Paper 3

How wrong are climate field reconstruction techniques in reconstructing a climate with long-range memory?

Manuscript in preparation, to be submitted to Climate of the Past.

How wrong are climate field reconstruction techniques in reconstructing a climate with long-range memory?

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

[1]Department of Mathematics and Statistics, University of Tromsø 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. Modern approaches to validate paleoclimatic reconstruction techniques often rely on GCM climate model data as the basis for pseudoproxy experiments. In this study, pseudoproxy experiments are performed using idealized input data, where ensembles of targets are generated from fields of 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 extremely simplistic compared with data extracted from GCM simulations, yet the essential component in focus is the prescribed spatiotemporal structure. Using the Bayesian BARCAST climate field reconstruction technique, ensemble-based temperature reconstructions are generated representing the European landmass for a millennial time period. Hypothesis testing in the spectral domain is used to investigate if the field and spatial mean reconstructions are consistent with either the fGn null hypothesis used for the target data, or the AR(1) 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 only for noise-free or weak noise scenarios. 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. Furthermore, there are local differences in scaling characteristics, while the spatial mean reconstructions are less distorted. However, none of the reconstructions were found to be 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 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 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 the associated confidence intervals, because of incorrect estimates of the reconstructed parameters.

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 a tem-

poral coverage of the proxy network varied 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 (2013); Luterbacher et al. (2016) present reconstructions of surface air temperatures (SAT) for different spatial and temporal domains. Due to apparent differences in the initial settings, the available reconstructions often tend to disagree on a number of crucial aspects such as timing, duration and amplitude of warm/cold periods. There are also alternative viewpoints on a more fundamental basis considering the level of high frequency versus low frequency variability, see e.g. Christiansen (2011); Tingley and Li (2012). Such discrepancies in the Fourier domain can occur among other things due to shortcomings of the reconstruction techniques, such as regression dilution. This describes variance losses back in time and bias of the target variable mean. These artifacts appear as a consequence of noisy measurements used as predictors in regression techniques based on ordinary least squares (Christiansen, 2011; Wang et al., 2014). 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 general circulation models (GCMs) first became available, and has been developed and applied over the last decade, (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 simulate real proxy networks and instrumental data. The target data representing the proxy period is then perturbed with noise to simulate real proxy data in a systematic manner, while the pseudo instrumental data are left unchanged or weakly perturbed with noise of magnitude typical for the real world instrumental data. The surrogate pseudoproxy and pseudoinstrumental 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 using a suite of metrics. The reconstruction skill is quantified through statistical metrics, both for a calibration- and a much longer validation interval.

The available pseudoproxy studies have to a great 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. Tingley and Huybers (2010a) use instrumental temperature data for North America and construct pseudoproxy data from some of the longest time series. Pseudoproxy experiments are then performed with the intention of testing the BARCAST climate field reconstruction technique and comparing it against the RegEM method used earlier by Mann et al. (2008, 2009). Werner and Tingley (2015) generated idealized target data based on the BARCAST model equations introduced in Sect. 2.1. In the present paper we extend the domain of pseudoproxy experiments. Instead of employing surrogate data from paleoclimate GCM simulations, ensembles of target fields are drawn from a field of 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 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 reconstructed data, compared with

surrogates with a prescribed spatiotemporal covariance structure. In addition 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 continuous ranked probability score (CRPS), which is a suitable skill metric for ensemble-based reconstructions in contrast to the often used coefficient of efficency (CE) and reduction of error (RE).

Temporal dependence in a stochastic process is described as persistence or memory, where a long-range memory (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 flat power spectrum. 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)) (Hasselmann, 1976). This process has a Lorentzian power spectrum 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). The exact strength of persistence varies between data sets and depends on the degree of spatial averaging, but in general $0 < \beta < 2$ is adequate (Fredriksen and Rypdal, 2016). 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).

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 scaling exponent $0<\beta<1$. The synthetic target data are designed as ensembles of LRM-processes in time, with an exponentially decaying 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 over land, typically $0 < \beta < 1$. 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). 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 mean 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 paleocli-

mate 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, 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 timeseries in (a) follows an fGn structure. The corresponding power spectrum is plotted in blue in (c). Furthermore, using the incorrect null hypothesis we estimate the AR(1) parameters from the timeseries in (a) using Maximum Likelihood estimation. A realization of an AR(1) process with these parameters is plotted in (b), with the power spectrum shown in red in (c). 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. Earlier, the reconstruction skill of BARCAST was tested and compared against another climate field reconstruction technique known as canonical correlation analysis (CCA) in Werner et al. (2013). 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.

In the following, we describe the methodology of BARCAST and the target data generation in Sect. 2. The spectral estimator used for scaling analyses is also introduced here. Sect. 3 comprises an overview of the experiment setup and explains the hypothesis testing procedure. Results are presented in Sect. 4 after performing hypothesis testing of scaling properties in the local and spatial mean reconstructions. The skill metric results are also summarized. Finally, Sect. 5 discuss 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 notion of a hierarchical model implies that there are multiple parameters that are related through the structure of the problem (Gelman et al., 2003, Chapter 5). 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:

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

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 positions $\mathbf{x}_i$ 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 10 year average of Siberian temperature station data is >5000 km. On the other hand, Tingley and Huybers (2010a) estimate a decorrelation length of 1800 km for global land data. They further use 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 area. 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.

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 of the proxy records relative to the temperatures. Note that these two parameters have no relation to the spectral persistence 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. This means the posteriors are dominated by the information in the observations and not the priors. The prior distributions are themselves given by parametrized distributions. These parameters are denoted as hyperparameters in the framework of a Bayesian hierarchical model. The priors for all parameters except ϕ are conditionally conjugate, meaning the prior and the posterior distribution has the same parametric form. Table A1 sums up the prior distributions and the choice of hyperparameters for the eight scalar parameters in BARCAST.

Likelihood functions for the observations given the true field values and all parameters $\boldsymbol{\Theta}$ are formulated:

$$P(\mathbf{W}_1, \mathbf{W}_2, ...\mathbf{W}_k | \mathbf{T}_1, \mathbf{T}_2, ...\mathbf{T}_k, \boldsymbol{\Theta}) = \prod_{n=1}^{k} P(\boldsymbol{W}_n | \boldsymbol{T}_n, \Theta) \tag{4}$$

Where k is the number of years for which there are any observations. Bayes' rule is applied and the posterior distribution is formulated as:

$$P(\mathbf{T}_0, \mathbf{T}_1, ...\mathbf{T}_k, \mathbf{\Theta}|\mathbf{W}_1, \mathbf{W}_2, ...\mathbf{W}_k) \propto P(\mathbf{T}_0) \cdot P(\mathbf{\Theta}) \cdot \prod_{n=1}^{k} P(\mathbf{W}_n|\mathbf{T}_n, \tau_I^2, \tau_P^2, \mu, \beta_0, \beta_1) P(\mathbf{T}_n|\mathbf{T}_{n-1}, \sigma^2, \phi, \alpha) \quad (5)$$

The hierarchical structure and the complexity of BARCAST makes the posterior simulation more complicated than it is for simpler Bayesian models. For all $\mathbf{T}_n$ and all parameters except ϕ it is possible to draw directly from the posterior distribution. The samples are drawn using the Markov-Chain Monte Carlo algorithm known as the Gibbs sampler. This is an iterative algorithm where values of $\mathbf{T}$ and $\mathbf{\Theta}$ are drawn from approximate distributions, and convergence to the correct distributions is reached after a sufficient number of sequential draws. For the total of 8 parameters, the value of parameter θ_{j} at each iteration is sampled from the conditional distribution given the current values of all the other parameters, temperature estimates and temperature observations:

