Clim. Past, 10, 1–19, 2014
www.clim-past.net/10/1/2014/
doi:10.5194/cp-10-1-2014

Climate of the Past
Open Access

Evaluating climate field reconstruction techniques using improved emulations of real-world conditions

J. Wang[1], J. Emile-Geay[1], D. Guillot[2], J. E. Smerdon[3], and B. Rajaratnam[2]

[1]University of Southern California, Los Angeles, California, USA
[2]Stanford University, Stanford, California, USA
[3]Lamont-Doherty Earth Observatory of Columbia University, Palisades, New York, USA

Correspondence to: J. Wang (jianghaw@usc.edu)

Received: 20 May 2013 – Published in Clim. Past Discuss.: 7 June 2013
Revised: 20 November 2013 – Accepted: 27 November 2013 – Published: 6 January 2014

Abstract. Pseudoproxy experiments (PPEs) have become an important framework for evaluating paleoclimate reconstruction methods. Most existing PPE studies assume constant proxy availability through time and uniform proxy quality across the pseudoproxy network. Real multiproxy networks are, however, marked by pronounced disparities in proxy quality, and a steep decline in proxy availability back in time, either of which may have large effects on reconstruction skill. A suite of PPEs constructed from a millennium-length general circulation model (GCM) simulation is thus designed to mimic these various real-world characteristics. The new pseudoproxy network is used to evaluate four climate field reconstruction (CFR) techniques: truncated total least squares embedded within the regularized EM (expectation-maximization) algorithm (RegEM-TTLS), the Mann et al. (2009) implementation of RegEM-TTLS (M09), canonical correlation analysis (CCA), and Gaussian graphical models embedded within RegEM (GraphEM). Each method's risk properties are also assessed via a 100-member noise ensemble.

Contrary to expectation, it is found that reconstruction skill does not vary monotonically with proxy availability, but also is a function of the type and amplitude of climate variability (forced events vs. internal variability). The use of realistic spatiotemporal pseudoproxy characteristics also exposes large inter-method differences. Despite the comparable fidelity in reconstructing the global mean temperature, spatial skill varies considerably between CFR techniques. Both GraphEM and CCA efficiently exploit teleconnections, and produce consistent reconstructions across the ensemble. RegEM-TTLS and M09 appear advantageous for reconstructions on highly noisy data, but are subject to larger stochastic variations across different realizations of pseudoproxy noise. Results collectively highlight the importance of designing realistic pseudoproxy networks and implementing multiple noise realizations of PPEs. The results also underscore the difficulty in finding the proper bias-variance tradeoff for jointly optimizing the spatial skill of CFRs and the fidelity of the global mean reconstructions.

1 Introduction

Over the past few decades, multiple methods have been proposed to estimate hemispheric and global temperature variability from proxy data over the Common Era (see Jones et al., 2009; Tingley et al., 2012, for comprehensive reviews). Such reconstructions provide an important test bed for understanding multidecadal to centennial climate variability and the climate sensitivity to exogenous forcing, while providing an extended context prior to the instrumental era for anthropogenic warming (Jansen et al., 2007). The majority of such reconstructions target an index (e.g., northern hemispheric mean temperature, Briffa et al., 2001; Crowley and Lowery, 2000; Mann and Jones, 2003; D'Arrigo et al., 2006), while a few are derived from climate field reconstruction (CFR) methods that aim to estimate the spatial, as well as the temporal aspects of large-scale temperature variability (Mann et al., 1998, 1999; Mann et al., 2009; Evans et al., 2002; Luterbacher et al., 2004; Rutherford et al., 2005; Tingley and Huybers, 2013).

A leading challenge in producing credible real-world climate reconstructions is the assessment of their uncertainties. The uncertainty of a real-world reconstruction is a mixture of two sources: the uncertainty associated with using necessarily imperfect proxy and target data, and the uncertainty associated with the employed statistical methodologies. *Data uncertainties* include measurement errors in the proxies, uncertainty in proxy-temperature relationships, sampling errors in instrumental climate fields, chronological uncertainties, and the uncertainty resulting from the network's coarse spatiotemporal coverage. *Methodological uncertainties* include a given method's sensitivity to input data (type of data, resolution, noise level, and spatiotemporal variability), its sensitivity to model parameters, and the uncertainty associated with the choice of these parameters.

Until recently, assessments of reconstruction uncertainties have primarily relied on cross-validation (CV, Cook et al., 1994), which consists of calibrating CFR methods over a subset of the instrumental period, and then validating the methods with the remaining observations. This method has the advantage of being firmly grounded in statistical theory (e.g., Hastie et al., 2008, Chap. 7) and it relies solely on actual observations; however, it was recently shown that shortening the calibration interval can lead to estimates of low-frequency skill that are biased low (Emile-Geay et al., 2013a). Temporal variations in reconstruction skill may be crudely estimated from "frozen network" experiments (Jones et al., 1998; Crowley and Lowery, 2000; Mann and Jones, 2003; Hegerl et al., 2006; Mann et al., 2007; Emile-Geay et al., 2013a), but because instrumental records are only available since the 1850s, it is impossible to directly estimate skill prior to the 19th century. Reconstruction uncertainty, particularly on multidecadal to centennial timescales, is thus difficult to quantify.

In this study, we use pseudoproxy experiments (PPEs) to extend our skill assessments of CFRs to decadal and centennial timescales and to isolate the impacts of the two principal uncertainty sources discussed above. PPEs were originally proposed by Bradley (1996) and adopted by Mann and Rutherford (2002) as a means of methodological assessment, and have been widely used to assess the performance of different CFRs in reconstructing global or hemispheric temperature (see Smerdon, 2012, and references therein for more details). Only a few of these PPEs, however, have focused on comprehensive assessments of CFR spatial skill (Tingley and Huybers, 2010b; Smerdon et al., 2011; Li and Smerdon, 2012; Annan and Hargreaves, 2012; Werner et al., 2013). In keeping with these earlier investigations, we focus herein on direct assessments of the spatial skill associated with leading CFR methods. Our approach nevertheless relies on more realistically designed pseudoproxy networks that give us better insights into the true spatial and temporal uncertainties in currently available CFR products.

Pseudoproxies typically are derived from the output of GCM simulations. The synthetic proxy data mimic some aspects of real-world proxy networks, and reconstruction algorithms are applied to the data to backcast the GCM-simulated climate conditions. Thus PPEs allow for controlled assessments of reconstruction methods with regard to the geographical and temporal distribution of proxies, their quality, and the spectral characteristics of the noise (Smerdon, 2012). However, most PPEs to date have constructed pseudoproxies that are temporally invariant throughout the reconstruction interval and have uniform proxy quality. Such networks under-represent the complexity of real-world proxy networks, limiting the direct applicability of their results to real-world reconstructions.

Here we construct more realistic pseudoproxy networks that mimic the key spatiotemporal characteristics of the multiproxy network used by Mann et al. (2008) (hereinafter M08). Two novelties in pseudoproxy design are introduced in this work: (1) the decrease in proxy availability over time follows that of the M08 network; and (2) the spatial variations of proxy quality mimic those found in M08. The more realistic pseudoproxy design allows a more stringent test on the performance of different CFR techniques, and provides insights into at least three aspects: (1) assessing how the spatiotemporal characteristics of the proxy network affects reconstruction skill, (2) tracing factors that contribute to the spatial variations of reconstruction skill, and (3) evaluating a method's ability to produce skillful index and field reconstructions. The four reconstruction techniques that we evaluate are (1) truncated total least squares regression embedded within the regularized expectation-maximization algorithm (Schneider, 2001, hereinafter RegEM-TTLS), (2) the Mann et al. (2009) implementation of RegEM-TTLS (hereinafter M09), (3) canonical correlation analysis (Smerdon et al., 2010, hereinafter CCA), and (4) Gaussian graphical models embedded within the EM algorithm (Guillot et al., 2013, hereinafter GraphEM). We first explore the spatiotemporal characteristics of M08 proxies in Sect. 2, and then describe the employed CFR techniques in Sect. 3. We present results in Sect. 4, followed by a discussion (Sect. 5) and a summary of our findings (Sect. 6).

2 Properties of real-world proxy networks

We consider the M08 proxy network as the basis for our pseudoproxy emulation of a real-world proxy network. The M08 proxy network has a relatively extensive spatial coverage over land, and most proxies have a temporal resolution of less than 10 yr. More importantly, the M08 network has recently been used to derive real-world CFRs (Mann et al., 2009), in which the authors reconstructed spatial patterns of surface temperature over the past 1500 yr and explored their associated dynamical causes. Out of the total 1209 proxies in the network, we exclude 71 European composite surface temperature reconstruction records (Luterbacher et al., 2004), so that only true natural proxies are used to as a basis for our

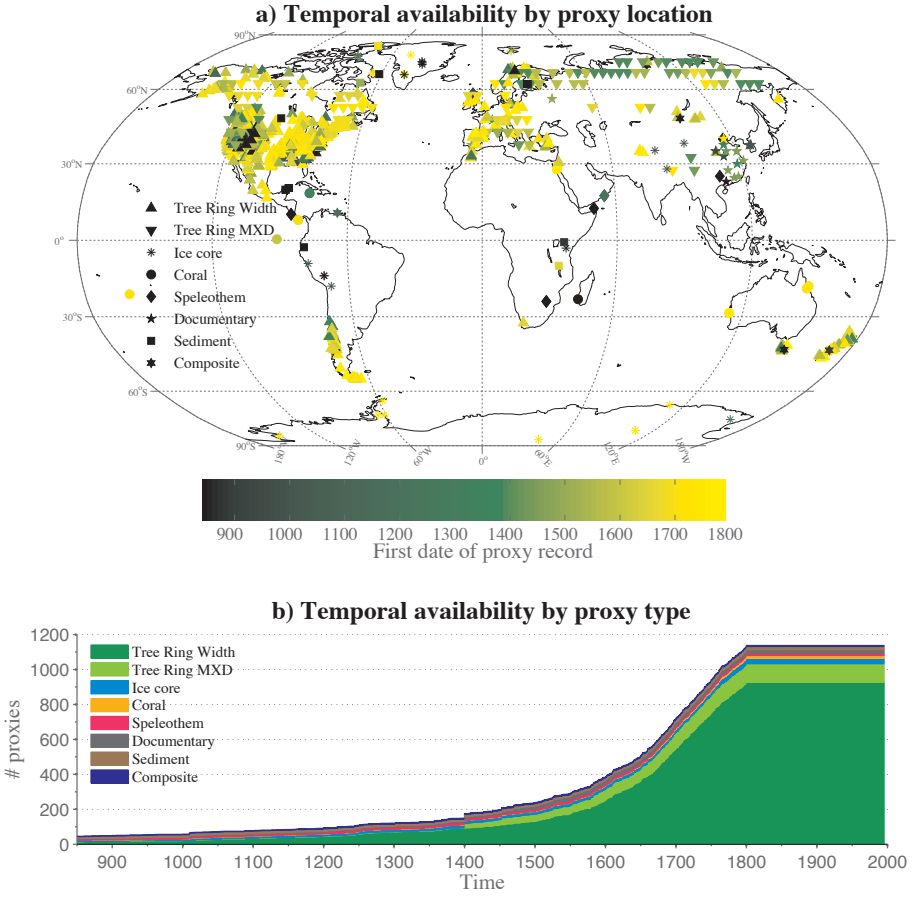

Fig. 1. Temporal and spatial availability of the M08 proxy network between 850–1800 AD. Top panel: the spatial distribution of the M08 proxies, with colored dots indicating the first year that each proxy record becomes available. Each marker represents a proxy class, as indicated in the legend. Bottom panel: the temporal availability of each proxy class.

emulation. Figure 1 shows the spatiotemporal distribution of the remaining 1138 proxies. Most proxies are concentrated in extra-tropical land regions of the Northern Hemisphere, particularly across North America and western Europe (Fig. 1a). Tree-ring width is the dominant proxy class, and fewer than 200 proxies in total are available prior to 1400 AD (Fig. 1b).

2.1 Spatial characteristics

Spatial relationships between proxies (P) and temperature (T) are explored by calculating the Pearson's correlation coefficient (ρ) between each proxy and the HadCRUT3v surface temperature field (Brohan et al., 2006, the temperature target used in the M08 and M09 studies). As in M08, temperature grid boxes less than 10 % complete were removed from the HadCRUT3v data set, and missing values were infilled with the RegEM algorithm using ridge regression (Schneider, 2001) during the 1850–2006 AD period. ρ is calculated between annually averaged HadCRUT3v temperature observations and annual proxy data[1] over the 1850–1995 AD period. The statistical significance of ρ is also taken into account, using a spectrum-preserving, nonparametric test (Ebisuzaki, 1997).

