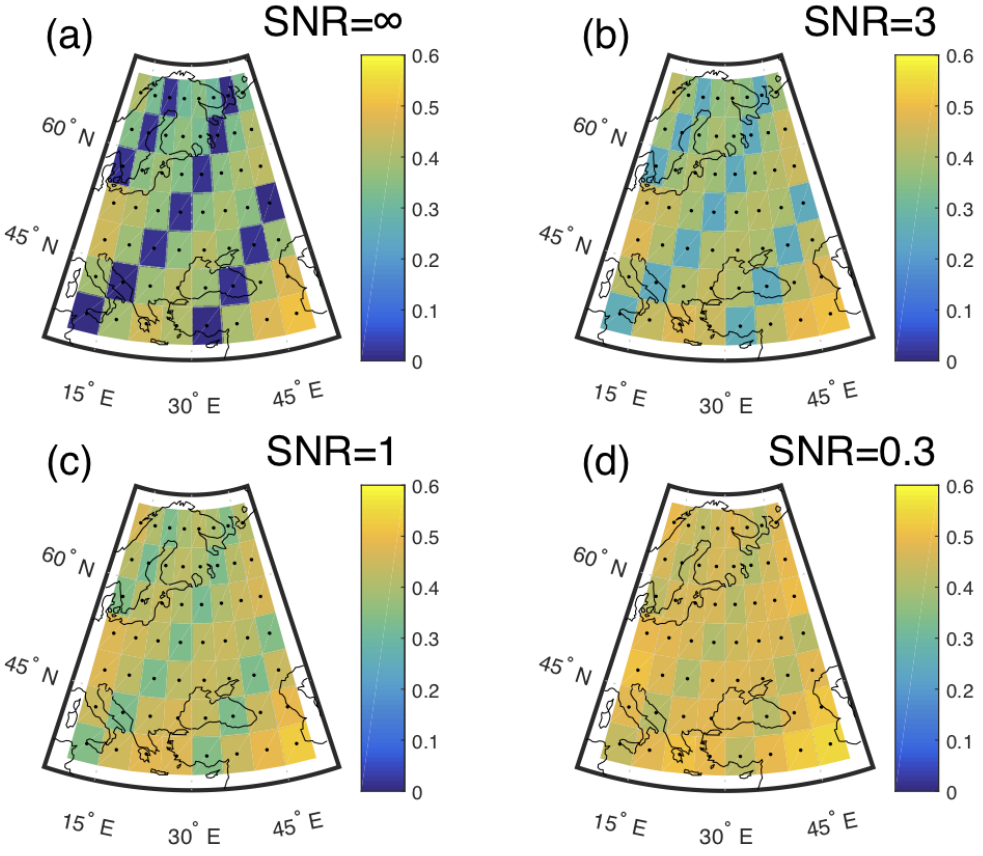

Figure 1: $\overline{\mathrm{CRPS}}_{\mathrm{pot}}$ for $\beta = 0.75$

of the theoretical and empirical CIs. The scores are now consistently below 0.1. The $\overline{\mathrm{Reliability}}$ scores are now of such small orders that we don't find it meaningful to discuss them in the revised manuscript. Hence, Sect. 4.4.1 will have to be rewritten, and figure 9 will be removed. However, the general distinction between the $\overline{\mathrm{CRPS}}_{\mathrm{pot}}$ and the $\overline{\mathrm{Reli}}$ will still be included in 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.

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.

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.

H. Hersbach. Decomposition of the continuous ranked probability score for ensemble prediction systems. *Weather and Forecasting*, 15(5):559–570, 2000. doi: 10.1175/1520-0434(2000)015⟨0559:DOTCRP⟩2.0.CO;2.

J. E. Smerdon, A. Kaplan, E. Zorita, J. F. González-Rouco, and M. N. Evans. Spatial performance of four climate field reconstruction methods targeting the Common Era. *Geophysical Research Letters*, 38(11):n/a–n/a, 2011. doi: 10.1029/2011GL047372. L11705.

J. Tipton, M. Hooten, N. Pederson, M. Tingley, and D. Bishop. Reconstruction of late holocene climate based on tree growth and mechanistic hierarchical models. *Environmetrics*, 27(1):42–54, 2015. doi: 10.1002/env.2368.

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

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.

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.

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 theoretical and practical applications.

Generation of target data begins with reformulating 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) (with $\mu = 0$) is the stochastic (ordinary) differential equation:

$$d\mathbf{T}_t = -\lambda \mathbf{T} dt + d\mathbf{W}_t, \tag{4}$$

where $\mathbf{T}_t = (T_{t,1}, \dots, T_{t,n})$, and $T_{t,i}$ is the temperature at time t and spatial position $\mathbf{x}_i$. The noise term $d\mathbf{W}_t = (dW_{t,1}, \dots, dW_{t,n})$ 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 on comprise of 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) and the multi-scale response kernel in Eq. (6) can be approximated by a power-law function to yield:

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

This expression describes the long-memory response to the noise forcing. 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. (7) in contrast to Eq. (5) is no longer a solution to an ordinary differential equation, but to a fractional differential equation. By neglecting the contribution from the noisy forcing prior to $t = 0$ 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

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

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 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 $\tau_0 = \beta/2$.

Summations over time steps $s = 1, 2, ...N$ of (9) results in the matrix 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. If we omit the spatial index i from Eq. (10), the model for the target temperature field **T** at time t can be written in the compressed form

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

Suggested revision of Sect.4.4 and inclusion of appendix B:

4.4 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). The two former skill metrics will be used in this study, but the CE metric is not a proper scoring rule and is therefore unsuitable for ensemble-based reconstructions in general (Gneiting and Raftery, 2007; Tipton et al., 2015), see also Appendix B. Instead, the continuous ranked probability score (CRPS) will be used (Gneiting and Raftery, 2007), the metric was used for probabilistic climate reconstructions in (Tipton et al., 2015; Werner and Tingley, 2015; Werner et al., 2018). Since the CRPS and its subcomponents are less well-known than the two former methods, we define these metrics in Appendix B and refer to Gneiting and Raftery (2007); Hersbach (2000) for further details.

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 point measures within a probability distribution P are are 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 $x_i < x < x_{i+1}$ refers 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 $\overline{CPRS}$ 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 distance 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.