$p(\theta_j|\theta_1, ...\theta_{j-1}, ...\theta_{j+1}, ...\theta_7, \mathbf{T}_0, ...\mathbf{T}_k, \mathbf{W}_1, ...\mathbf{W}_k)$

The parameter ϕ represent the spatial correlation length, and due to the exponentially decreasing structure it is chosen to use the lognormal prior distribution for this parameter. This prior distribution is not conditionally conjugate, and therefore the Gibbs sampler cannot be used for sampling. Instead, ϕ is updated using a single Metropolis step, see details in Gelman et al. (2003, chapter 11). The Metropolis algorithm involves drawing an initial value θ^1 from the prior/starting distribution, and then in the next iteration draw a proposed value θ^* from the so-called jumping distribution. This distribution must be symmetric. For iteration t a rejection/acceptance rule is applied, where the ratio of the present and the former probability density is estimated:

$$r = \frac{p(\theta^*|W)}{p(\theta^{t-1}|W)} \quad (6)$$

If this ratio is higher than one, the new draw θ^* is accepted as θ^t. If not, the jumping distribution is adjusted to account for the weighted probability. In sum, the draw θ^t is given by the probability:

$$\theta^t = \begin{cases} \theta^* & \text{with probability min (r,1)} \\ \theta^{t-1} & \text{otherwise} \end{cases} \quad (7)$$

The Metropolis-coupled MCMC algorithm is run for 5000 iterations in our experiments, running three chains in parallel. By chains we refer to the number of parallel computations initiated from the same set of input data. In our study BARCAST is run on a desktop computer using three cores to generate three sets of reconstructions and three sets of parameters for one input data set. Each chain is assumed equally representative for the temperature reconstruction if the parameters converge. By convergence we refer to how well the draws of $\mathbf{T}$ and $\mathbf{\Theta}$ follow their respective posterior distributions. 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. It is also possible to study the pdf for each parameter after discarding the first

n arbitrary steps, see Tingley and Huybers (2010a); 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 $\hat{R}$, (Werner et al., 2013; 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 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 reconstruction is estimated as the mean of the 500 independent iterations, the values of the mean parameters are estimated the same way.

The output temperature field is reconstructed also in grid cells without observations, which 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 time scales and spatiotemporal processes. BARCAST is so far only configured to deal with continental land data, (Tingley and Huybers, 2010a). The version of BARCAST used here is updated as described in Werner and Tingley (2015). Notably, the updated version generates different parameter values for τ_P^2, β_0 and β_1 for each proxy record, to be compatible with the latest multiproxy datasets used in paleoreconstructions (Werner et al., 2017; PAGES 2k Consortium, 2017).

2.2 Target data generation

While generating ensembles of synthetic LRM processes in time is straightforward using statistical software packages such as R or Mathematica, 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 has 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 ordinary differential equation:

$$\frac{d\mathbf{T}}{dt} = -(1-\alpha)\mathbf{I}\mathbf{T} + \boldsymbol{\epsilon}_t, \qquad \text{Where } \mathbf{I} \text{ Is the identity matrix.} \tag{8}$$

with the solution:

$$\mathbf{T}(t) = \int_0^t \exp^{-(1-\alpha)\mathbf{I}(t-s)} \boldsymbol{\epsilon}_s \mathrm{ds} \tag{9}$$

The exponential kernel is then replaced by a power-law function to yield:

$$\mathbf{T}(t) = \int_0^t (t-s)^{\beta/2-1} \boldsymbol{\epsilon}_\mathrm{s} \mathrm{ds} \tag{10}$$

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 **T**(t) in (10) in contrast to (9) is no longer a solution to an ordinary differential equation, but rather a fractional differential equation, whose solution for $t > 0$ depends not only on the initial condition but the entire time history of **T**(t), $t \in (-\infty, 0)$. Eq. 10 effectively corresponds to neglecting the contribution from the noisy forcing prior to $t = 0$.

In discrete form, the convolution integral in Eq. 10 is approximated over an index s:

$$\sum_{s=0}^{t} (t - s + \tau_0)^{\beta/2-1} \epsilon_\mathbf{s} \tag{11}$$

Note that here 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 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 $\tau = t - s$ of (11) results in the matrix $\mathbf{G}$ with terms:

$$\mathbf{G}(\tau) = (\tau + \beta/2)^{(\beta/2-1)} \mathrm{H}(\tau) \tag{12}$$

Where H(τ) is the unit step function:

$$\mathrm{H}(\tau) = \begin{cases} 0 & \tau < 0 \\ 1 & \tau \geq 0 \end{cases}$$

$\boldsymbol{\epsilon}_t$ is kept identical as in eq. (1) and (2). The target temperature field $\mathbf{T}$ at time t can be calculated as:

$$\mathbf{T}_t = \mathbf{G}_t \boldsymbol{\epsilon}_t \tag{13}$$

Pseudoinstrumental data are given by:

$$\mathbf{W}_{I,t} = \mathbf{T}_t + \mathbf{e}_{I,t}, \qquad \text{Where the noise terms } \mathbf{e}_{I,t} \text{ are IID normal draws } \mathbf{e}_{I,t} \sim N(0, \tau_I^2 \mathbf{I}) \tag{14}$$

While for the pseudoproxy data we have:

$$\mathbf{W}_{P,t} = \beta_1 \mathbf{T}_t + \beta_0 \mathbf{1} + \mathbf{e}_{P,t}, \qquad \text{With IID normal draw noise terms } \mathbf{e}_{P,t} \sim N(0, \tau_P^2 \mathbf{I}) \tag{15}$$

The true parameter values used to generate our target data are listed in Table A2.

2.3 Estimation of power-spectral density

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. Scaling properties of target data, pseudoproxies and the reconstruction are compared and analyzed in the spectral domain using the periodogram as the estimator.

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 scaling exponent 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 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 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 $5°$x$5°$. 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. Luterbacher et al. (2016) present an SAT reconstruction of the spatial mean and field for this area back to 138 BCE, and 755 CE, respectively. The reconstruction period for the present pseudoproxy study is 1000 years, reflecting the last millennium. By construction the target fGn data are meant to be an analogue of the detrended SAT field and hence can be considered as representing unforced GCM 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 pseudoproxy data covers every fourth grid cell for the time period 1-1000. The temporal resolution for all types of data is annual.

Our set of experiments is summarized in Table 1 and comprises target data with three different strengths of persistence, $\beta = 0.55, 0.75, 0.95$. The pseudoproxies are perturbed with white noise and four different signal to noise ratios by standard deviation (SNR): 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 involves a tripling of the ensemble size for the reconstructions, generating 60 realizations for each β and SNR. A total of 720 reconstructions are generated corresponding to 240 input data sets.