In Fig. 2, $|\rho|_{\max}$ is plotted as a function of P–T distance, where $|\rho|_{\max}(i) = \max_{j\in[1,p]} |\rho(P_i, T_j)|$ is the highest absolute value of the estimated correlation coefficients between the ith proxy and all temperature grid points. The total number of temperature grid cells is $p = 1732$, as in M08. All the temperature and proxy data can be downloaded at http://www.ncdc.noaa.gov/paleo/pubs/mann2008/mann2008.html.

Contrary to common assumptions (e.g., Jones and Mann, 2004), we find that $\rho(P, T)$ is not a monotonically decreasing function of distance. As in Fig. 2, the distribution of P–T distance is bimodal: one cluster of proxies is well correlated to local temperature (distance shorter than 2000 km, similar to findings in Hansen and Lebedeff, 1987), but the majority of proxies are at least 8000 km away from the temperature

[1] M08 interpolated non-annually resolved proxies to annual resolution.

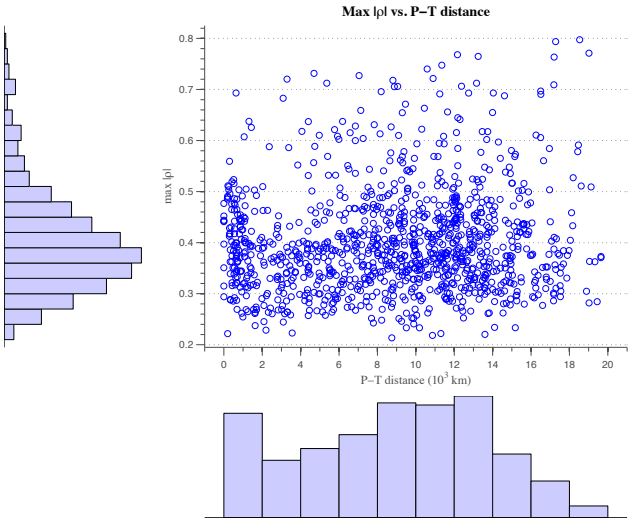

Fig. 2. Maximum absolute correlation coefficient $|\rho|$ between proxies of the M08 network and the HadCRUT3v grid point temperatures vs. the corresponding distance between the proxy location and the grid point. On the y axis is the histogram of the maximum $|\rho|$; on the x axis is the histogram of distance between each proxy P_i and the corresponding temperature grid T_j that gives the highest $|\rho|$.

point yielding the highest $|\rho|$. On the other hand, the distribution of $|\rho|_{\max}$ is unimodal and positively skewed. The distribution exhibits a mode near 0.4, while high values are quite rare (95 % of ρ values are below 0.76). The average $|\rho|_{\max}$ is 0.45, corresponding to a P–T distance of 11 000 km.

The counterintuitive pattern of Fig. 2 is a consequence of two effects: many proxies indeed are primarily sensitive to local temperature, but the probability of finding a spurious (non-physical) correlation also increases as the search radius increases. Some of the high non-local correlations may be reflective of long-range temperature dependencies (teleconnections, c.f. Liu and Alexander, 2007), such as precipitation proxies in the southwestern United States (e.g., Cook et al., 2004, 2007, see also Fig. S23, Supplement). Others may arise by chance alone. Since we lack a theoretical criterion to distinguish real teleconnections from spurious correlations, we constructed pseudoproxies representing two end-member possibilities: one corresponding to local temperature associations, and the other mimicking each proxy's highest potential to capture large-scale teleconnections. An alternative network design, balancing the two extreme scenarios, is explored in the supplementary information (SI, Sect. 3).

Traditionally, pseudoproxies $P(x,t)$ are generated according to

$$P(x,t) = T_{\mathrm{s}}(x,t) + \frac{1}{\mathrm{SNR}} \cdot \varepsilon(x,t), \qquad (1)$$

where T_{s} is a time-standardized[2] version of T. The primary data of T are grid cells extracted from GCM fields in a way that mimics instrumental data availability. $\varepsilon(x,t)$ are independent realizations of a Gaussian white-noise process with zero mean and unit variance, and the signal-to-noise ratio (SNR) controls the amount of noise in the pseudoproxies (Mann and Rutherford, 2002; von Storch et al., 2004; Mann et al., 2005, 2007; Rutherford et al., 2005; Küttel et al., 2007; Smerdon et al., 2008; Smerdon et al., 2011; Christiansen et al., 2009; Emile-Geay et al., 2013a). SNR is related to proxy-temperature correlations ρ via

$$\mathrm{SNR} = \frac{|\rho|}{\sqrt{1-\rho^2}}. \qquad (2)$$

(Mann et al., 2007). While most studies have heretofore considered spatially uniform SNRs, it is clear that $|\rho|$ is quite variable (Fig. 2), requiring pseudoproxy networks to contain such variability. In this study, spatial variations of SNRs and the bimodal pattern of Fig. 2 are explored via two end-members:

1. Local SNR: each proxy record is regressed onto temperature at the closest HadCRUT3v grid point over the 1850–1995 AD period, exposing each proxy's ability to record local temperature conditions (Fig. 3, top panel);

2. Max SNR: proxies are regressed onto all temperature points in the HadCRUT3v data set. The highest $|\rho|$ for each proxy is selected to construct $P(x,t)$ at that point (Fig. 3, bottom panel).

Temperature grid points selected in each design are then used to calculate SNR via Eq. (2). Their locations are also used to select grid cells from the simulated temperature field, which are assigned to T_{s} in Eq. (1). In addition, the statistical significance of ρ is also incorporated into the construction of the PPEs. Proxies showing spurious correlations are excluded based on the Ebisuzaki (1997) significance test. Only the ones with a significant relationship to annual temperature in the HadCRUT3v data set are retained. Based on the significance test, only 312 out of 1138 M08 proxies exhibit a significant correlation with local temperature, whereas 1121 proxies are significantly correlated to at least one temperature point on Earth. Pseudoproxies are therefore sampled only at these proxy sites. Unique temperature grid points being used in the local and max SNR networks reduce to 128 and 551, respectively, i.e., only locations of these temperature grids are used to sample T_{s} in Eq. (1). As illustrated in Fig. 3, even in the best-case scenario (max SNR), SNR is on average lower than 0.5, with fewer than 30 proxies exhibiting an SNR above 1.0. In the local SNR case, the mean SNR is even lower (0.27), close to the low end of SNRs usually considered in pseudoproxy studies (0.25).

[2] In this case, over the period 850–1995 AD.

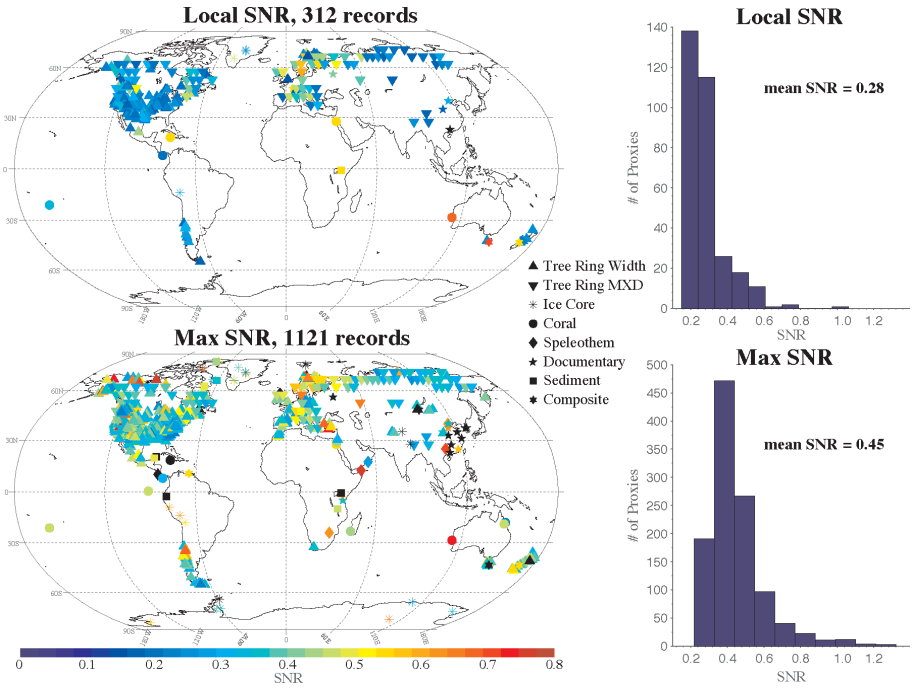

Fig. 3. Estimated signal-to-noise ratio (SNR) for proxies in the M08 network. Top panel: Local SNR scenario, in which SNR is calculated between each proxy and its closest temperature grid. Bottom panel: max SNR scenario, in which the highest SNR for each proxy is chosen from its correlations with all temperature grids available. Colors reflect the value of SNR assigned to each pseudoproxy, as per Eq. (2).

We emphasize that neither choice of SNR design is physically realistic – instead, each may only be viewed as an end-member experiment of real-world conditions. A middle-ground scenario, balancing locality with the ability to capture the largest correlations, is explored in the SI (Sect. 3). Results based on this intermediate SNR design are found to be very similar to the max SNR case, and thus are not shown in the manuscript.

Similar characteristics were also considered in PPEs by Christiansen et al. (2009), in which empirical SNRs and noise values were used to reflect the heterogeneous proxy quality in the Mann et al. (1998) proxy network. Their work, however, did not model the temporal heterogeneity in proxy availability, and the spatial skill evaluation was not their main focus. Our study here seeks to evaluate the impact of spatial heterogeneity in multiproxy networks and its impact on derived CFRs. Six networks are designed to address this problem, of which two model the spatial variation of SNRs in real-world proxies (local SNR and max SNR) and four have uniform SNRs. Following previous studies (Mann and Rutherford, 2002; von Storch et al., 2004; Mann et al., 2005, 2007; Rutherford et al., 2005; Küttel et al., 2007; Smerdon et al., 2008; Smerdon et al., 2011; Werner et al., 2013), the four networks of homogenous quality are designed by assigning constant SNRs (SNR = ∞, 1.0, 0.5, 0.25) in Eq. (1). These six networks together provide the basis for our experiments.

2.2 Temporal characteristics

Another realistic characteristic we incorporate into the PPE design is the temporal heterogeneity of proxy availability. As shown in Fig. 1b, data availability decreases steeply back in time, and a staircase pattern is evident for all proxy classes. In a similar manner, the effective SNR, which is the average SNR of all proxies available at a given time point, also declines back in time (Fig. 4). The pattern is consistent with properties of the M08 network: most of the high SNR proxies (such as tree rings and corals) are only available for several decades or centuries prior to widespread observational data. For instance, most tree-ring chronologies drop out of the network prior to the 16th century, with fewer than 100 (out of an original 1031) still available before the 14th century. Overall, only 47 proxies are available throughout the entire reconstruction period, of which 19 have decadal or lower resolution.

To isolate the impact of temporal availability, we specify two types of pseudoproxy networks:

1. M08 “flat” network: pseudoproxy availability is uniform through time.

2. M08 “staircase” network: pseudoproxy availability matches the pattern of the M08 database (Fig. 1b).

Contrasting these two cases will therefore characterize the impact of temporal heterogeneity on CFR performance.

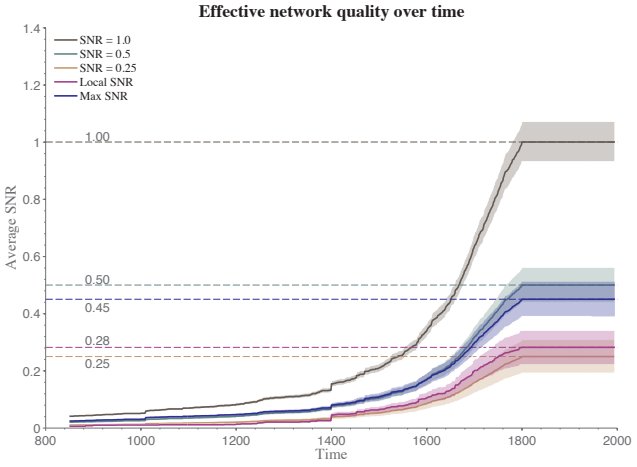

Fig. 4. Effective network quality expressed as SNR (Eq. 2) for idealized pseudoproxy networks and observed networks. For the observed SNRs, we consider two possible scenarios as described in Sect. 2.1. Shaded areas correspond to the 2.5–97.5 percentile interval, providing a complete picture of effective SNR across the 100-member noise ensemble. Dashed lines correspond to the values of SNR for temporally invariant networks. SNR = ∞ is not plotted.

2.3 Limitations

Despite the spatiotemporal characteristics that are modeled in the pseudoproxies, the networks are still idealized in various respects. The pseudoproxy networks do not model the temporal auto-correlation (persistence) present in real-world temperature and proxy data, nor do they consider the effect of using low-resolution data for reconstructions of annual temperature. All pseudoproxies are generated on an annual basis without regard for the proxies' actual resolution (as done in most other pseudoproxy studies). To more realistically model real-world proxies, future PPE designs should represent the actual resolution of each proxy (Christiansen, 2011).

Using Gaussian white noise for ε in Eq. (1) is a natural first step, but a more complex noise model could be used to better reflect real-world noisy proxies (Smerdon, 2012, and references therein), as done in Tingley and Huybers (2010a, b) and Christiansen and Ljungqvist (2012). Furthermore, mechanistic proxy models could be used to simulate synthetic proxy records with more realistic properties (Anchukaitis et al., 2006; Evans, 2007; Cobb et al., 2008; Thompson et al., 2011; Evans et al., 2013). Finally, the target field is assumed to be noise-free, yet in reality, gridded instrumental observations may contain substantial noise or interpolation errors, yielding a large influence on the derived calibrations and thus the reconstruction in the preinstrumental era (Emile-Geay et al., 2013a, b). While we explore herein the impacts of significant advancements in PPE design, incorporation of the above considerations will further improve the degree to which PPEs can be interpreted as representative of real-world CFR performance.

3 Methodology

3.1 CFR techniques

Two classes of statistical methods are commonly used to perform CFRs. One is based on multivariate linear regression models, where inference is performed in a frequentist framework (e.g., Mann et al., 1998; Mann et al., 2008, 2009; Schneider, 2001; Luterbacher et al., 2004; Smerdon et al., 2010; Guillot et al., 2013) and the other uses Bayesian hierarchical models (BHMs, e.g., Li et al., 2010; Tingley and Huybers, 2010a). We restrict our attention to frequentist regression-based methods since only those have heretofore been used to derive global/hemispheric CFRs.

Let P be an $n_\mathrm{p} \times p_\mathrm{p}$ matrix of proxy values and T be an $n_t \times p_t$ matrix of instrumental temperature records, where n_p and n_t are the number of years of available data (i.e., number of observations), p_p and p_t are the number of spatial locations (i.e., number of variables), and the subscripts p and t denote proxies and instrumental data, respectively. Traditional regression-based CFR methods assume a multivariate linear relationship between proxies and the climate variable of interest: e.g., temperature (Jones and Mann, 2004; National Research Council, 2006; Jones et al., 2009; Tingley et al., 2012). Additionally, each year is often treated as an independent observation. In this context, temperature may be estimated from the proxies via the regression equation:

$$T = B\,P + \varepsilon, \tag{3}$$

where ε is an error term following a multivariate normal distribution with zero mean. In the sample-rich setting familiar to classic regression problems (e.g., Anderson, 2003), the optimal estimate of B would be given by the ordinary least squares (OLS) estimate:

$$\widehat{B} = \left(\mathbf{P}_\mathrm{c}^\top \mathbf{P}_\mathrm{c}\right)^{-1} \mathbf{P}_\mathrm{c}^\top T, \tag{4}$$

where $\mathbf{P}_\mathrm{c}$ is the submatrix of P spanning the calibration interval. This formulation is such that in order to estimate $\widehat{B}$, $\mathbf{P}_\mathrm{c}^\top \mathbf{P}_\mathrm{c}$ must be invertible (non-singular). In paleoclimate applications, however, it is often the case that $\mathbf{P}_\mathrm{c}^\top \mathbf{P}_\mathrm{c}$ is rank-deficient, (i.e., not invertible): instrumental temperature records are only available for the past 150 yr or so ($n_t \approx 150$), and the number of proxies p_p is on the order of 10^3 (high dimension, low-sample size). In this setting, the OLS estimate is no longer optimal, and may be wildly erroneous. Some form of regularization is needed to make $P^\top P$ invertible, and thus solving Eq. (4) amounts to finding a regularized least squares solution (Hansen, 1998). Each regression-based CFR technique accomplishes this in a different manner. Our study focuses on four such techniques: RegEM-TTLS, M09, GraphEM and CCA.

3.1.1 RegEM

RegEM (Schneider, 2001) is a variant of the EM algorithm (Dempster et al., 1977; Little and Rubin, 2002) designed for the imputation of missing values in spatiotemporal data sets typically encountered in climatology. Under the multivariate normal assumption, given an initial estimate of the mean $\hat{\mu}$ and the covariance matrix $\widehat{\Sigma}$, the RegEM algorithm reduces to regressing the missing values onto the available ones (instrumental temperature data and overlapping proxies)[3]. The estimates of $\hat{\mu}$ and $\widehat{\Sigma}$ are updated at each iteration until convergence is achieved. Two regularization methods have been considered in RegEM: one is *ridge regression* (Tikhonov and Arsenin, 1977; Hoerl and Kennard, 1970a, b), used in the paleoclimate context by Mann and Jones (2003), Rutherford et al. (2003), Mann et al. (2005) and Rutherford et al. (2005); the other is *truncated total least squares* (Van Huffel and Vandewalle, 1991; Fierro et al., 1997, hereinafter TTLS), used in the paleoclimate context by Mann et al. (2007, 2008, 2009), Wilson et al. (2010) and Emile-Geay et al. (2013a, b). Rutherford et al. (2010), Christiansen et al. (2010) and Tingley et al. (2012) have also discussed the relative merits of these methods. As explained in Smerdon and Kaplan (2007), Mann et al. (2007) and Smerdon et al. (2008), ridge regression can lead to overly damped reconstructions in very data-sparse scenarios such as paleoclimate reconstructions. TTLS mitigates such variance losses, and therefore is chosen as our regularization method for this study. We employ two different styles of RegEM-TTLS: one following the standard formulation described in Schneider (2001), and the other following its paleoclimate-specific implementation described in Mann et al. (2009).

3.1.2 M09 implementation of RegEM-TTLS

The M09 implementation of TTLS uses a hybrid version of RegEM-TTLS (Mann et al., 2007) that treats low-frequency and high-frequency signals separately (the domains are split at a 20 yr period). The reconstruction is then performed in a forward stepwise approach century by century. For each step, the target data are compressed in that only the first M leading modes of surface temperature are retained, where M is an estimate of the number of degrees of freedom in the proxy network, determined by a fit to its log-eigenvalue spectrum. Additionally, semi-adaptive choices are adopted for both low-frequency k_{l} and high-frequency truncation parameters k_{h}. The method selects (1) k_{l} to retain 33 % of the low-frequency multivariate data variance (Mann et al., 2009; Rutherford et al., 2010); and (2) k_{h} by detecting the first break in the log-eigenvalue spectrum of high-frequency multivariate data variance.

Although these ad hoc modifications are not grounded in statistical theory, the M09 implementation of TTLS has proven effective in practice (Mann et al., 2009; Emile-Geay et al., 2013a, b). Indeed, the M09 implementation is the only technique that has ever been used for global-scale, real-world CFRs, so it is taken as the benchmark of our study. By comparing M09 to TTLS with fixed regularization (i.e., RegEM-TTLS), we investigate the merits of the M09 approach to parameter selection in an ensemble framework, which is a novel assessment of this heuristic approach.

3.1.3 CCA

CCA (Christiansen et al., 2009; Smerdon et al., 2010) is based on ideas presented in Barnett and Preisendorfer (1987). As discussed in Smerdon et al. (2010), CCA employs singular value decomposition (SVD) to perform dimensional reductions separately on T, P, and B. The basic assumption, as in most paleoclimate applications, is that the first few leading modes of EOF-PC pairs contain most of the variance in the target climate field and the multiproxy network. The algorithm seeks an optimal set of truncation parameters (d_t, d_{p}, d_{cca}) that yields good approximations of T, P, and B, respectively. These truncation parameters are chosen by minimizing the area-weighted root mean square error (RMSE) of the reconstruction relative to the target field using a leave-half-out cross-validation procedure (e.g., Chap. 7, Hastie et al., 2008).

Smerdon et al. (2010); Smerdon et al. (2011) have only applied CCA on pseudoproxy networks with constant temporal availability. For this study, we modify the original CCA code to make it applicable to real networks with variable temporal availability, and also made it more computationally efficient. Readers are referred to the SI for more details.

3.1.4 GraphEM

GraphEM (Guillot et al., 2013) is based on the theory of Gaussian graphical models (GGMs, a.k.a. Markov random fields, Whittaker, 1990; Lauritzen, 1996). A GGM makes use of the conditional independence[4] structure of the climate field, in order to reduce the dimensionality and obtain a parsimonious estimate of $\Omega \equiv \Sigma^{-1}$. The conditional independence relations are estimated by solving an ℓ_1-penalized maximum likelihood problem (Friedman et al., 2008). Σ is then estimated in accordance with these conditional independence relations. The resulting $\widehat{\Sigma}$ is sparse and better-conditioned, and therefore is applicable within the OLS framework. This procedure is implemented within the standard EM algorithm without further need for regularization. GraphEM was extensively tested against RegEM-TTLS in

[3] $\mu \in \Re^{1\times p}$ is the temporal mean of the $p = p_{\mathrm{p}} + p_t$ variables to be estimated, while $\Sigma \in \Re^{p\times p}$ is the corresponding covariance matrix.

[4] *Conditional independence*: two random variables X and Y are "conditionally independent" given a random variable Z if, once Z is known, the value of Y does not add additional information about X.

Guillot et al. (2013), albeit with the more idealized proxy design of Smerdon et al. (2011). One goal of this study is to document GraphEM's performance in a more realistic context.

3.2 Numerical experiments

As in Smerdon et al. (2011), all of our pseudoproxy experiments target the annual surface temperature field computed from the NCAR CSM1.4 integration of Ammann et al. (2007), using the correctly oriented version of the CSM1.4 field (Smerdon et al., 2010). Although multiple last millennium simulations are becoming publicly available as part of the Paleoclimate Modelling Intercomparison Project Phase 3 (PMIP3) and through other projects (Fernández-Donado et al., 2013), we selected the CSM1.4 model to enable comparisons with previous work (Mann et al., 2005, 2007; Lee et al., 2008; Mann et al., 2009; Li et al., 2010; Ammann et al., 2010; Smerdon et al., 2010; Smerdon et al., 2011). The target field was spatially masked to approximate the availability of the HadCRUT3v data set used in the Mann et al. (2009) study. We generated 100 realizations of pseudoproxy series for each SNR case (SNR = ∞, 1.0, 0.5, 0.25, local SNR and max SNR) by varying ε in Eq. (1), and then performed reconstructions with the four CFR techniques. The first four cases have fixed SNR, corresponding to networks of homogeneous quality.

The ensemble approach allows us to identify spurious reconstruction skill arising from random noise, and to test the robustness of each method. We also conducted experiments using both the flat and staircase networks. Experiments using the flat network, in which the spatial distribution of pseudoproxies is temporally invariant, serve as the control group in this study. Experiments with the staircase network, correspondingly serve as the test group. By comparing results between these two experiments, we test the null hypothesis that temporal heterogeneity does not affect reconstruction skill.

Reconstruction and calibration intervals are 850–1849 AD and 1850–1995 AD, respectively. Annual means of temperature and pseudoproxies are used for the reconstructions. The same input data are assigned to each CFR method, and standardization (if necessary) is performed internally in each method. Reconstructions are referenced to zero over the calibration interval. Possible trends during the calibration interval are not removed. We also consider computational cost, and investigate numerical methods to speed up each CFR technique. Further details are provided in the SI.

4 Results

4.1 Skill metrics

As is common practice in paleoclimatology, we evaluate reconstruction skill using the coefficient of efficiency (CE), reduction of error (RE) and the coefficient of determination (R^2) (Cook et al., 1994; Bürger, 2007). These validation statistics (Table 1) are related to the mean squared error (MSE) that is commonly used for statistical analysis. They are calculated for both the global mean temperature index and the spatial field; the former provides an aggregate summary of a method's ability to track global climate fluctuations, and the latter evaluates a method's spatial performance. The global mean enables comparisons with index reconstructions, which comprise the majority of published reconstructions of hemispheric and global temperatures (Fig. 6.10 in Jansen et al., 2007).