3.1 Hypothesis testing

Hypothesis testing is used to determine which pseudoproxy/reconstructed data sets can be classified as fGn with the prescribed scaling parameter. The null hypothesis is that the data sets under study can be described using an fGn with the prescribed scaling parameter for the target data, $\beta_{\mathrm{target}} = 0.55, 0.75$ and 0.95 respectively. For testing we generate a Monte Carlo 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 Monte Carlo ensemble. The null hypothesis is rejected if the log-binned spectrum of the data is outside of the confidence range for the fGn model at any point. If the null hypothesis is rejected for the reconstructed data we formulate a new hypothesis that the data can be described as an AR(1) process with the parameters (α, σ^2) estimated from BARCAST. The 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 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 of 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 all the data used to generate spectra are standardized by subtracting the mean and normalizing by the standard deviation. A different normalization could potentially be used, which would shift the spectra horizontally. If the reconstructed/target data are not perfectly scaling, the normalization of the confidence range spectrum is particularly relevant. For instance, the spectrum of the confidence range could be normalized to have the same power as the reconstructed/target data only for fre-

quencies lower than a certain threshold. Our null hypotheses do not require that the data are consistent with the fGn/AR(1) model specified for a certain spectral power, only that they are consistent with an fGn/AR(1) model with given parameters. Our hypothesis testing results below are based on a uniform, subjective choice of normalization, but additional testing using other normalizations could also be included.

4 Results

BARCAST successfully estimates posterior distributions for all reconstructed temperature fields and the scalar parameters. Convergence is reached for the scalar parameters despite a substantial inconsistency of the input data temporal covariance structure with the default assumption of BARCAST. Table A2 lists the true parameter values used for the target data generation, and Tab. A3 summarizes the mean or 95% confidence range of the posterior distributions estimated from BARCAST. For the parameters $\alpha, \phi, \mu\sigma^2, \tau_I^2$ the 95% confidence ranges for the posterior distributions are very narrow. Tab. A3 therefore summarizes the mean values of the respective posteriors. The posterior distributions of α and σ^2 depend on the prescribed β and SNR for the target data. The mean values of the distributions were used to generate Monte Carlo ensembles of AR(1) processes used for hypothesis testing. The estimates of the parameters μ, ϕ, τ_I^2 and β_0 are all stable with increasing β and SNR, and resemble the true target variables satisfactorily. τ_P^2 does not follow the expected dependence on the SNR. For the remaining parameter β_1, it decreases with decreasing SNR.

For the remaining parameters τ_P^2, β_0 and β_1 the distributions are wider, and 95% confidence ranges are therefore provided. 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 following correspond to mean spectra. Analyses of the reconstruction skills presented below are performed on a grid point basis as well as 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 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. Figure 3(a) shows the spectrum for SNR$= \infty$, which is the unperturbed fGn signal corresponding to ideal proxies. Panels 3b, c and 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. Panels 3b, c and d also show that the spectral slope of the pseudoproxy data is reduced compared to the ensemble mean on all 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 $\beta_{\mathrm{target}} = 0.55$ and 0.95 are the same.

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: is the reconstruction consistent with the fGn structure in the target data?
2: is the reconstruction consistent with the AR(1) model used in BARCAST?

Figure 3 shows the mean power spectra generated for one arbitrary proxy grid cell of the reconstruction in red. The fGn model is adequate only for the ideal proxy (SNR=∞) in panel 3(a), while for the lower SNR presented in panels b-d the reconstruction spectrum falls outside the confidence range of the theoretical spectrum in the higher frequency range. 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. All reconstructions tend to overestimate the high-frequency variability compared with the fGn Monte Carlo ensemble.

The hypothesis testing results vary moderately between the individual grid cells. PSD analyses of the reconstructions in the arbitrary non-proxy location displayed in Fig. 4 suggest better performance of BARCAST to preserve LRM properties of the target data between proxies than directly at proxy locations. The high-frequency end of the spectra is consistent with the fGn null hypothesis for $\beta = 0.75$, SNR=3.

Table 2 summarizes the results for all experiment configurations. At proxy sites, the null hypothesis 1 is rejected for all values of β, SNR=3, 1 and 0.3. Between proxies, the reconstructions are consistent with the fGn hypothesis for all values of β, SNR=∞ and 3, and also for $\beta_{\mathrm{target}} = 0.55$, SNR=1.

The BARCAST algorithm attempts to use the information from the input data to generate a reconstruction following an AR(1) model. We further test the null hypothesis 2 for consistency of the reconstructions with the AR1 model, (figure not shown). Distributions for all scalar parameters including the AR(1) parameter α and the variance σ^2 were estimated through the BARCAST algorithm. The mean of these two parameters were used to generate a Monte Carlo ensemble of AR(1) processes. Table 2 demonstrates that the null hypothesis was rejected for all of the numerical experiments considered.

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. Figure 5 shows the raw and log-binned periodogram of the spatial mean reconstruction for $\beta_{target} = 0.75$ in red, together with the 95% confidence range of fGn generated with $\beta = 0.75$. Hypothesis testing results are summarized in Table 3. Results show that the fGn null hypothesis is suitable for all values of β, SNR$= \infty, 3$ and 1, but not for SNR=0.3. Same as for the field reconstruction, the AR(1) null hypothesis is rejected for all data.

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, the root-mean squeared error (RMSE), and the coefficient of efficiency (CE) (Smerdon et al., 2011; Wang et al., 2014). However, the CE metric is improper for reconstructions based on the Bayesian framework (Werner et al., 2017), and will not be used. Instead we will use the continuous ranked probability score (CRPS), (Gneiting and Raftery, 2007), which was earlier used in (Werner and Tingley, 2015; Werner et al., 2017). The validation metrics are summarized below. 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 following metrics, in addition to the Pearson's correlation coefficient (r):

Root mean squared error (RMSE): Continuous ranked probability score (CRPS):

$$\mathrm{RMSE} = \sqrt{\frac{\sum_{i=1}^{N}(\hat{y}_{iv} - y_{iv})^2}{n}} \qquad \mathrm{CRPS}(\{F_t\}_{t=1}^{N}, \hat{y_v}) = -\sum_{t=1}^{N}\int_{-\infty}^{\infty}(F_t(y) - \mathbf{I}_{\{y_v \geq \hat{y_v}, t\}})^2 dy$$

Where N is the length of the timeseries. y_i is the target value, $\hat{y}_i$ is the reconstructed value, and the subscript v denotes the verification period (year 1-849). F is the cumulative distribution function. $\mathbf{I}_{\{y_v \geq \hat{y_v}, t\}}$ is a binary matrix of zeros and ones where the ones indicate locations where the field target value is greater than the field reconstruction value.

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 will use the temporally averaged score metric, called the average CRPS potential, ($\overline{\mathrm{CRPS}}_{\mathrm{pot}}$). This metric is akin to the Mean Absolute Error of a deterministic forecast, see Gneiting and Raftery (2007). We will also use the Reliability score, which represent the validity of the confidence range (Hersbach, 2000). This metric tests whether for all cases in which a certain probability p was forecast, on average, the event occurred with that fraction p. A perfectly reliable system has Reliability=0.

4.4.1 Skill measure results

The figures 6, 7, 8 and 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 at proxy sites except in Fig. 9. Figure 6 shows the local correlation coefficient r between the target and the localized reconstruction for the verification period 1-1849. The correlation is highest for the ideal-proxy experiment in Fig.6a, and gradually decreases at all locations as the noise level rises in panels b-d. Fig. 7 shows the local RMSE. Note that Fig. 7 use the same color bar as in Fig. 6, but best skill is achieved where the RMSE is low. Fig. 8 shows the distribution of $\overline{\mathrm{CRPS}}_{\mathrm{pot}}$. The minimum estimate for the proxy locations in 8a is 0.15, which indicates a low error between the temporally averaged reconstruction and the target. For the remaining locations in 8a-d, the estimates are between 0.55-0.67. The temperature unit has not been given for our pseudoproxy reconstructions, but for real-world reconstructions the unit will typically be degrees Celsius ($^{\circ}$C). The Reliability shown in Fig.

9 is generally low if the proxy locations in 9a are neglected. Except for these grid cells, the local Reliability ranges between $1 * 10^{-}3$ to 0.32, with lowest (best skill) estimate for the lowest SNR. 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 9a is 0.93, which indicates poor reconstruction skill when the full confidence range is 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 full confidence range is not.

Table 4 summarizes the mean local skill values for all experiments and skill metrics. All skill values are positive, indicating that BARCAST is 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 insensitive 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 skill values of r and RMSE for the spatial mean reconstructions. The skill values are considerably higher than for the local field reconstructions. The CRPS scores have not been evaluated for the spatial mean reconstruction.

5 Discussion and conclusions

In this study we have tested the capability of BARCAST to preserve temporal long-range memory properties of reconstructed data. Pseudoproxy and pseudoinstrumental data were generated with a prescribed LRM temporal persistence and spatial covariance structure using a new method. The data were then used as input to the BARCAST reconstruction algorithm which by construction uses 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 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 an AR(1) process using the parameters estimated from BARCAST. We found that despite the default assumptions in BARCAST, none of the reconstructed spatiotemporal fields generated from noisy input data were consistent with the AR(1) model. Moreover, local reconstructed fields were found consistent with the fGn model for all experiments with ideal proxies, and also at locations between proxies for the experiments $\beta = 0.55$, 0.75, 0.95, SNR=3 and $\beta = 0.55$, SNR=1. For the local reconstructed data at proxy sites, both null hypotheses were rejected if the pseudoproxy data were too noisy.

From a first glance, the fact that the LRM properties are better preserved at local sites between proxies than at actual proxy sites is a counterintuitive result. Compared with the fGn target, more high-frequency variability is lost for the reconstruction at proxy sites than it is for locations between proxies. This feature is related to the effect of added proxy-noise of the input data, and noise cancellation at individual sites for the reconstructions. By construction BARCAST estimates the posterior distributions of the proxy error variance τ_P^2, which are related to the signal to noise ratio. Tab. A3 shows that the estimated mean values of τ_P^2 are underestimated compared to the true parameter values in Tab. A2, and that the bias is stronger for lower

SNR. When BARCAST estimates low values for τ_P^2 it implies that the proxy noise terms have a small variance, hence there is a positive bias in the SNR assumed by BARCAST. The reconstruction is designed to follow the model equations 1 and 2 of the true temperature field. The assumed proxy noise is therefore eliminated for the data at proxy sites. Due to the interdependence of the BARCAST parameters, the underestimation of τ_P^2 is accompanied by an increase in the estimated AR(1) parameter α and a decrease of β_1 for noisy input data. These erroneous estimates influence the resulting reconstruction.

The power spectra in Fig. 3 and 4 show that the temporal covariance structure of the reconstructions are altered compared with the target data for all experiments where noisy input data were used. This effect 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 data to a model that is not necessarily correct. Another reconstruction technique that may experience similar deficiencies is the regularized expectation-maximization algorithm (Reg-EM), (Schneider, 2001; Mann et al., 2007), which assumes observations at subsequent years are independent (Tingley and Huybers, 2010b).

Our results suggest that the spatial mean reconstructions exhibit better scaling properties than local values. This is clear when comparing the spatial mean reconstruction spectra (red lines in Fig. 3 and Fig. 5 and from comparing the hypothesis testing results in Table 2 and 3. The improvement in scaling behavior is expected, as the small-scale variability denoted by ϵ_t in Eq. 13 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).

The skill metrics used to validate the reconstruction skill are the RMSE, r and CRPS, the latter divided into the $\overline{\mathrm{CRPS}}_{\mathrm{pot}}$ and the Reliability. We stress that even though the estimates of RMSE, correlation and $\overline{\mathrm{CRPS}}_{\mathrm{pot}}$ 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/gained in the reconstruction compared with the target. This fraction is essentially 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 bias of the variance: $\mathrm{R_{var}} = \frac{\mathrm{var_{rec}}}{\mathrm{var_{target}}}$. A ratio less than unity implies that the reconstruction has lost variance compared with the target, and represents a negative bias. Our analyses for the total variance reveal that the ratio varies between 0.65-1.0 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~ 0.95 for all β, SNR$=\infty$ and progressively decreases to R=0.65, 0.74 and 0.84 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, and an expected increase in variance bias with increasing noise levels. We also find that the negative bias is frequency-dependent by dividing the frequency range into three sections as shown in Fig. 10. The sections

separate low frequencies corresponding approximately to centennial time scales, mid frequencies corresponding to time scales between decades and centuries, and high frequencies corresponding to time scales shorter than decadal. The results show that the negative variance bias is most pronounced at high- and low frequencies. The mid-frequency range contain a bias which is similar to the total mean bias or even larger.

Previously, the scaling properties of millennium-long paleoclimate reconstructions have been studied in e.g. Lovejoy and Schertzer (2012); Nilsen et al. (2016). 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 in 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 time-series 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. The present study indirectly support the conclusions of Nilsen et al. (2016), that the two-regime model is redundant for the Holocene period when a simpler model is available using only one scaling regime.

The spectral shape of the input pseudo proxy 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; Werner et al., 2017). 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. (2017) 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 scaling properties of the input proxy records and the reconstructed temperatures were investigated using the same spectral techniques as here. Fig. A4 in Werner et al. (2017) 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 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 Fig. 3. 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). The tree ring width is a superposition of the age-

dependent curve, which is individual for a tree, and a climatic signal. 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); Helama et al. (2017)). This technique attempts to eliminate biological age effects on tree-growth while preserving low frequency variability. As an example, consider tree-ring width as a function of age. 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 curve. In particular, there may be slightly different climate processes affecting the growth of different trees, causing localized nonlinearities that are not well represented by the negative exponential growth curve. Hence, the negative exponential curve may be representative for some trees, but not all. To conclude there is probably some trade-off between the age-related error in width growth and the full proxy noise. 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.

A natural continuation of the pseudoproxy study presented here would be to generate target data using a more complicated model. The stochastic-diffusive models described in North et al. (2011); Rypdal et al. (2015) makes interesting candidates because of the alternative method for generating spatial covariance. The reconstruction technique used in this paper generates a signal without spatial dynamics, where the spatial covariance is defined through the noise term. On the other hand, the stochastic-diffusive models generates 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 experiments where 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.

An extension proposed for BARCAST already in Tingley and Huybers (2010a) was to generalize the spatial covariance structure using the Matérn covariance function. This would make the assumptions of BARCAST more realistic with respect to teleconnections. 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. Wang et al. (2014) later performed pseudo proxy experiments using four different climate field reconstruction 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). None of the reconstruction techniques were found to outperform the others, they have individual strengths

and drawbacks. 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 this study.

The pseudoproxy study presented here is based on simplistic target data generated using a novel technique. 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 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.