Table 1. Comparison between R^2, RE, and CE. y_i denotes the ith temperature grid point, and $\hat{y}_i$ denotes the estimation of y_i. $\overline{y}_{\mathrm{c}}$ and $\overline{y}_{\mathrm{v}}$ refer to the mean of $(y_1, y_2, \ldots, y_n)$ over the calibration period and verification period, respectively. Adapted from National Research Council (2006), Chap. 6.

Metric	Expression	Range	Tracks
R^2	$\frac{\left[\sum_{i=1}^{n}(y_i-\overline{y}_{\mathrm{v}})\left(\hat{y}_i-\overline{\hat{y}}_{\mathrm{v}}\right)\right]^2}{\sum_{i=1}^{n}(y_i-\overline{y}_{\mathrm{v}})^2\sum_{i=1}^{n}\left(\hat{y}_i-\overline{\hat{y}}_{\mathrm{v}}\right)^2}$	[0,1]	Phase
RE	$1-\frac{\sum_{i=1}^{n}(y_i-\hat{y}_i)^2}{\sum_{i=1}^{n}(y_i-\overline{y}_{\mathrm{c}})^2}$	$[-\infty, 1]$	Phase and amplitude
CE	$1-\frac{\sum_{i=1}^{n}(y_i-\hat{y}_i)^2}{\sum_{i=1}^{n}(y_i-\overline{y}_{\mathrm{v}})^2}$	$[-\infty, 1]$	Phase, amplitude and mean

As indicated in Fig. 3, the quality of pseudoproxies in the local SNR case, on average, is comparable with the SNR = 0.25 network, and the average SNR of pseudoproxies in the max SNR case is similar to the SNR = 0.5 network. Based on these considerations, we only show results from the spatially heterogenous pseudoproxy networks. The reader is referred to the SI for results on spatially homogenous networks.

4.2 Reconstructing the global mean

In the global mean temperature reconstructions, the overall shape of low-frequency variability is reasonably well reconstructed by all CFR methods (Fig. 5). Warm biases nevertheless are present in all of the reconstructed estimates (von Storch et al., 2004; Smerdon et al., 2011), as expected from regression dilution (e.g., Frost and Thompson, 2000; Tingley et al., 2012). In the local SNR case (Fig. 5a), GraphEM and CCA, in particular, underestimate the amplitude of variability by a factor of 3–5. The variance-bias decomposition (Hastie et al., 2008, Chap. 7) further shows that for GraphEM and CCA (Fig. 6, right column), the bias contributes more than 75 % to their MSE. RegEM-TTLS and M09, on the other hand, have a similar variance but a much smaller bias,

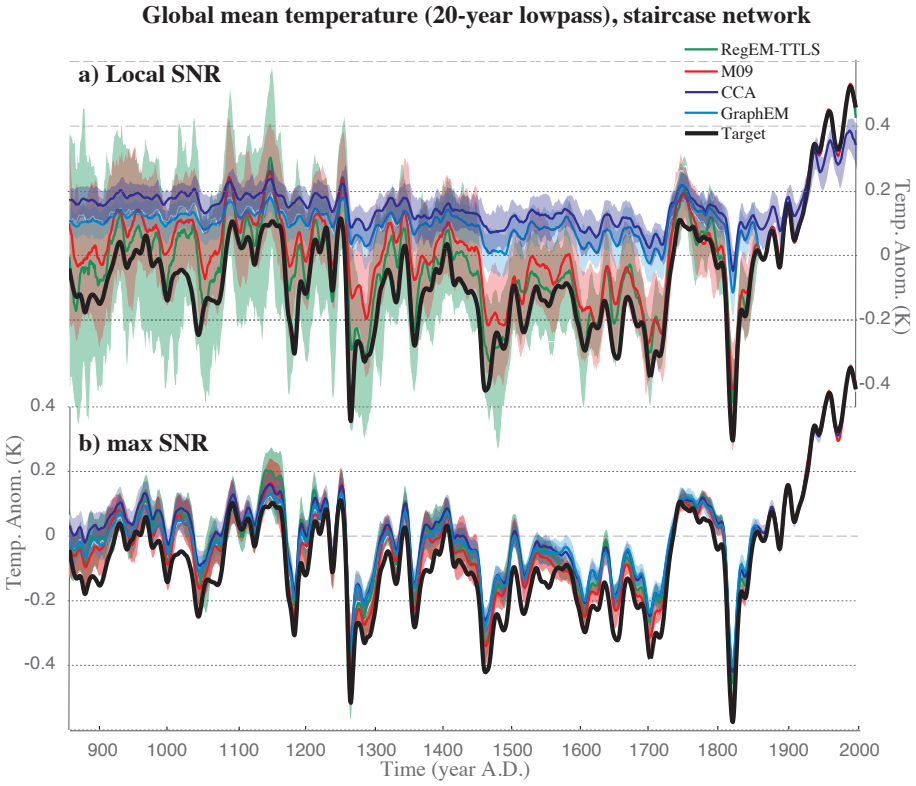

Fig. 5. Area-weighted global mean time series comparison of the four CFR methods, with the staircase network. **(a)** local SNR, **(b)** max SNR. Only the low-frequency (20 yr lowpass) signal is plotted. Black line: target temperature from the CSM1.4 model output; colored lines: reconstructed temperature from median of the reconstruction ensembles; shaded areas: 2.5–97.5 percentiles derived from the reconstruction ensembles.

and thus a correspondingly smaller MSE. Overall, M09 produces the most skillful global mean temperature series in both cases, closely followed by RegEM-TTLS. It would be erroneous, however, to conclude that these two methods produce the closest match to the target, as there is a large spread between different noise realizations.

In assessing the risk properties of each method (Fig. 5a), we find that GraphEM and CCA are more consistent estimators than RegEM-TTLS and M09: their ensemble spreads are much narrower, especially for early reconstruction intervals (prior to 1600 AD). This indicates that any given RegEM-TTLS or M09 reconstruction may yield an inaccurate depiction of the true temperature, and this risk should be kept in mind especially when using M09 and RegEM-TTLS for real-world reconstructions.

4.3 Spatial performance

We now examine the spatial performance of the four CFRs. In Figs. 7 and 8, we summarize the century-by-century skill variation and each CFR's ensemble spread using box plots of the globally averaged CE statistic. CE is shown because it is the most stringent metric among the three in Table 1 (Cook et al., 1994; Ammann and Wahl, 2007), and thus exposes significant contrast between methods (Table S1, Supplement). Box plots characterize the full distribution of the 100-member ensemble: the median of each distribution assesses the tendency for temporal proxy availability to affect reconstruction skill, while the spread yields information about the consistency of each method. The impact of spatial heterogeneities is isolated by plotting spatial maps of CE using the flat network in Figs. 9 and 10, which help us trace spatial errors and inter-method similarities and differences.

4.3.1 Effect of temporal heterogeneities

As evident in Figs. 7 and 8, reconstruction skill varies substantially from century to century, even when proxy availability is time-invariant (gray box plots in Figs. 7, 8, S13, S14, and S3–S12, Supplement). In general, reconstruction skill is highest in the most recent 100 yr slice (1750–1849 AD) but does not decrease monotonically prior to this slice. For instance, reconstruction skill decreases from 1050 to 1450 AD, but increases from 850 to 1049 AD. During this interval (broadly coincident with the "Medieval Climate Anomaly"), the NCAR CSM1.4 model is forced by relatively high solar irradiance and frequent, high-amplitude volcanic eruptions, in particular during the 13th century. Such high-amplitude

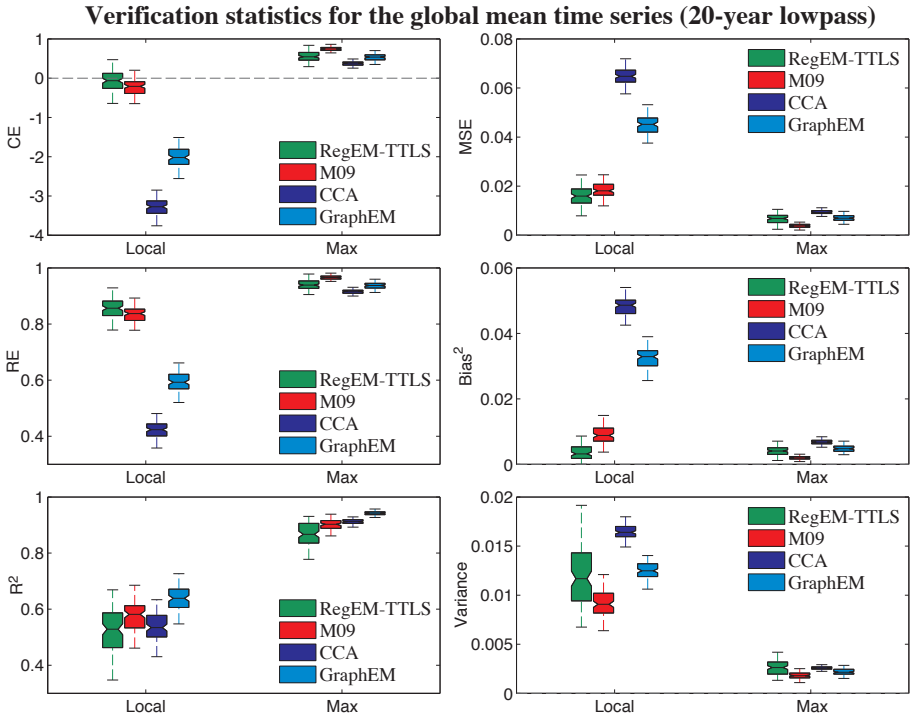

Fig. 6. Summary of verification statistics of the global mean temperature reconstruction ensemble. Note the range of ordinates for each metric is different. MSE = variance + bias^2.

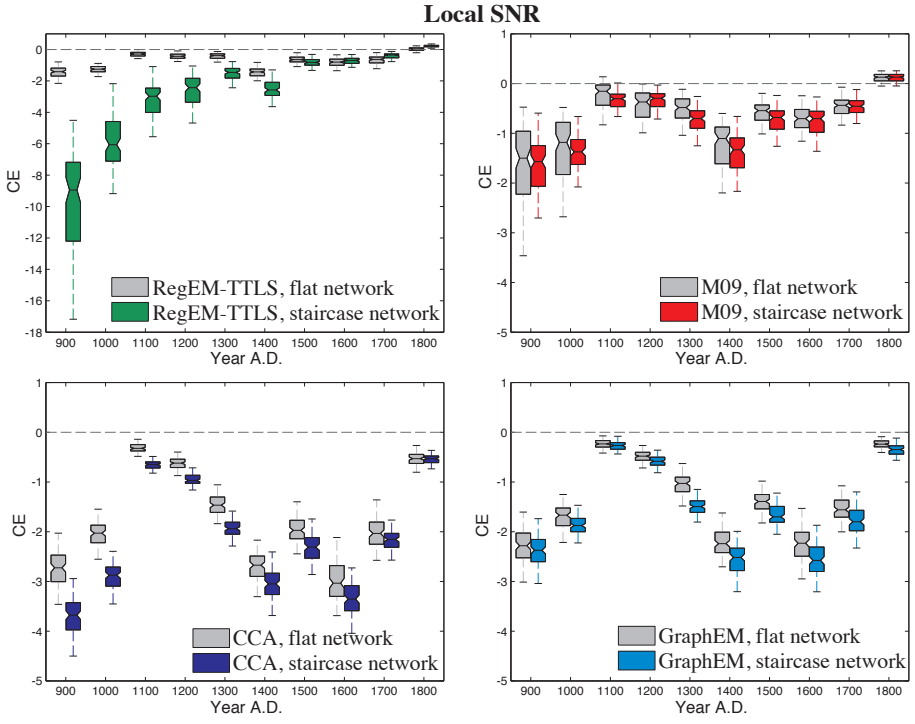

Fig. 7. Temporal variation of globally averaged CE, within CFRs derived from the local SNR network. Spatial CE is first calculated for each grid box, and then global averages are calculated using area-weighted means. Each box plot represents CE scores from the 100-member ensemble for each 100 yr slice between 850 and 1850 AD. For example, the box plot with time slice 900 corresponds to the global mean CE between 850- and 950 AD. Note that the y scale for RegEM-TTLS is different from the other ones.