6 Appendix A: 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.A1.

The parameter values prescribed for the target data are listed in Tab.A2. 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: $\mathrm{SNR} = \frac{1}{\tau_P^2}$

The mean of the posterior distributions of the BARCAST parameters $\alpha, \mu, \sigma^2, 1/\phi$ and τ_I^2 are listed in Tab.A3, together with the 95% confidence ranges for the remaining parameters τ_P^2, β_0 and β_1. The confidence range was calculated for these three latter parameters since BARCAST estimate the posterior distributions for each proxy observation.

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

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

D. V. D was partly supported by TromsøResearch 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.

References

Beran, J., Feng, Y., Ghosh, S., and Kulik, R.: Long-Memory Processes, Springer, 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.

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, http://dx.doi.org/10.1002/2016RG000521, 2016RG000521, 2017.

Christiansen, B., Schmith, T., and Thejll, P.: A Surrogate Ensemble Study of Climate Reconstruction Methods: Stochasticity and Robustness, Journal of Climate, 22, 951–976, doi:10.1175/2008JCLI2301.1, 2009.

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.

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.

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

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.

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, 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, 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, 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.

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, 24, 5850–5862, doi:10.1175/ 2011JCLI4199.1., 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.

PAGES 2k Consortium: A global multiproxy database for temperature reconstructions of the Common Era, Scientific Data, 4, 170 088 EP –, doi:10.1038/sdata.2017.88, http://dx.doi.org/10.1038/sdata.2017.88, 2017.

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, 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.

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.

Smerdon, J. E., Coats, S., and Ault, T. R.: Model-dependent spatial skill in pseudoproxy experiments testing climate field reconstruction methods for the Common Era, Climate Dynamics, 46, 1921–1942, doi:10.1007/s00382-015-2684-0, 2016.

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.

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.

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 two millennia: an overview of the last major climate anomalies, Climate of the Past Discussions, 2017, 1–43, doi:10.5194/cp-2017-29, 2017.

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.

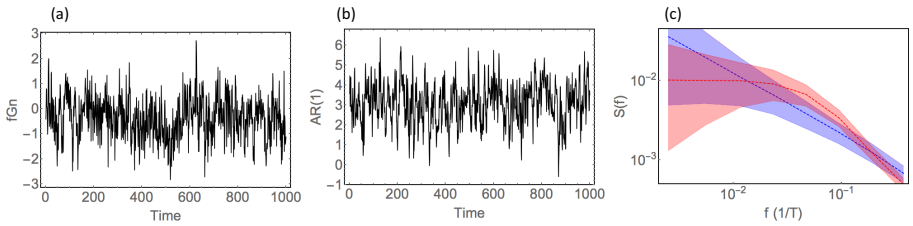

Figure 1. (a) Arbitrary fGn timeseries with $\beta = 0.75$. (b) Arbitrary timeseries of an AR(1) process with parameters estimated from the timeseries in (a) using Maximum Likelihood. (c) Log-log spectrum showing 95% confidence ranges based on Monte Carlo ensembles of fGn with $\beta = 0.75$ (blue shaded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

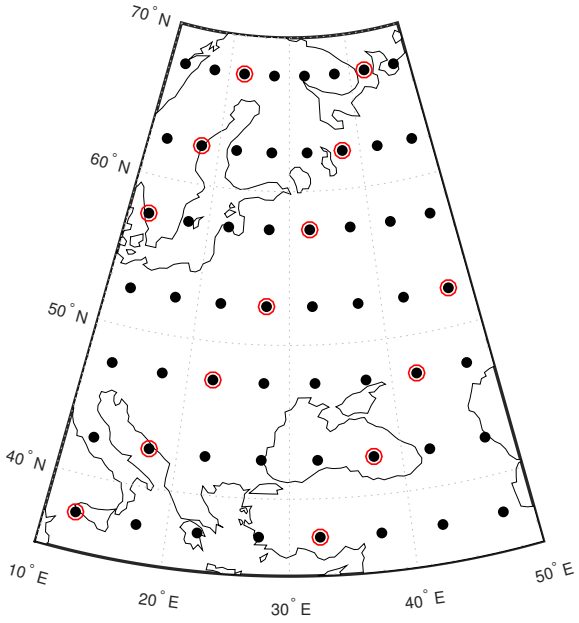

Figure 2. The spatial domain of the reconstruction experiments. Dots mark locations of instrumental sites, proxy sites are highlighted by red circles. The superimposed map of Europe provides a spatial scale.

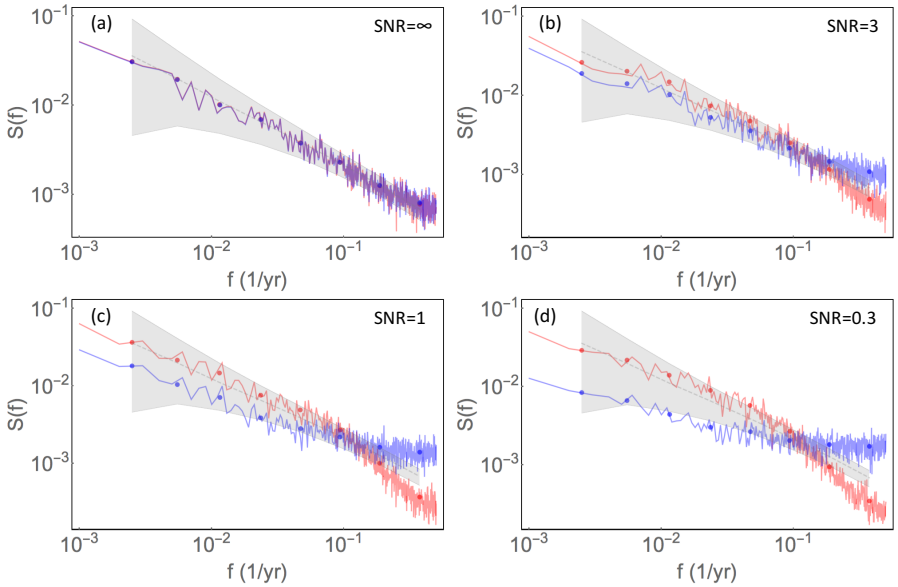

Figure 3. Mean raw and log-binned PSD for pseudoproxy data (blue curve and dots, respectively) and reconstruction at the same site (red curve and dots, respectively) generated from β_{target} =0.75 and different SNR. Colored gray shadings and dashed, gray lines indicate 95% confidence range and the ensemble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

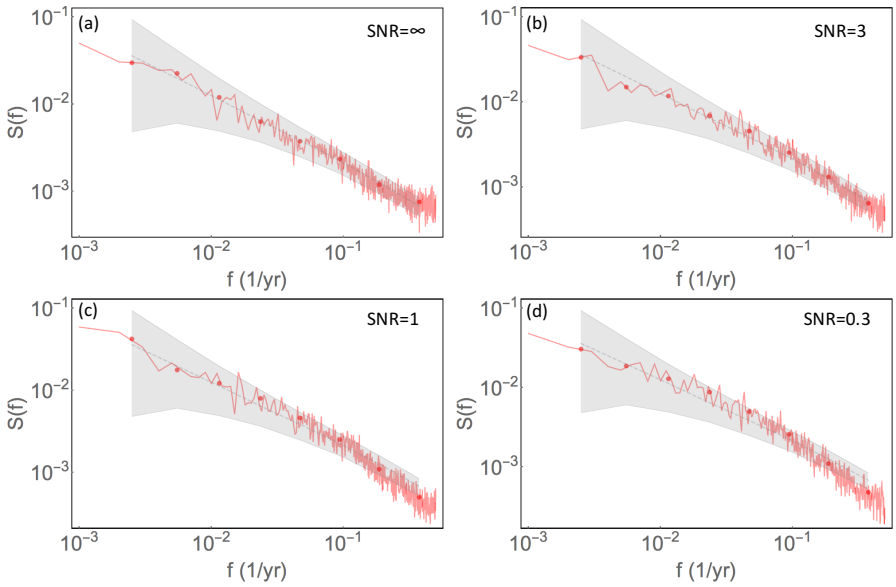

Figure 4. Mean raw and log-binned PSD for local reconstructed data at a site between proxies (red curve and dots, respectively) generated from β_{target} =0.75 and different SNR. Colored gray shadings and dashed, gray lines indicate 95% confidence range and the ensemble mean, respectively, for a Monte Carlo ensemble of fGn with β =0.75.