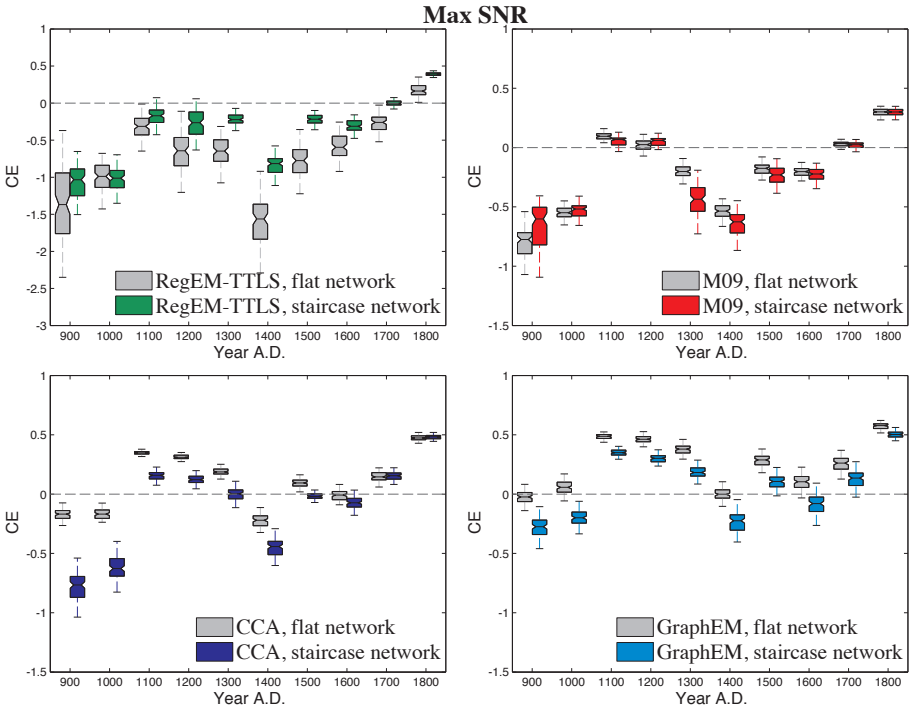

Fig. 8. Same as Fig. 7, but for the max SNR network.

forcing events, which may have more coherent spatial expressions than other fluctuations, appear to be more easily captured by the proxy network. This suggests that reconstruction skill is not only affected by proxy availability and quality, but is also a function of the type and amplitude of climate variations (i.e., internally generated vs. externally forced variations). See Sect. 4 in the SI for more discussions.

An additional important observation to note is that RegEM-TTLS in the local SNR case is distinct from the other three methods (Fig. 7) in that the temporal availability of input data dominates the reconstruction skill, which is a monotonically increasing function of time. The ensemble spread becomes wider back in time as well (consistent with Fig. 5a). The decreasing trend back in time for RegEM-TTLS is partially due to the fact that the method uses a fixed truncation parameter for the estimation of $\widehat{\Sigma}$, despite the declining availability (Fig. 3) and quality (Fig. 4) of pseudoproxies back in time. Consequently, the TTLS solution tends to be less regularized, and hence is dominated by noise (Sima and Van Huffel, 2007). For GraphEM, a fixed graph is used for the entire reconstruction interval. The graph identifies most of the significant proxy-temperature relationships and thus GraphEM is able to efficiently use the relationships for reconstruction. M09 and CCA, on the other hand, have semi-adaptive and adaptive criteria respectively when performing reconstructions, and are less sensitive to temporal heterogeneities in the pseudoproxies. In the max SNR case (Fig. 8), RegEM-TTLS shows a similar pattern to the other CFRs. This implies that for RegEM-TTLS, the reconstruction skill vs. data availability relationship is conditionally dependent on data quality: when proxy quality is relatively high (max SNR), reconstruction skill is relatively insensitive to the choice of truncation parameters in RegEM-TTLS, but – like other CFRs – the skill is still sensitive to high-amplitude climate events.

Despite similarities across reconstructions of global mean temperature (Figs. 5b, 6; max SNR columns), the spatial metrics reveal large discrepancies among methods. Although M09 and RegEM-TTLS perform well reconstructing the global mean temperature, their globally averaged spatial skill only breaches zero in the last two centuries of the reconstruction, even in the max SNR case (Fig. 8). GraphEM and CCA, on the other hand, display high spatial skill for most of the reconstruction period (in the max SNR case, Fig. 8).

4.3.2 Effect of spatial heterogeneities

To isolate the effect of spatial heterogeneities, and to better visualize spatial patterns, Figs. 9 and 10 display the spatial pattern of CE using the flat network over the first (850–949 AD) and last centuries (1750–1849 AD) of the reconstruction[5]. In Fig. 9, a band of high CE scores connecting the eastern equatorial Pacific to North America is evident in all cases, and appears to be a feature of CSM1.4's climate (Smerdon et al., 2011). Similarly, there is some reconstruction skill over other oceans where no proxies are available.

[5]Spatiotemporal maps for each century of the entire reconstruction interval are available in the Supplement, Figs. S3–S12

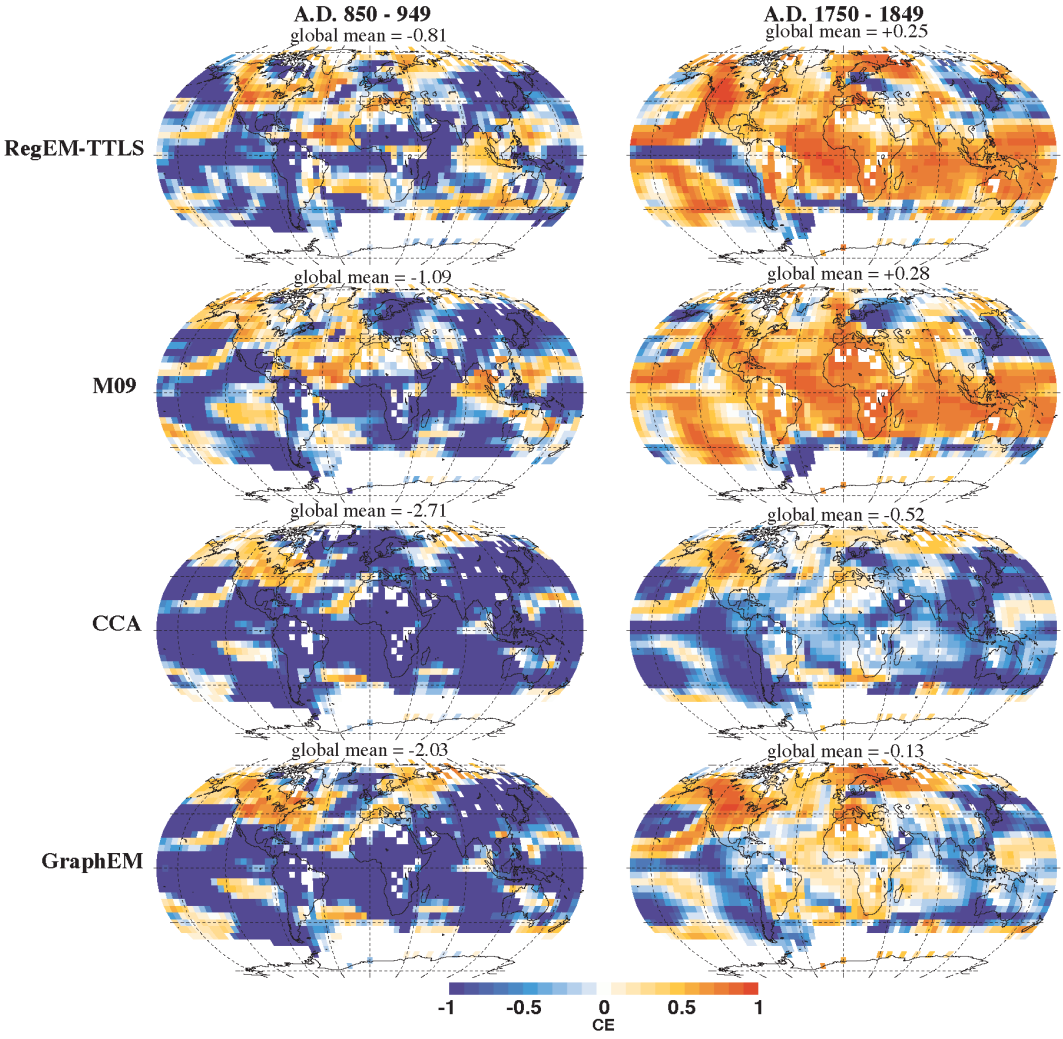

Fig. 9. Spatial pattern of CE, based on ensemble median using the local SNR flat network. 850–949 and 1750–1849 AD represent periods with the minimum and maximum global mean CE over the entire reconstruction interval, respectively.

Collectively, this indicates that modeled teleconnections are effectively exploited by all methods to reconstruct surface temperature in regions with little to no proxy coverage. However, the pattern of enhanced skill associated with ENSO (El Niño–Southern Oscillation) teleconnections vanishes in the 850–949 AD interval when employing RegEM-TTLS or M09 on the max SNR network (Fig. 10), but is still visible in CFRs using CCA and GraphEM. This suggests that the latter two methods are more skillful in resolving such spatial patterns.

Using the local SNR network, we find that RegEM-TTLS and M09 both produce more skillful reconstructions than GraphEM and CCA. In the case of max SNR, however, the results are opposite: reconstructions with CCA and GraphEM are more skillful. In particular, GraphEM is the most skillful method almost everywhere (Fig. 10) and across all time intervals (Figs. 8, S10, Supplement). The goal of dimension reduction in CCA is to increase the solution stability, and is achieved by pre-filtering noise and retaining only a few leading modes. Similarly, GraphEM filters out spurious proxy-temperature relationships and noise by assigning zeros in the precision matrix (Friedman et al., 2008; Hastie et al., 2008; Guillot et al., 2013). In the case of the local SNR network, most proxies have SNRs lower than 0.3 (or equivalently, more than 92 % noise[6]), and hence are dominated by random noise. As a consequence, both CCA and GraphEM tend to treat those proxies as noise and filter them out, shrinking the reconstruction closer to the calibration mean. Figures 9 and 10 suggest that RegEM-TTLS and M09 are likely to be more powerful when data are very noisy (the local SNR network). Nevertheless, the poor risk properties of these two

[6]%noise: the fraction of the variance in the proxy accounted for by the noise component alone, formulated as $\frac{1}{1+\mathrm{SNR}^2}$ (Mann et al., 2007).

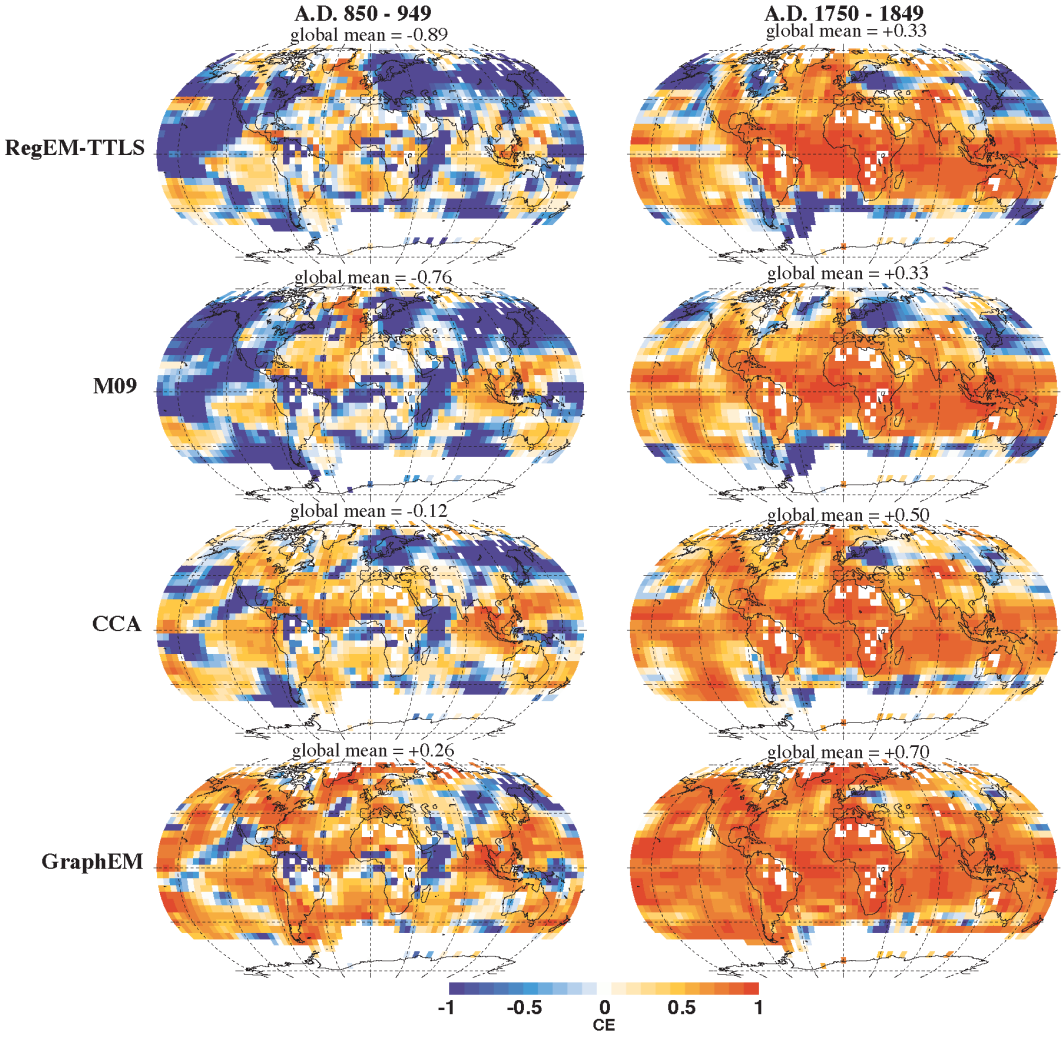

Fig. 10. Same as Fig. 9, but for the max SNR flat network.

methods, as discussed above, indicate that single inferences drawn with these methods should be treated with caution.

Despite the differences among methods described above, some common features emerge when comparing reconstructions with the realistic SNRs to uniform SNR networks (Figs. S9–S12, Supplement)[7]. Compared with CFRs using the SNR = 0.25 network (Fig. S11, Supplement), CFRs using the local SNR network (Fig. S9, Supplement) produce similar results but are less skillful. This is primarily due to the fact that the local SNR network contains only 312 records, while the SNR = 0.25 network includes all 1138 proxies in the M08 database. The comparison between CFRs with SNR = 0.5 (Fig. S12, Supplement) and max SNR networks (Fig. S10, Supplement), however, shows that reconstructions using the max SNR network are much more skillful, especially during early reconstruction periods. We discuss this point below.

[7]Reconstructed patterns are relatively consistent between methods, thus we only present results from GraphEM in this paper.

5 Discussion

In exploring the proxy-temperature relationship, we find that even in the max SNR network, where the highest possible SNR is taken for each proxy, the average SNR is 0.47 (close to the SNR = 0.5 case used in previous studies). Nevertheless, some of the proxies on the right end of the distribution (Fig. 3) may exhibit SNRs higher than 1.0; these proxies drive reconstruction skill upward, so that reconstructions derived from the max SNR network are more skillful than using the SNR = 0.5 network (Figs. S10 and S12 in the Supplement, Table 2). In other words, a small number of high-quality proxies may contribute to a majority of the reconstruction skill. This is encouraging and suggests that global surface temperature may be skillfully reconstructed without requiring uniform spatial sampling over the entire globe. Nevertheless, in our experimental settings, the number of unique temperature grid cells used for sampling T in Eq. (1) is only 128 and 551, in the local and max SNR networks, respectively.

Sampling the same grid cells multiple times effectively increases the SNR for those grid cells, and thus may contribute to the reconstruction skill. This feature is expected to operate in nature as well, however, and provides additional motivation for replicating proxy records. Additionally, we note again that such conclusions might be model-dependent. The skill observed here may result from the low internal variability in the NCAR CSM1.4 model (partially a consequence of its low resolution). To confirm these findings, similar experiments will be conducted with PMIP3-generation last millennium simulations (http://pmip3.lsce.ipsl.fr/).

We also find that differences across CFRs are much smaller in the case of the max SNR network (Fig. 10), indicating that, not surprisingly, reconstructions are much less sensitive to methodology when data quality is high. The spatial reconstruction skill, as expected, is highest in regions of dense proxy availability (e.g., North America), which is consistent with previous findings by Smerdon et al. (2011). The contrast between Figs. 9 and 10, both assuming constant time availability, suggests that our CFR methods, in particular GraphEM and CCA, are more sensitive to data quality than to temporal availability.

As mentioned in Sect. 4.3.1, the ensemble spread for each method is quite different. RegEM-TTLS consistently yields the largest spread, followed by M09, CCA and GraphEM. As an error-in-variable model (EVM), TTLS is designed to minimize the variance of residuals from both the predictands ($\hat{T} - T$) and the predictors ($\hat{P} - P$). The minimization is subject to the estimates of regression coefficients, which in turn depend crucially on the choice of the truncation parameter. Since the noise ε in Eq. (3) is randomly generated, it is sometimes spuriously high in the calibration interval and makes the true signal too noisy for CFR methods to identify, especially in the local SNR network. Under such circumstances, reconstructions using TTLS with fixed truncation may therefore be over-fitting to noise. This confirms an important point made by Christiansen et al. (2009): there is a substantial element of stochasticity in the reconstructions. Hence, one might obtain very different results with the same method applied to different pseudoproxy noise realizations or to different jackknifed proxy networks in real-world reconstructions. In order to improve reconstruction skill when employing RegEM-TTLS, we suggest the development of an algorithm that adaptively selects the regularization parameters using standard statistical theory.

Compared with RegEM-TTLS, M09 produces more skillful reconstructions. In particular, M09 appears advantageous at very high noise levels (local SNR; Figs. 7, S6, Supplement). This is not surprising in part due to the heuristic truncation choices. M09 strongly benefits from the hybrid-frequency approach to perform reconstructions, namely white noise contained in the low-frequency pseudoproxies is effectively filtered out and hence the low-frequency components are better reconstructed in the PPEs. The noise-filtering advantage is likely not present in real-world reconstructions, and the absence of a theoretical justification for the selection criterion makes it vulnerable. Given a different data set or given a different noise model (by varying ε in Eq. (1), for instance, using red noise instead), the 20 yr frequency split might no longer be optimal, and the "33 %-truncation" criterion might also need to change. For the global mean temperature, M09 produces the closest fit to the target overall (Fig. 5), which is due to the fact that truncation parameters are optimized to fit the global mean. However, our results suggest that the optimization comes at the expense of spatial skill, especially during the early reconstruction period (Figs. 10, S5, Supplement).

Reconstructions derived from CCA and GraphEM in general show very similar results, with GraphEM slightly outperforming CCA. In particular, as shown in Figs. 8 and 10, in the max SNR case, GraphEM outperforms other methods at all locations across all time intervals. This indicates that given enough high-quality data, GraphEM can produce the most skillful reconstructions. The strength of GraphEM is especially noticeable in regions of dense proxy sampling. For instance, over North America (Fig. S10, Supplement), the other three methods display negative CE scores prior to 1650 AD, yet GraphEM displays positive CE scores over the entire reconstruction interval. Nevertheless, as noted in Fig. 9, GraphEM does not perform as well in the local SNR setting as it does using the max SNR network. GraphEM's superior performances in the former case and unsatisfying performances in the latter are both largely due to the graphical structure selected by the GraphEM algorithm. In the max SNR case, pseudoproxies have, by design, much higher SNR than they do in the local SNR case. Thus most of the significant proxy-temperature relationships are effectively detected and exploited by the method. In the local SNR case, on the other hand, very few significant relationships are detected and thus GraphEM fails to produce meaningful CFRs. Despite sensitivity to data quality, another potential cause of GraphEM's poor performance in the local SNR cases is the choice of the graph. Currently, as described in Guillot et al. (2013), the graph is a fixed choice for the entire reconstruction period, which is based solely on instrumental proxy-temperature relationships. To improve the performance of GraphEM-based CFRs, adaptive choices of the graph should be made for each century of the reconstruction. Through this approach, available proxies from each century will be more effectively used. Alternative methods should also be explored to estimate the graph.

As illustrated in Fig. 6, both GraphEM and CCA suffer from large mean biases, which contribute to more than 75 % of the total MSE associated with reconstructions from each method. It is well-known that introducing a certain amount of bias can lead to a reduction in MSE (compared to using an unbiased estimator). However, this often results in the reduction of the corresponding variance, i.e., the amplitude of past climatic variations are underestimated (Fig. 5). CCA often does best by measure of R^2 values (Table 2). This is

Table 2. Verification statistics summary for the global mean temperature time series, using the staircase network. CE, RE, R^2 and bias are computed for the low-frequency component (20 yr low-pass) of the reconstructed global mean. All numbers given outside of parentheses are the mean of the 100-member ensemble; numbers in parentheses are the corresponding standard deviation.

Method	SNR	CE	RE	R^2	bias
RegEM-TTLS	∞	$+0.92_{0.00}$	$+0.99_{0.00}$	$+0.92_{0.00}$	$+0.02_{0.00}$
	1.0	$+0.84_{0.02}$	$+0.98_{0.00}$	$+0.91_{0.01}$	$+0.03_{0.01}$
	0.5	$+0.78_{0.07}$	$+0.97_{0.01}$	$+0.87_{0.03}$	$+0.03_{0.02}$
	0.25	$+0.53_{0.13}$	$+0.94_{0.02}$	$+0.70_{0.06}$	$+0.04_{0.02}$
	local	$-0.12_{0.39}$	$+0.85_{0.05}$	$+0.52_{0.09}$	$+0.06_{0.03}$
	max	$+0.55_{0.13}$	$+0.94_{0.02}$	$+0.87_{0.04}$	$+0.07_{0.01}$
M09	∞	$+0.93_{0.00}$	$+0.96_{0.00}$	$+0.88_{0.00}$	$+0.02_{0.00}$
	1.0	$+0.74_{0.02}$	$+0.97_{0.00}$	$+0.88_{0.01}$	$+0.05_{0.02}$
	0.5	$+0.69_{0.07}$	$+0.96_{0.01}$	$+0.85_{0.02}$	$+0.05_{0.01}$
	0.25	$+0.30_{0.21}$	$+0.91_{0.03}$	$+0.67_{0.08}$	$+0.07_{0.02}$
	local	$-0.24_{0.26}$	$+0.83_{0.04}$	$+0.57_{0.06}$	$+0.11_{0.02}$
	max	$+0.74_{0.06}$	$+0.97_{0.01}$	$+0.90_{0.03}$	$+0.05_{0.01}$
CCA	∞	$+0.96_{0.00}$	$+1.00_{0.00}$	$+0.97_{0.00}$	$+0.01_{0.00}$
	1.0	$+0.78_{0.04}$	$+0.97_{0.01}$	$+0.94_{0.01}$	$+0.05_{0.01}$
	0.5	$-0.18_{0.12}$	$+0.84_{0.02}$	$+0.88_{0.02}$	$+0.13_{0.01}$
	0.25	$-2.92_{0.28}$	$+0.47_{0.04}$	$+0.61_{0.06}$	$+0.24_{0.01}$
	local	$-3.28_{0.25}$	$+0.42_{0.03}$	$+0.54_{0.06}$	$+0.25_{0.01}$
	max	$+0.37_{0.07}$	$+0.92_{0.01}$	$+0.91_{0.01}$	$+0.09_{0.01}$
GraphEM	∞	$+0.95_{0.00}$	$+0.99_{0.00}$	$+0.97_{0.00}$	$+0.01_{0.00}$
	1.0	$+0.68_{0.07}$	$+0.96_{0.01}$	$+0.95_{0.01}$	$+0.06_{0.01}$
	0.5	$+0.53_{0.10}$	$+0.94_{0.01}$	$+0.90_{0.01}$	$+0.08_{0.01}$
	0.25	$-0.43_{0.38}$	$+0.81_{0.05}$	$+0.75_{0.04}$	$+0.14_{0.02}$
	local	$-2.00_{0.30}$	$+0.60_{0.04}$	$+0.64_{0.05}$	$+0.21_{0.01}$
	max	$+0.53_{0.09}$	$+0.94_{0.01}$	$+0.94_{0.01}$	$+0.08_{0.01}$

expected: the method regularizes by maximizing the cross-correlation between the proxy and target matrices, but without further constraining the variance. One possible modification would be matching the variance of CCA reconstructions to the variance of the target data in each grid cell during the calibration interval. In doing this, more variance would be preserved for networks affected by declining data availability. However, this modification would inflate errors and the solution could no longer be interpreted as minimizing the calibration misfit.