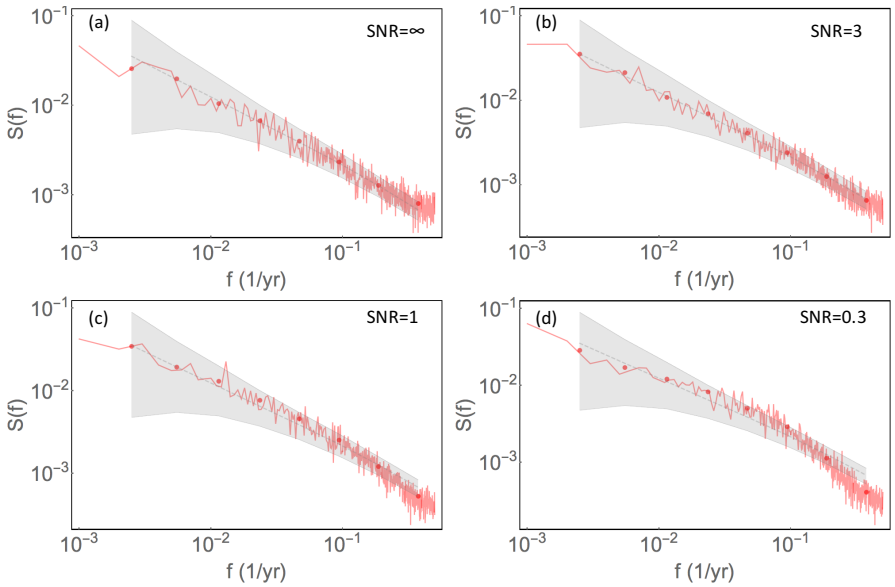

Figure 5. Mean raw and log-binned PSD for the spatial mean reconstruction (red curve and dots, respectively), generated from β_{target} =0.75 and different SNR. Colored gray shadings and dashed, gray lines indicate 95% confidence range and the ensemble mean, respectively, for a Monte Carlo ensemble of fGn with β =0.75.

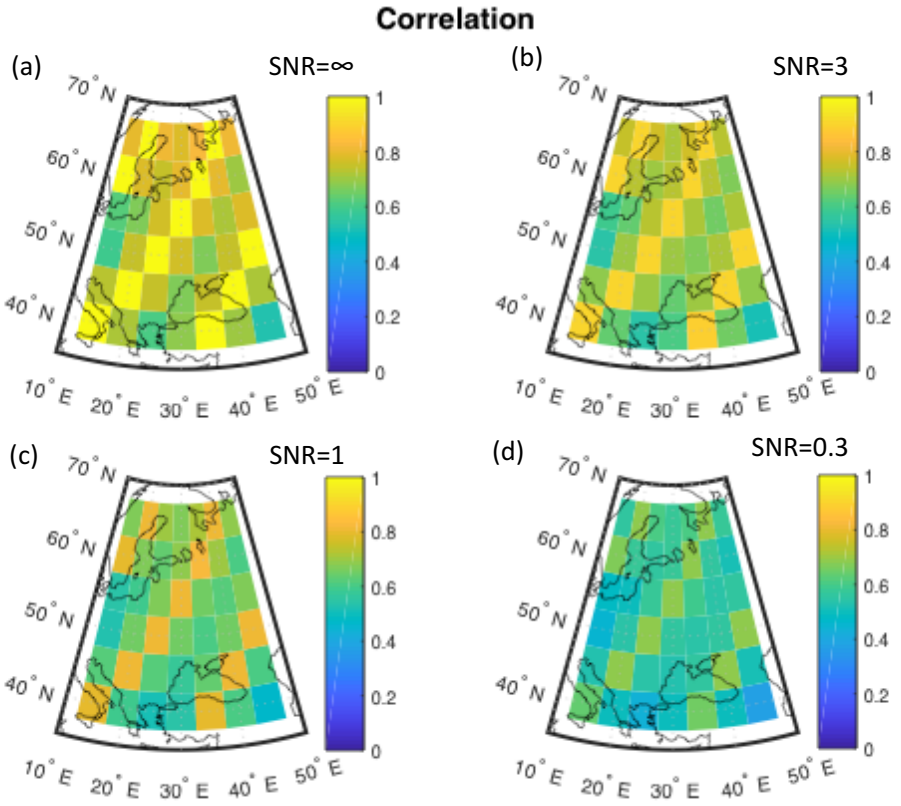

Figure 6. Mean local correlation coefficient between reconstructed temperature field and target field for the reconstruction period. β =0.75 and signal to noise ratios: (a)SNR=∞, (b) SNR=3, (c) SNR=1, (d) SNR=0.3.

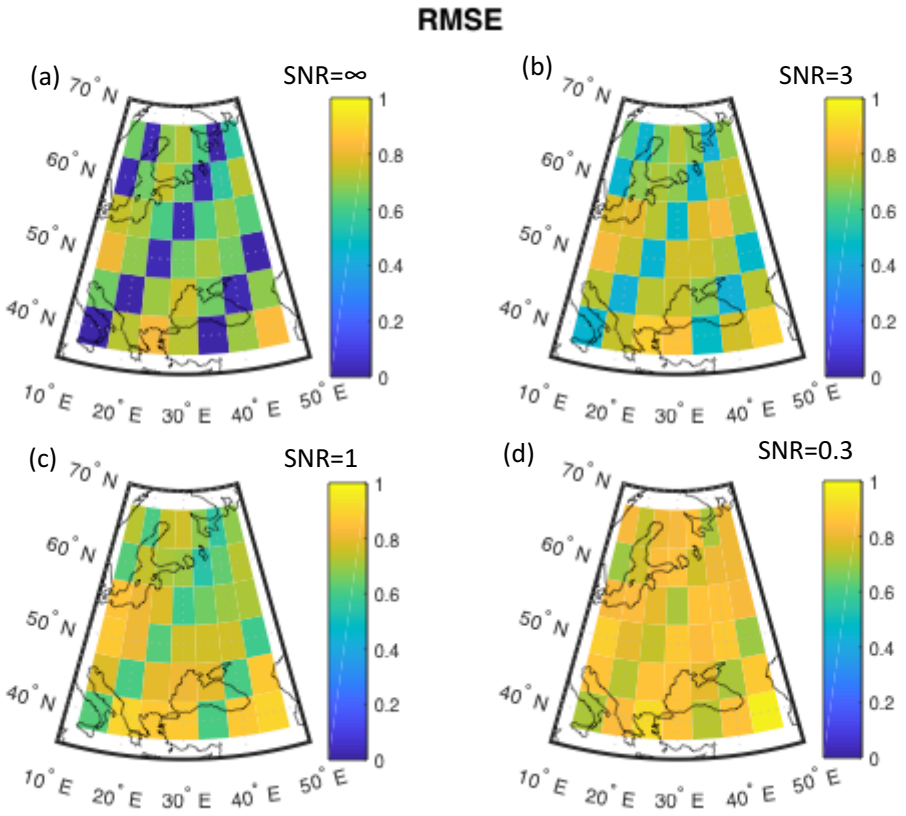

Figure 7. Mean local root-mean square error (RMSE) between reconstructed temperature field and target field for the reconstruction period. β =0.75 and signal to noise ratios: (a)SNR=∞, (b) SNR=3, (c) SNR=1, (d) SNR=0.3.

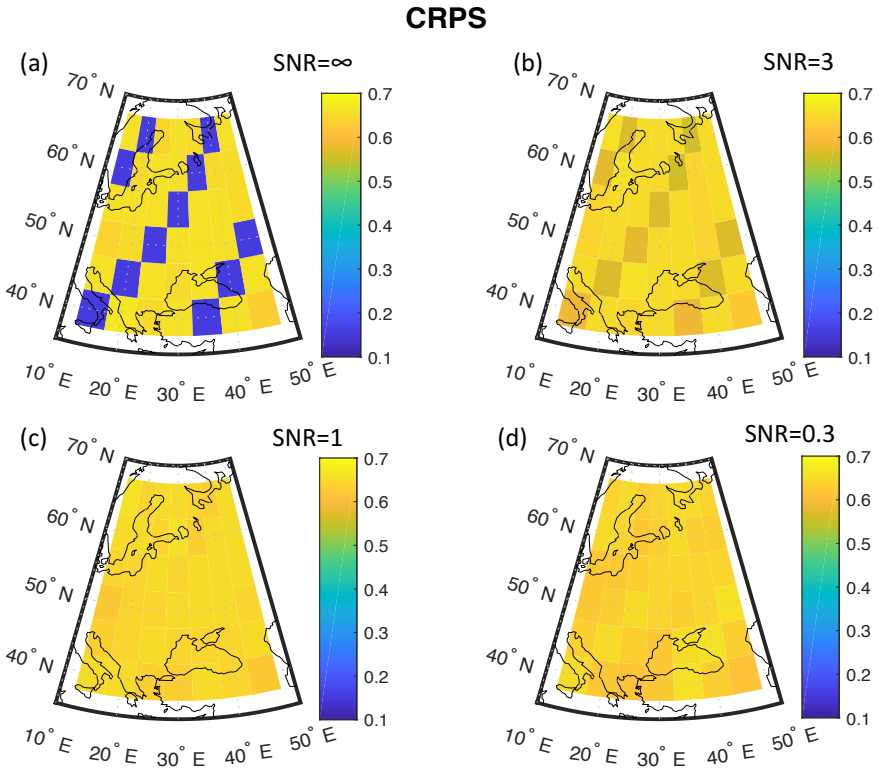

Figure 8. Mean local $\overline{\mathrm{CRPS}}_{\mathrm{pot}}$ between reconstructed temperature field and target field. β =0.75 and signal to noise ratios: (a)SNR=∞, (b) SNR=3, (c) SNR=1, (d) SNR=0.3.

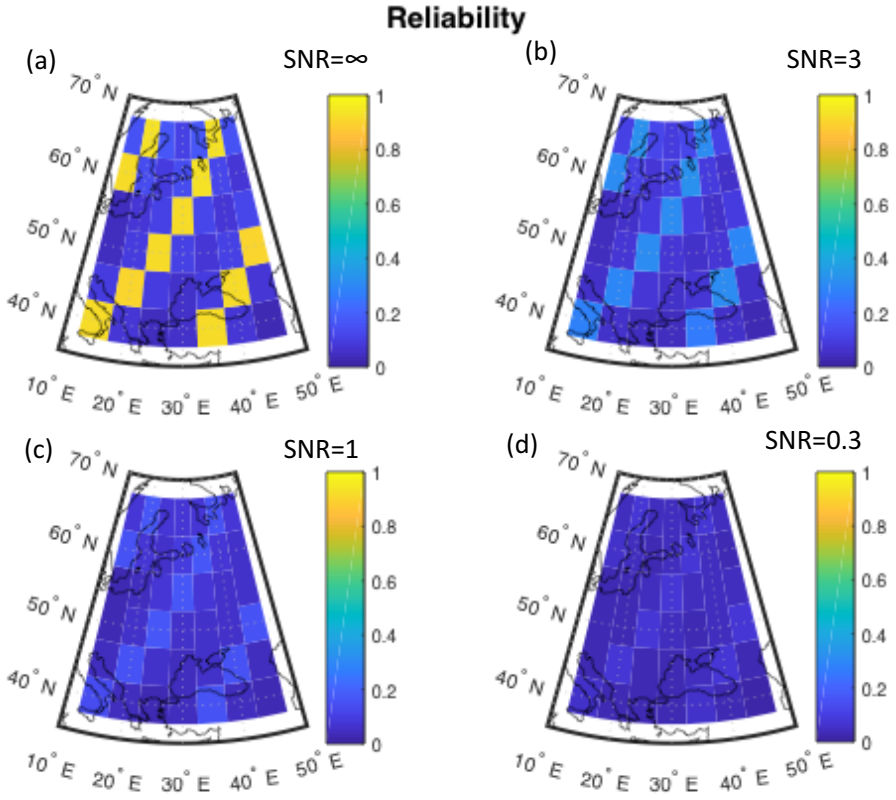

Figure 9. Mean local Reliability between reconstructed temperature field and target field. β =0.75 and signal to noise ratios: (a)SNR=∞, (b) SNR=3, (c) SNR=1, (d) SNR=0.3.

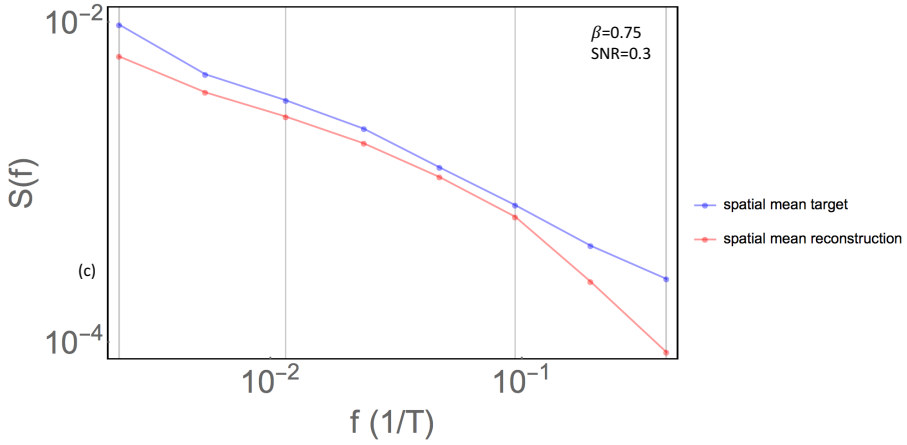

Figure 10. Log-log plot showing log-binned power spectra of spatial mean target (blue) and reconstruction (red) for the experiment $\beta = 0.75$, SNR=0.3 Vertical, gray lines mark the frequency ranges used to estimate bias of variance as referred to in Sec. 5.