By contrasting the CFRs derived from four methods in both the spatial and temporal context, we find that, despite some general agreements (Fig. 5b) and reasonable skill (Table 1) in the global mean temperature reconstruction, the four methods yield large spatial differences, and their validation scores in terms of CE can still be large locally. This confirms previous findings in Smerdon et al. (2011), that the global mean temperature series is a poor indicator of spatial skill, and that spatial performance metrics are crucial for the assessment of different CFR techniques (e.g., Li and Smerdon, 2012). The results also highlight the difficulty in jointly optimizing the spatial skill and the global mean temperature. Fundamentally, reconstruction skill can be assessed using MSE. As discussed in the previous paragraph, in order to find the lowest MSE (thus the best reconstruction), a tradeoff must be found between bias and variance. It therefore is most likely that the lowest MSE for the spatial field and the global mean time series are not given by the same set of regularization parameters.

We also calculated a suite of diagnostics for reconstruction skill and its dependence on (1) the number of proxies, (2) the average SNR, and (3) the sum of SNR in each grid box. No apparent relationships between these variables and spatial skill were found (Figs. S15–S20, Supplement). Our experiments also highlight the need for methodological refinements, since no method can consistently perform well in all cases for both index and field reconstructions. We find that both RegEM-TTLS and M09 produce meaningful global mean reconstructions, but do not perform as well in the spatial field. The disagreement between the field and index reconstruction was explored in Guillot et al. (2013), in which it is found that the skillful performance of TTLS-based global mean temperature reconstructions involves considerable cancellation between positive and negative deviations from the true field at any given grid point (see Supplement). Hence, the fidelity of the reconstructed global mean is a poor indication of spatial skill (Figs. S21, S22, Supplement).

Additionally, we note that all methods consistently introduce a warm bias to the global mean temperature reconstruction, even in the max SNR setting. As previously found in von Storch et al. (2004), Christiansen et al. (2009), and Smerdon et al. (2011), regression-based CFR methods are generally associated with variance losses and large mean biases. These are an inevitable by-products of linearly regressing temperature onto proxies, and are especially severe if proxies are subject to extensive errors. This well-known problem is called regression dilution (e.g., Frost and Thompson, 2000; Tingley et al., 2012). It commonly translates into reconstructions that are always biased towards the mean of the calibration interval. Ammann et al. (2010) has proposed a correction to regression dilution in the context of index reconstructions, which minimizes out-of-sample bias over subsets of the calibration interval. An alternative is to consider methods that respect the proxies' physical dependence on temperature, and express proxies as a function of temperature (Christiansen, 2011). Tingley and Li (2012) find that this leads to a reduced bias but may also lead to infinite variance in very noisy cases. They suggest an alternative solution leveraging Bayesian hierarchical models, wherein a proxy's dependence on temperature is formulated using process-based forward models at the data level, allowing for an elimination of the variance inflation and an internally consistent quantification of uncertainties (see also Christiansen, 2012, for more discussions).

6 Conclusions

An updated pseudoproxy network design has been constructed with more realistic characteristics: for the first time, pseudoproxies were sampled with spatiotemporal characteristics that reflect heterogeneities in proxy quality and proxy attrition back in time. The updated network has allowed an assessment of the spatial performance of four different CFR techniques using a comprehensive suite of experiments.

Results based on the max SNR network show relatively small CFR sensitivity to the choice of methodology when SNR is high. However, results are strongly method-dependent in sample-starved settings. Overall, reconstructions are generally better in regions with dense proxy sampling, although teleconnections are also exploited by these CFR methods, in particular CCA and GraphEM, to derive spatial skill outside of directly sampled regions.

The effect of temporal heterogeneities of proxy availability is counterintuitive. We find that despite the declining data availability back in time, reconstruction skill does not necessarily follow suit. Rather, our experiments show that forced, high-amplitude climatic events have a larger impact on reconstruction skill and are more easily resolved by these methods, even when data availability is low. This conclusion is nevertheless model-dependent, and needs to be verified with PPEs using output from other GCMs.

Our experiments also show that no method universally outperforms another, and that each method has its own strengths and weaknesses. Overall, RegEM-TTLS and M09 produce more skillful index reconstructions (global mean temperature, Fig. 6), and retain a higher skill than other methods when proxies are very noisy (local SNR network, Fig. 9). However, RegEM-TTLS displays large ensemble spreads, partially due to its fixed choice of truncation parameter and high sensitivity to noise in the data. This emphasizes the high risk associated with conclusions from a single noise realization (such as a real-world reconstruction). The stochasticity of a reconstruction method should therefore always be seriously considered when evaluating a real-world reconstruction (Christiansen et al., 2009).

The heuristic parameter choices proposed by M09 show the potential for RegEM-TTLS to produce meaningful global mean temperature reconstructions. The setup nevertheless deviates greatly from the standard implementation of RegEM-TTLS. Additionally, for global mean reconstructions, both RegEM-TTLS and M09 involve error cancellations that are not readily noticeable (Sect. 2.4 in the Supplement, Figs. S21, S22). This might explain some of the observed divergence between the quality of index and field reconstructions using these methods.

CCA and GraphEM generally produce very similar results, but the former suffers from larger variance losses and associated mean biases. This can be attributed to the manner in which the method selects for the optimal estimates of regression coefficients $\widehat{B}$. Given enough high-quality data, reconstructions using GraphEM display a higher spatial skill than the other three methods everywhere in the field, and in particular over the oceans and regions with denser proxy sampling. This suggests that the reconstruction strongly benefits from the improved covariance estimation induced by the use of Gaussian graphical models.

Given the large performance differences among various CFR methods in the pseudoproxy context, we emphasize that unless reconstructions with various methods provide very similar spatiotemporal information, real-world reconstructions derived from a single method should be viewed with caution. In agreement with Smerdon et al. (2011), we recommend applying as many methods as possible to make robust conclusions. Additionally, the exact pattern of spatial skill varies according to the GCM simulation used as the basis of the PPEs. Multiple PMIP3 last millennium simulations should ideally be used to validate the present results. Future studies should also rigorously model real-world conditions, including persistence, noise characteristics, and a mechanistic representation of climate proxies. Finally, we emphasize the fundamental difficulty in finding a bias-variance tradeoff that optimizes the reconstruction of both the temperature field and its global mean. Future studies should explore solutions that jointly minimize spatial and temporal errors.

Supplementary material related to this article is available online at http://www.clim-past.net/10/1/2014/cp-10-1-2014-supplement.pdf.

Acknowledgements. The authors thank Sylvia Dee and Adam Vaccaro for comments that improved the presentation of this manuscript, and also thank Sandrah Eckel and Gareth James for their statistical insight. The authors acknowledge NSF awards AGS1003818 and AGS0902436, NOAA grant NA10OAR4320137, and computational resources from the USC High Performance Computing Center.

Edited by: S. Bronnimann

References

Ammann, C. M. and Wahl, E.: The importance of the geophysical context in statistical evaluations of climate reconstruction procedures, Climatic Change, 85, 71–88, doi:10.1007/s10584-007-9276-x, 2007.

Ammann, C. M., Joos, F., Schimel, D. S., Otto-Bliesner, B. L., and Tomas, R. A.: Solar influence on climate during the past millennium: Results from transient simulations with the NCAR Climate System Model, Proc. Nat. Acad. Sc., 104, 3713–3718, doi:10.1073/pnas.0605064103, 2007.

Ammann, C. M., Genton, M. G., and Li, B.: Technical Note: Correcting for signal attenuation from noisy proxy data in climate reconstructions, Clim. Past, 6, 273–279, doi:10.5194/cp-6-273-2010, 2010.

Anchukaitis, K. J., Evans, M. N., Kaplan, A., Vaganov, E. A., Hughes, M. K., Grissino-Mayer, H. D., and Cane, M. A.: Forward modeling of regional scale tree-ring patterns in the southeastern United States and the recent influence of summer drought, Geophys. Res. Lett., 33, L04705, doi:10.1029/2005GL025050, 2006.

Anderson, T.: An Introduction to Multivariate Statistical Analysis, 3rd Edn., John Wiley & Sons, Inc., New York, 2003.

Annan, J. D. and Hargreaves, J. C.: Identification of climatic state with limited proxy data, Clim. Past, 8, 1141–1151, doi:10.5194/cp-8-1141-2012, 2012.

Barnett, T. P. and Preisendorfer, R.: Origins and Levels of Monthly and Seasonal Forecast Skill for United States Surface Air Temperatures Determined by Canonical Correlation Analysis, Mon. Weather Rev., 115, 1825–1850, 1987.

Bradley, R. S.: Are there optimum sites for global paleotemperature reconstruction?, vol. 41 of NATO ASI, chap. Climate variations and forcing mechanisms of the last 2000 years, Springer, Berlin, Heidelberg, New York, 603–624, 1996.

Briffa, K. R., Osborn, T. J., Schweingruber, F. H., Harris, I. C., Jones, P. D., Shiyatov, S. G., and Vaganov, E. A.: Low-frequency temperature variations from a northern tree ring density network, J. Geophys. Res., 106, 2929–2942, doi:10.1029/2000JD900617, 2001.

Brohan, P., Kennedy, J. J., Harris, I., Tett, S. F. B., and Jones, P. D.: Uncertainty estimates in regional and global observed temperature changes: A new data set from 1850, J. Geophys. Res., 111, D12106, doi:10.1029/2005JD006548, 2006.

Bürger, G.: On the verification of climate reconstructions, Clim. Past, 3, 397–409, doi:10.5194/cp-3-397-2007, 2007.

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

Christiansen, B.: Reply to "Comments on 'Reconstructing the NH Mean Temperature: Can Underestimation of Trends and Variability be Avoided?"', J. Climate, 25, 3447–3452, doi:10.1175/JCLI-D-11-00162.1, 2012.

Christiansen, B. and Ljungqvist, F. C.: Reply to "Comments on 'Reconstruction of the Extratropical NH Mean Temperature over the Last Millennium with a Method That Preserves Low-Frequency Variability"', J. Climate, 25, 7998–8003, doi:10.1175/JCLI-D-11-00642.1, 2012.

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

Christiansen, B., Schmith, T., and Thejll, P.: Reply, J. Climate, 23, 2839–2844, doi:10.1175/2010JCLI3281.1, 2010.

Cobb, K. M., Kiefer, T., Lough, J. M., Overpeck, J. T., and Tudhope, A. W.: Final Report, Tech. rep., CLIVAR-PAGES Workshop on representing and reducing uncertainties in high-resolution climate proxy data, Trieste, Italy, 2008.

Cook, E. R., Briffa, K. R., and Jones, P. D.: Spatial regression methods in dendroclimatology: A review and comparison of two techniques, Int. J. Climatol., 14, 379–402, 1994.

Cook, E. R., Woodhouse, C. A., Eakin, C. M., Meko, D. M., and Stahle, D. W.: Long-Term Aridity Changes in the Western United States, Science, 306, 1015–1018, doi:10.1126/science.1102586, 2004.

Cook, E. R., Seager, R., Cane, M. A., and Stahle, D. W.: North American drought: Reconstructions, causes, and consequences, Earth Sci. Rev., 81, 93–134, doi:10.1016/j.earscirev.2006.12.002, 2007.

Crowley, T. J. and Lowery, T. S.: How Warm Was the Medieval Warm Period?, AMBIO, 29, 51–54, doi:10.1579/0044-7447-29.1.51, 2000.

D'Arrigo, R., Wilson, R., and Jacoby, G.: On the long-term context for late twentieth century warming, J. Geophys. Res., 111, D03103, doi:10.1029/2005JD006352, 2006.

Dempster, A. P., Laird, N. M., and Rubin, D. B.: Maximum Likelihood from Incomplete Data via the EM Algorithm, J. Roy. Stat. Soc. B , 39, 1–38, 1977.

Ebisuzaki, W.: A Method to Estimate the Statistical Significance of a Correlation When the Data Are Serially Correlated, J. Climate, 10, 2147–2153, doi:10.1175/1520-0442(1997)010<2147%3AAMTETS>2.0.CO%3B2, 1997.

Emile-Geay, J., Cobb, K. M., Mann, M. E., and Wittenberg, A. T.: Estimating Central Equatorial Pacific SST variability over the Past Millennium, Part 1: Methodology and Validation, J. Climate, 26, 2302–2328, doi:10.1175/JCLI-D-11-00511.1, 2013a.