Table 1. Summary of the experiment setup. 12 different input data experiments are run, using the strength of persistence and level of noise listed below. 20 ensemble members of the input data are generated for each experiment, and this number is tripled for the output reconstructions. A total of 240 input data sets are used, and 720 output reconstructions.

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	60
Total number of ensemble members	240	720

Table 2. Hypothesis testing results for local reconstructed data compared to Monte Carlo ensembles of fGn and AR(1) processes. The null hypotheses 1 and 2 listed to the left in the table are:

1: Is the reconstruction consistent with the fGn structure in the target data?

2: Is the reconstruction consistent with the AR(1) assumption from BARCAST?

Local field values				
SNR	∞	3	1	0.3
$\beta = 0.55$ Proxy site				
1	Yes	No	No	No
2:	No	No	No	No
$\beta = 0.55$ Between proxy sites				
1:	Yes	Yes	Yes	No
2:	No	No	No	No
$\beta = 0.75$ Proxy site				
1:	Yes	No	No	No
2:	No	No	No	No
$\beta = 0.75$ Between proxy sites				
1:	Yes	Yes	No	No
2:	No	No	No	No
$\beta = 0.95$ Proxy site				
1:	Yes	No	No	No
2:	No	No	No	No
$\beta = 0.95$ Between proxy sites				
1:	Yes	Yes	No	No
2:	No	No	No	No

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

Spatial mean values				
SNR	∞	3	1	0.3
$\beta = 0.55$				
Q1:	Yes	Yes	Yes	No
Q2:	No	No	No	No
$\beta = 0.75$				
Q1:	Yes	Yes	Yes	No
Q2:	No	No	No	No
$\beta = 0.95$				
Q1:	Yes	Yes	Yes	No
Q2:	No	No	No	No

Table 4. Mean local skill measures

SNR	r	RMSE	$\overline{\mathrm{CRPS}}_{\mathrm{pot}}$	Reliability
	$\beta = 0.55$			
∞	0.79	0.51	0.53	0.30
3	0.72	0.67	0.63	0.12
1	0.64	0.75	0.64	$6.6 * 10^{-2}$
0.3	0.53	0.85	0.62	$3.1 * 10^{-2}$
	$\beta = 0.75$			
∞	0.79	0.52	0.53	0.3
3	0.71	0.66	0.63	0.13
1	0.66	0.74	0.64	$7.6 * 10^{-2}$
0.3	0.56	0.82	0.63	$3.8 * 10^{-2}$
	$\beta = 0.95$			
∞	0.79	0.51	0.53	0.30
3	0.73	0.65	0.63	0.14
1	0.68	0.71	0.65	$9.2 * 10^{-2}$
0.3	0.6	0.79	0.64	$5.1 * 10^{-2}$

Table 5. Mean skill measures for spatial mean

SNR	r	RMSE
	$\beta = 0.55$	
∞	0.97	0.13
3	0.93	0.2
1	0.87	0.27
0.3	0.74	0.37
	$\beta = 0.75$	
∞	0.97	0.13
3	0.93	0.2
1	0.88	0.27
0.3	0.77	0.36
	$\beta = 0.95$	
∞	0.98	0.12
3	0.93	0.2
1	0.88	0.26
0.3	0.81	0.33

Table A1. List of parameters defined in BARCAST, form of prior and hyperparameters

Parameter	Form	Hyperparameters
α	Normal	N(α_μ, α_σ), $\alpha_\mu = 0.5$, $\alpha_\sigma = 0.1$
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($\beta_{1,\mu}$, $\beta_{1,\sigma}$), $\beta_{1,\mu} = 1$, $\beta_{1,\sigma} = 0.4$

Table A2. List of parameter values defined for the target data set. The four values of τ_P^2 listed correspond to the four different signal-to-noise ratios. 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, 0.333, 1, 3.33
β_0	0
β_1	1
β	0.5, 0.75, 0.95

Table A3. Mean (single value) or 95% confidence ranges (range within parantheses) for posterior distribution of parameters

Persistence	SNR	α	μ	σ^2	$1/\phi$	τ_I^2	τ_P^2	β_0	β_1
$\beta = 0.55$	∞	0.40	$-2.2 * 10^{-2}$	0.83	1020	$2.3 * 10^{-3}$	$(8.4 * 10^{-3}, 8.7 * 10^{-3})$	$(-7.9 * 10^{-4}, 1.9 * 10^{-3})$	1
	3	0.43	$-2.6 * 10^{-2}$	0.78	1053	$2.4 * 10^{-2}$	(0.23, 0.27)	(-7.4 $* 10^{-3}$, $1.2 * 10^{-2}$)	(0.88, 0.9)
	1	0.44	$-3.3 * 10^{-2}$	0.75	1064	$3.0 * 10^{-2}$	(0.48, 0.52)	$(-2.1 * 10^{-3}, 1.2 * 10^{-2})$	(0.73, 0.76)
	0.3	0.44	$-2.7 * 10^{-2}$	0.75	1053	$3.3 * 10^{-2}$	(0.72, 0.76)	$(-6.6 * 10^{-3}, 1.9 * 10^{-2})$	(0.53, 0. 56)
$\beta = 0.75$	∞	0.57	$-3.5 * 10^{-2}$	0.68	1020	$2.3 * 10^{-3}$	$(8.5 * 10^{-3}, 9.1 * 10^{-3})$	$(-2.9 * 10^{-3}, 1.3 * 10^{-3})$	1,
	3	0.62	$-4.5 * 10^{-2}$	0.61	1111	$2.8 * 10^{-2}$	(0.26, 0.28)	$(-1 * 10^{-2}, 1.8 * 10^{-2})$	(0.87, 0.9)
	1	0.64	$-5.6 * 10^{-2}$	0.59	1136	$3.3 * 10^{-}2$	(0.49, 0.53)	$(1.3 * 10^{-4}, 2.2 * 10^{-2})$	(0.71, 0.74)
	0.3	0.64	$-5.1 * 10^{-2}$	0.59	1136	$3.4 * 10^{-2}$	(0.73, 0.76)	$(-9.5 * 10^{-3}, 2.4 * 10^{-2})$	(0.51, 0.55)
$\beta = 0.95$	∞	0.71	$-4.8 * 10^{-2}$	0.5	1020	$2.4 * 10^{-3}$	$(8.6 * 10^{-3}, 9.1 * 10^{-3})$	$(-4.8 * 10^{-3}, 3.3 * 10^{-3})$	1
	3	0.77	$-8.5 * 10^{-2}$	0.44	1205	$2.8 * 10^{-2}$	(0.27, 0.29)	$(-8.1 * 10^{-3}, 1.7 * 10^{-2})$	(0.84, 0.88)
	1	0.79	$-9.7 * 10^{-2}$	0.41	1235	$3.1 * 10^{-2}$	(0.51, 0.53)	$(-3.7 * 10^{-3}, 2.6 * 10^{-2})$	(0.69, 0.73)
	0.3	0.77	$-1.0 * 10^{-1}$	0.42	1190	$2.9 * 10^{-2}$	(0.74, 0.76)	$(-1.3 * 10^{-4}, 3.1 * 10^{-2})$	(0.51, 0.54)