Emile-Geay, J., Cobb, K. M., Mann, M. E., and Wittenberg, A. T.: Estimating Central Equatorial Pacific SST variability over the Past Millennium, Part 2: Reconstructions and Uncertainties, J. Climate, 26, 2329–2352, doi:10.1175/JCLI-D-11-00511.1, 2013b.

Evans, M. N.: Toward forward modeling for paleoclimatic proxy signal calibration: A case study with oxygen isotopic composition of tropical woods, Geochem. Geophy. Geosy., 8, Q07008, doi:10.1029/2006GC001406, 2007.

Evans, M. N., Kaplan, A., and Cane, M. A.: Pacific sea surface temperature field reconstruction from coral δ^{18}O data using reduced space objective analysis, Paleoceanography, 17, 7-1–7-13, doi:10.1029/2000PA000590, 2002.

Evans, M. N., Tolwinski-Ward, S., Thompson, D., and Anchukaitis, K. J.: Applications of proxy system modeling in high resolution paleoclimatology, Quaternary Sci. Rev., 76, 16–28, doi:10.1016/j.quascirev.2013.05.024, 2013.

Fernández-Donado, L., González-Rouco, J. F., Raible, C. C., Ammann, C. M., Barriopedro, D., García-Bustamante, E., Jungclaus, J. H., Lorenz, S. J., Luterbacher, J., Phipps, S. J., Servonnat, J., Swingedouw, D., Tett, S. F. B., Wagner, S., Yiou, P., and Zorita, E.: Large-scale temperature response to external forcing in simulations and reconstructions of the last millennium, Clim. Past, 9, 393–421, doi:10.5194/cp-9-393-2013, 2013.

Fierro, R. D., Golub, G. H., Hansen, P. C., and O'Leary, D. P.: Regularization by truncated total least squares, SIAM J. Sci. Comput., 18, 1223–1241, 1997.

Friedman, J., Hastie, T., and Tibshirani, R.: Sparse inverse covariance estimation with the graphical lasso, Biostat, 9, 432–441, doi:10.1093/biostatistics/kxm045, 2008.

Frost, C. and Thompson, S. G.: Correcting for regression dilution bias: comparison of methods for a single predictor variable, J. Roy. Stat. Soc. A, 163, 173–189, doi:10.1111/1467-985X.00164, 2000.

Guillot, D., Rajaratnam, B., and Emile-Geay, J.: Statistical Paleoclimate Reconstructions via Markov Random Fields, http://arxiv.org/abs/1309.6702, Ann. Appl. Stat., submitted, 2013.

Hansen, J. and Lebedeff, S.: Global trends of measured surface air temperature, J. Geophys. Res., 92, 13345–13372, doi:10.1029/JD092iD11p13345, 1987.

Hansen, P. C.: Rank-Deficient and Discrete III – Posed Problems: Numerical Aspects of Linear Inversion, SIAM Monogr. on Mathematical Modeling and Computation, Society for Industrial and Applied Mathematics, Philadelphia, PA, 1998.

Hastie, T., Tibshirani, R., and Friedman, J.: The elements of statistical learning: data mining, inference and prediction, 2nd Edn., Springer, New York, 2008.

Hegerl, G. C., Crowley, T. J., Hyde, W. T., and Frame, D. J.: Climate sensitivity constrained by temperature reconstructions over the past seven centuries, Nature, 440, 1029–1032, doi:10.1038/nature04679, 2006.

Hoerl, A. E. and Kennard, R. W.: Ridge regression: Biased estimation for non-orthogonal problems, Technometrics, 12, 55–67, 1970a.

Hoerl, A. E. and Kennard, R. W.: Ridge regression: Applications to non-orthogonal problems, Technometrics, 12, 69–82, correction, 12, 723, 1970b.

Jansen, E., Overpeck, J., Briffa, K., Duplessy, J.-C., Joos, F., Masson-Delmotte, V., Olago, D., Otto-Bliesner, B., Peltier, W., Rahmstorf, S., Ramesh, R., Raynaud, D., Rind, D., Solomina, O., Villalba, R., and Zhang, D.: Climate Change 2007: The Physical Science Basis, Contribution of Working Group I to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change, in: chap. Palaeoclimate, Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA, 2007.

Jones, P. D. and Mann, M.: Climate over past millennia, Rev. Geophys., 42, RG2002, doi:10.1029/2003RG000143, 2004.

Jones, P. D., Briffa, K., Osborn, T., Lough, J., van Ommen, T., Vinther, B., Luterbacher, J., Wahl, E., Zwiers, F., Mann, M., Schmidt, G., Ammann, C., Buckley, B., Cobb, K., Esper, J., Goosse, H., Graham, N., Jansen, E., Kiefer, T., Kull, C., Kuttel, M., Mosley-Thompson, E., Overpeck, J., Riedwyl, N., Schulz, M., Tudhope, A., Villalba, R., Wanner, H., Wolff, E., and Xoplaki, E.: High-resolution palaeoclimatology of the last millennium: a review of current status and future prospects, Holocene, 19, 3–49, doi:10.1177/0959683608098952, 2009.

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, Holocene, 8, 455–471, doi:10.1191/095968398667194956, 1998.

Küttel, M., Luterbacher, J., Zorita, E., Xoplaki, E., Riedwyl, N., and Wanner, H.: Testing a European winter surface temperature reconstruction in a surrogate climate, Geophys. Res. Lett., 34, L07710, doi:10.1029/2006GL027907, 2007.

Lauritzen, S. L.: Graphical Models, Clarendon Press, Oxford, 1996.

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

Li, B. and Smerdon, J. E.: Defining spatial comparison metrics for evaluation of paleoclimatic field reconstructions of the Common Era, Environmetrics, 23, 394–406, doi:10.1002/env.2142, 2012.

Li, B., Nychka, D. W., and Ammann, C. M.: The value of multiproxy reconstruction of past climate, J. Am. Stat. Assoc., 105, 883–911, doi:10.1198/jasa.2010.ap09379, 2010.

Little, R. J. A. and Rubin, D. B.: Statistical analysis with missing data, Wiley series in probability and statistics, New York, NY, 2002.

Liu, Z. and Alexander, M. A.: Atmospheric bridge, oceanic tunnel, and global climatic teleconnections, Rev. Geophys., 45, RG2005, doi:10.1029/2005RG000172, 2007.

Luterbacher, J., Dietrich, D., Xoplaki, E., Grosjean, M., and Wanner, H.: European Seasonal and Annual Temperature Variability, Trends, and Extremes Since 1500, Science, 303, 1499–1503, doi:10.1126/science.1093877, 2004.

Mann, M. E. and Jones, P. D.: Global surface temperatures over the past two millennia, Geophys. Res. Lett., 30, 1820, doi:10.1029/2003GL017814, 2003.

Mann, M. E. and Rutherford, S.: Climate reconstruction using "Pseudoproxies", Geophys. Res. Lett., 29, 139-1–139-4, doi:10.1029/2001GL014554, 2002.

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, doi:10.1038/33859, 1998.

Mann, M. E., Bradley, R. S., and Hughes, M. K.: Northern hemisphere temperatures during the past millennium: Inferences, uncertainties, and limitations, Geophys. Res. Lett., 26, 759–762, doi:10.1029/1999GL900070, 1999.

Mann, M. E., Rutherford, S., Wahl, E., and Ammann, C.: Testing the fidelity of methods used in proxy-based reconstructions of past climate, J. 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, doi:10.1029/2006JD008272, 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, P. Natl. Acad. Sci., 105, 13252–13257, 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, doi:10.1126/science.1177303, 2009.

National Research Council: Surface Temperature Reconstructions for the Last 2,000 Years, The National Academies Press, Washington, D.C., 2006.

Rutherford, S., Mann, M. E., Delworth, T. L., and Stouffer, R. J.: Climate Field Reconstruction under Stationary and Nonstationary Forcing, J. Climate, 16, 462–479, doi:10.1175/1520-0442(2003)016<0462:CFRUSA>2.0.CO;2, 2003.

Rutherford, S., Mann, M. E., Osborn, T. J., Bradley, R. S., Briffa, K. R., Hughes, M. K., and Jones, P. D.: Proxy-Based Northern Hemisphere Surface Temperature Reconstructions: Sensitivity to Method, Predictor Network, Target Season, and Target Domain, J. Climate, 18, 2308–2329, doi:10.1175/JCLI3351.1, 2005.

Rutherford, S., Mann, M., Ammann, C., and Wahl, E.: Comment on: "A surrogate ensemble study of climate reconstruction methods: Stochasticity and robustness" by Christiansen, Schmith and Thejll, J. Climate, 23, 2832–2838, doi:10.1175/2009JCLI3146.1, 2010.

Schneider, T.: Analysis of Incomplete Climate Data: Estimation of Mean Values and Covariance Matrices and Imputation of Missing Values, J. Climate, 14, 853–871, doi:10.1175/1520-0442(2001)014<0853%3AAOICDE>2.0.CO%3B2, 2001.

Sima, D. M. and Van Huffel, S.: Level choice in truncated total least squares, Computational Statistics & Data Analysis, 52, 1103–1118, doi:10.1016/j.csda.2007.05.015, 2007.

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. and Kaplan, A.: Comments on "Testing the Fidelity of Methods Used in Proxy-Based Reconstructions of Past Climate": The Role of the Standardization Interval, J. Climate, 20, 5666–5670, doi:10.1175/2007JCLI1794.1, 2007.

Smerdon, J. E., Kaplan, A., and Chang, D.: On the Origin of the Standardization Sensitivity in RegEM Climate Field Reconstructions*, J. Climate, 21, 6710–6723, doi:10.1175/2008JCLI2182.1, 2008.

Smerdon, J. E., Kaplan, A., Chang, D., and Evans, M. N.: A Pseudoproxy Evaluation of the CCA and RegEM Methods for Reconstructing Climate Fields of the Last Millennium*, J. Climate, 23, 4856–4880, doi:10.1175/2010JCLI3328.1, 2010.

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, Geophys. Res. Lett., 38, doi:10.1029/2011GL047372, 2011.

Thompson, D. M., Ault, T. R., Evans, M. N., Cole, J. E., and Emile-Geay, J.: Comparison of observed and simulated tropical climate trends using a forward model of coral of δ^{18}O, Geophys. Res. Lett., 38, 14, doi:10.1029/2011GL048224, 2011.

Tikhonov, A. N. and Arsenin, V. Y.: Solution of Ill-Posed Problems, in: Scripta Series in Mathematics, V. H. Winston and Sons, Washington, 1977.

Tingley, M. P. and Huybers, P.: A Bayesian Algorithm for Reconstructing Climate Anomalies in Space and Time, Part 1: Development and applications to paleoclimate reconstruction problems, J. 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 2: Comparison with the Regularized Expectation-Maximization Algorithm, J. Climate, 23, 2782–2800, doi:10.1175/2009JCLI3016.1, 2010b.

Tingley, M. P. and Huybers, P.: Recent temperature extremes at high northern latitudes unprecedented in the past 600 ears, Nature, 496, 201–205, doi:10.1038/nature11969, 2013.

Tingley, M. P. and Li, B.: Comments on "Reconstructing the NH mean temperature: Can underestimation of trends and variability be avoided?", J. Climate, 25, 3441–3446, doi:10.1175/JCLI-D-11-00005.1, 2012.

Tingley, M. P., Craigmile, P. F., Haran, M., Li, B., Mannshardt, E., and Rajaratnam, B.: Piecing together the past: statistical insights into paleoclimatic reconstructions, Quaternary Sci. Rev., 35, 1–22, doi:10.1016/j.quascirev.2012.01.012, 2012.

Van Huffel, S. and Vandewalle, J.: The Total Least Squares Problem: Computational Aspects and Analysis, vol. 9 of Frontiers in Applied Mathematics, SIAM, Philadelphia, PA, 1991.

von Storch, H., Zorita, E., Jones, J. M., Dimitriev, Y., González-Rouco, F., and Tett, S. F. B.: Reconstructing Past Climate from Noisy Data, Science, 306, 679–682, doi:10.1126/science.1096109, 2004.

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.

Whittaker, J.: Graphical Models in Applied Multivariate Statistics, John Wiley and Sons, Chichester, UK, 1990.

Wilson, R., Cook, E., D'Arrigo, R., Riedwyl, N., Evans, M. N., Tudhope, A., and Allan, R.: Reconstructing ENSO: the influence of method, proxy data, climate forcing and teleconnections, J. Quaternary Sci., 25, 62–78, doi:10.1002/jqs.1297, 2010.